

SOME STATISTICAL ASPECTS OF HIGH-FREQUENCY NONSYNCHRONOUS STOCK PRICE DATA

Thesis submitted at the Indian Statistical Institute
in partial fulfilment of the requirements for
the degree of Doctor of Philosophy
in Statistics

by

Arnab Chakrabarti

Thesis Supervisor: Dr. Rituparna Sen

Applied Statistics Unit

Indian Statistical Institute, Bangalore

JULY 2019

Dedicated to

My Father Sri Debkumar Chakrabarti

&

Late Grandfather Sri Harendrakumar Chakrabarti

Acknowledgments

First and foremost, I would like to thank my supervisor Prof. Rituparna Sen for her guidance, and constant encouragement throughout the period of my Ph.D. life. She supported me in every possible way- both professional and personal- without which this thesis would not have materialized. She continuously motivated me to attend various conferences and workshops such that I can get enough food for innovative thoughts. She always encourages me to meet new people, to learn new things and to engage in various intellectual activities. Her influence on me goes much beyond this Ph.D. thesis- almost every other aspect of life.

I would like to thank Prof. Gopal Basak, Prof. Diganta Mukherjee and Prof. Sourish Das for helping me to come up with this thesis in its final shape. Their support and kind suggestions are valuable beyond measure. I am also grateful to my teachers during M.Sc., particularly Prof. Arindam Sengupta, Prof. Bhaswati Ganguly, Prof. Sugato Sen Roy and Prof. Gaurangadeb Chatterjee for teaching me valuable lessons in Statistics at the University of Calcutta.

I owe my deepest gratitude to the professors at Indian Statistical Institute (ISI) Chennai, especially Prof. Prabuddha Chakraborty , Prof. Sujata Ghosh and Prof. Amit Biswas for making my stay at Chennai really pleasurable.

The most vibrant part of my Ph.D. life was friends; without their love, affection and pampering, I would probably have finished my Ph.D. much earlier. They are no less important than the research or the degree I was supposed to focus at. One of the most important events during my Ph.D. life was the friendship with Saikat Palit, who was an M.Stat. student during my first and second year of Ph.D. The friendship went well beyond the ISI campus, continues to provide constant fuel for a dangerously dynamic life.

I would like to convey my deepest love, respect and gratitude for Dalu Jacob

who was also a Ph.D. student in ISI, Chennai. She made my life beautiful and colorful- beyond my imagination. I wish I could have spent more time with her at ISI Chennai.

I would also like to thank some of the M.Stat. students, particularly- Khyati Sharma, Anwesha Bhattacharyya, Tamoghna Halder, Ananya Nath, Animesh Mitra, Prantik Bera, Sanhita Sengupta, Manjima Bhattacharya, Kritika Gangnani, Soumajit Biswas, Samiran Kundu, Sukanya Bhattacharyya, Abhishek Mandol, Chirayata Kushari, Sayan Mukherjee, Kunal Das and Sambuddha Sekhar for their support and friendship. Some of the research fellows in ISI Chennai, like Aishwarya Ghosh, Sreelakshmi Narayanan and Suparna Biswas, were also instrumental for making it a wonderful place.

Last, but by no means least, I would like to thank my father Sri Debkumar Chakrabarti, my mother Ms. Sukla Chakraborty, my elder sister Ms. Anwesha Chakrabarti, grand mother Srimati Umarani Chakrabarti and late grandfather Sri Harendrakumar Chakrabarti- they are and will be the most important people in my life. I also thank my cousin Mr. Mainak Sadhukhan for his support and well wishes in all my pursuits.

I apologize if I have forgotten anyone here. I thank all for providing the support and help that I very much needed. For any errors or inadequacies that may remain in this work, the responsibility is entirely my own.

Contents

1 Introduction **12**

1.1 Asynchronicity 13

1.2 High-frequency data 16

1.3 Why copula? 16

1.4 Daily (co)volatility 17

1.5 High-Dimensional data 19

1.6 Contribution of the thesis 20

2 Effect of asynchronicity for Elliptical copula **22**

2.1 Introduction 22

2.2 Copula and asynchronicity 24

2.2.1 Pairing Method 24

2.2.2 Estimation of copula Parameter 27

2.3 Simulation study 34

2.4 Real data analysis 40

2.5 Conclusion 41

2.6 Proofs of the theorems 41

2.6.1 Proof of Theorem 3 41

2.6.2 Proof of Theorem 4 44

2.6.3 Proof of Theorem 5 47

3 Extension for more general copula 52

3.1 Dependence measures and relation with copula: 53

3.2 Asynchronicity and effect on Kendall's tau 55

3.2.1 Assumptions 57

3.2.2 Main Results: 59

3.2.3 Positive and negative connection 64

3.3 Simulation 65

3.4 Proofs of the Lemmas and Theorems 67

3.4.1 Proof of Lemma 6: 67

3.4.2 Proof of Lemma 8: 67

3.4.3 Proof of Theorem 10: 69

3.4.4 Proof of Theorem 11: 70

3.4.5 Proof of Theorem 12 73

4 Spectral Distribution of Hayashi-Yoshida Estimator 76

4.1 Introduction 76

4.2 Integrated Covariance Matrix and Asynchronicity 78

4.3 Spectral Distribution 81

4.4 Spectral Analysis of High-Dimensional Hayashi's Estimator . . . 82

4.5 Simulation 96

4.6 Conclusion and further direction 97

5 SPECTRAL DISTRIBUTION FOR $p/n \to 0$ 101

5.1 Introduction 101

5.2 Spectral Analysis of High-dimensional Hayashi's Estimator for $p/n \to 0$ 103

6 Conclusion **113**

7 Appendix **116**

7.1 Introduction to copula: 116

7.1.1 Some examples of copula 117

7.1.1.1 Elliptical copula family 117

7.1.1.2 Archimedean copula family 119

7.1.2 Estimation

120

7.2 Limiting Spectral Analysis of Large Dimensional Random Matrices 122

7.2.1 Stieltjes transform method 128

7.3 Chapter 5 130

7.4 Chapter 5 134

List of Figures

1.1.1 Nonsynchronous transaction times, calculated in seconds, for two stocks Facebook and Apple for first six ticks on a particular day 10/05/2017. 14

1.1.2 Epps effect- correlation decreases to 0 as sampling intervals (k) are getting smaller. Synchronization is done by previous tick sampling technique. x-axis represents sampling intervals and y-axis represents (realized) covariance between two stocks Google and Microsoft on 6th January 2006 [13]. 15

1.5.1 Plot of true (dotted line) and sample (solid line) eigenvalues. . . 20

2.2.1 Two consecutive pairs represented by the arrows. 26

2.2.2 Two consecutive pairs represented by the arrows. 26

2.2.3 Two consecutive pairs of log-returns are $(X_{t(k_{i-1}^1)}, Y_{t(k_{i-1}^2)})$ and $(X_{t(k_i^1)}, Y_{t(k_i^2)})$ with their corresponding transaction times $(t(k_{i-1}^1), t(k_{i-1}^2))$, $(t(k_i^1), t(k_i^2))$ 28

2.2.4 Two consecutive pairs of log-returns are $(X_{t(k_i^1)}, Y_{t(k_i^2)})$ and $(X_{t(k_{i-1}^1)}, Y_{t(k_{i-1}^2)})$ with their corresponding transaction times $(t(k_i^1), t(k_i^2))$, $(t(k_{i+1}^1), t(k_{i+1}^2))$ 28

2.3.1 Realizations (stock price) from one simulation with copula parameter -0.4. 34

2.3.2 Box plots for two cases $\rho = -0.4$ (left) and $\rho = 0.8$ (right); the yellow plot is corresponding to the corrected estimate and the red plot corresponding to uncorrected estimates. 35

2.3.3 Box plot for $\rho = 0.8$; The yellow plot is for the corrected estimates of 100 simulations with sample size 500. The red and green plots correspond to the estimates from refresh time sampling and previous tick sampling respectively. 36

2.3.4 Boxplot for proposed and TSCV estimates. 36

2.3.5 Normal QQ plot of $log(1 + \frac{\hat{\theta}}{\gamma})$ from 100 simulations. 38

2.3.6 Figure illustrating how change in λ_2 affects γ. λ_1 is fixed at 5. Increasing value of $\lambda_2$is plotted at x axis. The y-axis represents values of $1/\gamma$. 39

2.4.1 Facebook (black) and Apple (gray) data after subtracting the mean for two consecutive days 10.05.2017 and 11.05.2017. 40

2.6.1 Two situations depending on the position of $t(k_i^2)$ 47

2.6.2 An illustration of the first case. 48

2.6.3 A case with two perspectives. 50

3.2.1 Two independent observations (pairs) of two different configurations. 56

3.2.2 Two independent observations of a particular configuration. . . 57

3.2.3 The region where signs of true and calculated Kendall's Tau are opposite. 62

4.2.1 Positions of arrivals and the interarrivals of two assets- Asset a and Asset b. 80

4.5.1 Plot for empirical cumulative distribution functions and theoretical cumulative distribution functions for Hayashi Yoshida estimator. 97

4.5.2 The histogram of eigenvalues for 60 stocks and sample size 600 (Left) and 1000 (Right). 98

4.5.3 Kernel density plot of the empirical cumulative distribution function of eigenvalues for 50 stocks with sample size 1000. 99

List of Tables

2.1 Mean and sd of estimators from 100 simulations for different ρ and sample size n. The standard (uncorrected) estimators (with previous tick and refresh time synchronization) and the corrected estimator, prescribed in this chapter, are reported. 37

2.2 Calculated mean-square-error for the results in Table2.3 37

2.3 Simulation for t copula for different marginals with $\rho = -0.4$. The standard estimates with refresh time sampling (uncorrected) and the corrected estimates are reported. 39

2.4 Copula estimation for the joint distribution of Apple and Facebook data. 41

3.1 Simulation result for Clayton, Gumbel and Normal copula is presented. In all the cases the marginals taken to be Normal with mean 0 and variance 1. The uncorrected estimates correspond to refresh time synchronization. Sample size is 1000 in all the cases. 66

7.1 Different Archimedean copulas with the corresponding parameter range. 120

7.2 Different Archimedean copulas with the generators. 120

Chapter 1

Introduction

In this thesis, we are concerned about some modeling issues of intraday financial data. Univariate market risk models have been thoroughly investigated for intraday data [1]. However, univariate modeling (that considers one stock at a time), although important for addressing certain kinds of problems, cannot be enough to unveil the nature and dynamics of financial market because it does not take into account the interdependence of different financial instruments. Correlation dynamics models, therefore, have become an important aspect of theory and practice in finance. Correlation trading, which is a trading activity to exploit the changes in dependence structure of financial assets, and correlation risk that captures the exposure to losses due to changes in correlation, have attracted the attention of many practitioners [2] . Needless to say, poor models of dependence structure may lead to an unexpected collapse of the market. Correlation is also of paramount interest in a range of practical scenarios. For example, basket options are widely used although accurate pricing of the basket option is challenging. The primary reason is that they are cheaper to use for portfolio insurance. The cost-saving relies on the dependence structure between

the assets [3] . In the actuarial world also, Monte Carlo-based approaches to joint modelling of risks, like *dynamic financial analysis*, rely heavily on the dependence structure [4] . The choice of model and correlation has significant impact on the tail of the loss distribution and measures of extreme risks[5] [6]. Short term stock price prediction based on limit order book dynamics and high frequency data is discussed in [7].

This thesis is organized as follows. In the rest of the chapter 1, we introduce the basic notions and relevant existing results. In chapter 2, we will discuss the problem of estimation of Elliptical copula in presence of asynchronicity . In chapter 3, we will extend the results for a general class of copulas. Chapter 4 and 5 will deal with determining limiting spectral distribution for Hayashi-Yoshida estimator for high-dimension.

1.1 Asynchronicity

It follows from the above discussion that we need accurate multivariate modeling and analysis. But in order to perform multivariate analysis, we need multivariate data. This means that we need to have data for all $p(\geq 2)$ variables on n (sufficiently large) time points like days or months. For example, in case of daily financial data we would expect to observe the price of all p stocks on a particular day [1]. We call this kind of data as synchronously observed data. On the other hand, if we don't have observations for one or several variables (or stocks) on a particular time point then we call it nonsynchronous /asynchronous data. An example of such data is intraday stock price data because within a particular day when we observe a transaction in one stock we cannot expect to observe transactions in all other stocks at the exact same time point. In other words, it is not reasonable to assume that financial transactions are made syn-

[1] In this Thesis, we will only focus on stocks and not other kinds of financial instruments

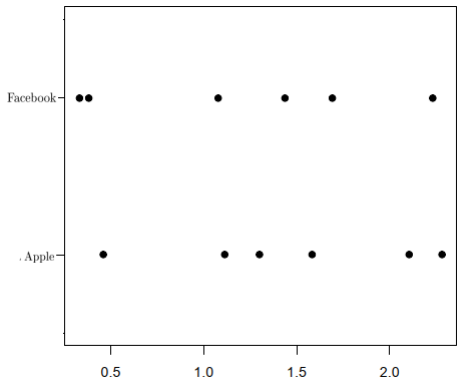


Figure 1.1.1: Nonsynchronous transaction times, calculated in seconds, for two stocks Facebook and Apple for first six ticks on a particular day 10/05/2017.

chronously for all stocks under considerations. In Figure 1.1.1, we have shown the occurrence /arrival times of two stocks within a small time interval. The figure illustrates the problem of asynchronicity.

The effect of this asynchronicity, if not handled properly, can be quite serious. One of such effect, reported by T W Epps [8], is called Epps effect. In 1979, T. W. Epps reported some empirical results showing that stock return correlations decrease as sampling frequency increases. Later the same phenomenon has been reported in several other studies of different stock markets [9] and foreign exchange market [10] [11]. It is also empirically shown that taking into account only the synchronous (or nearly synchronous) price ticks increase the estimated correlation [12]. The Epps effect is a result of asynchronicity of price observations and the lead-lag relation between asset prices [12]. In Figure 1.1.2, the Epps effect is illustrated for two stocks: Google and Microsoft.

This thesis is concerned about the specific effects of asynchronicity on models of dependence structure (based on high-frequency data). As we know, one very widely used method to model multivariate stock price data is through multivariate geometric Brownian motion with a linear correlation which is useful

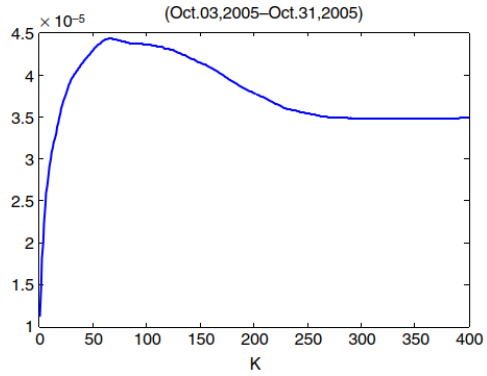


Figure 1.1.2: Epps effect- correlation decreases to 0 as sampling intervals (k) are getting smaller. Synchronization is done by previous tick sampling technique. x-axis represents sampling intervals and y-axis represents (realized) covariance between two stocks Google and Microsoft on 6th January 2006 [13].

for pricing of options under the Black Scholes model. An alternative way to model the dependence structure is through the use of copula. It is important to mention some important works concerned with asynchronicity. For example, the TSCV proposed in [14] which combines two-scale subsampling and previous tick method aimed to remove the bias due to the microstructure noise and asynchronicity. Fan, Li and Yu ([15]) studies TSCV and refresh time under the high-dimensional and high-frequency setting. Ait-Sahalia, Fan and Xiu ([16]) proposes QMLE with a generalized time synchronization method.

In this thesis, we will illustrate the effects of asynchronicity on multivariate modelling through copula and also how it presents challenges for estimation of intraday (co)variation, what would be the possible way out and few additional challenges in case of high-dimensional data. But before delving into the details of these we will briefly discuss:

1. What is "*high-frequency data*"?

2. Why copula is important in finance?

3. What is intraday (co)variation or (co)volatility, how can we capture it?

4. What is special about high-dimensional data analysis, where it is different from multivariate data analysis?

1.2 High-frequency data

High-frequency data refers to time series data collected at a very fine time scale. Its frequency reveals whether the data is collected monthly, weekly, daily, hourly or even finer time scale. In finance, it is the advances in data acquisition and processing techniques that make such data available to us. The highest frequency data available in finance is tick-by-tick data (or trade-by-trade data) in security markets. Since a large number of trades take place every day, the size of such data is huge. These data are irregularly spaced and temporally dependent. The availability of such large data helps to model the financial market more effectively but, at the same time, it poses some challenges. With the help of high-frequency data, in 1980, Merton developed a powerful estimator of (daily)volatility (which we will discuss later in details) [17]. High-frequency data at the one hand help us understand the market microstructure noise and introduce serious bias in volatility estimation on the other. In fact, the non-synchronous nature of intraday financial data is precisely the consequence of high-frequency data.

1.3 Why copula?

Modelling the dependence structure through copula has proved its merit over simple product moment correlation because correlation merely measures a

form of co-linearity [18]. Over the past few decades, copula models have become widely popular in the analysis of financial data. One of the many advantages of the copula is the flexibility it offers to model the complex relationship between variables in a rather simple manner. Copula allows us to model the marginal distributions as necessary and takes care of the dependence structure separately. It is also one of the most important tools to model the probability of extremely large negative (positive) return on one asset given that the other asset yielded an extremely large negative (positive) return- commonly known as tail dependence [19] . For this reason, copula is also a useful tool for modelling the joint distribution between default times. Therefore it is important for pricing Credit Spread Basket, *Credit Debt Obligation*, *First to Default*, *N-to Default* and other Credit derivative baskets [20]. It also helps to correlate the systematic risk to idiosyncratic risk [20]. Copula can also play a key role to estimate dynamic daily dependence [21]. On the other hand, realized copula calculated daily gives us a time series of copula parameter that helps us to capture time-varying dependency [22]. Recently a Mixed-copula VaR model has been proposed to accurately measure the portfolio risk and a novel investment strategy was developed [23]. All these studies show the power and usefulness of copula to model multivariate financial data.

1.4 Daily (co)volatility

It is well known that the (daily) returns of financial assets are hard, in fact almost impossible, to forecast. The volatility of the returns is comparatively easier to predict and leads to various models in financial econometrics. However, the conditional variance is not directly observable and has to be estimated. In order to do so, the ARCH and GARCH family of models and stochastic volatility models are extensively used in financial literature. But these models

cannot satisfactorily capture all the stylized facts of financial time series such as low but slowly decreasing autocorrelations in the squared returns or excess kurtosis of returns [24] [25] [26] [27]. The availability of high-frequency data has made it possible to estimate the volatility of intraday data accurately by realized volatility which is defined as the sum of squared realizations.

Suppose that on date t, the log-price of a given asset follows a continuous time diffusion process:

$$dp(t+\tau) = \mu(t+\tau)d\tau + \sigma(t+\tau)dW(t+\tau), \quad 0 \leq \tau \leq 1,$$

where $p(t+\tau)$ is the log-price at time $t+\tau$, $\mu(t+\tau)$ is the drift, $\sigma(t+\tau)$ is the instantaneous volatility and $W(t+\tau)$ (independent of $\sigma(t+\tau)$) is the standard Brownian motion. The independence of $\sigma(.)$ and $W(.)$ ensures that there is no leverage effect.

Then, given the sigma algebra generated by the sample paths of $\mu(t+\tau-1)$ and $\sigma(t+\tau-1)$, denoted as $\mathcal{F}_t$, the daily return $r(t) = p(t) - p(t-1)$ has the following Gaussian distribution:

$$r_t|\mathcal{F}_t \sim N(\int \mu(t+\tau-1)d\tau, \int_0^1 \sigma^2(t+\tau-1)d\tau).$$

The variance term $\int_0^1 \sigma^2(t+\tau-1)d\tau$ is called Integrated variance. It is a measure of volatility within a day.

Define,

$$RV_t = \sum_{i=1}^{n_t} r_{t,i}^2\,,$$

where n_t is the number of returns of the stock observed on day t. It can be showed that RV_t is a consistent estimator of the integrated variance [17]. We call RV_t as the Realized variance. It is worthwhile to note that realized variance is not consistent in the presence of microstructure noise [24]. In this thesis we

will throughout assume that the data is free of noise [2].

Now if we consider a multivariate model, we can similarly define Integrated covariance matrix. In order to perform a multivariate analysis we need synchronously observed data. But from the previous discussion, it is clear that we don't have synchronous data. As we will discuss in Chapter 4, in a multivariate but low dimensional set-up, the estimator, proposed by Hayashi and Yoshida. It is unbiased and has good large sample properties. But for the high-dimensional case, it is no longer a consistent estimator. Chapter 4 and 5 are dedicated to the study of this matrix in high-dimension from the perspective of random matrix theory. In the next two sections, we are going to discuss some challenges of high-dimensional data and some of the preliminary notions from Random matrix theory.

1.5 High-Dimensional data

In multivariate statistical theory, the sample covariance matrix is the most common and undisputed estimator because it is unbiased and good large sample properties with a growing number of observations when the number of variables is fixed. But if the ratio of the number of variables (p) and the number of observations (n) is large, then sample covariance does not behave as expected. It can be showed that if p grows at the same rate as n i.e. $p/n \to \lambda(> 0)$ the sample covariance matrix becomes inconsistent and therefore cannot be relied upon [28]. In Figure 1.5.1 eigenvalues of the population covariance matrix and sample covariance matrix are plotted for different values of p and n where the population covariance matrix is the identity matrix. It is evident that the true and sample spectra differ a lot as the ratio p/n grows. So for high-dimensional

[2] Microstructure noise is a deviation from fundamental value that is a result of the nature of the market under consideration, e.g. bid-ask bounce, the discreteness of price change, latency, and asymmetric information of traders. It makes realized volatility very unstable.

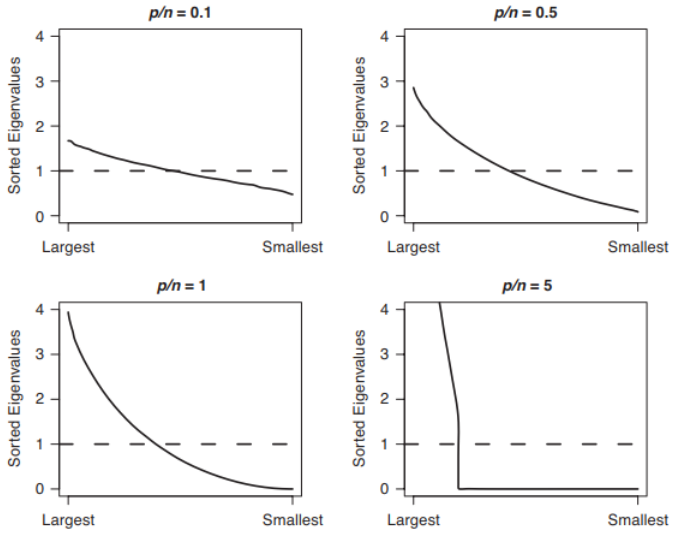


Figure 1.5.1: Plot of true (dotted line) and sample (solid line) eigenvalues.

data ($p/n \to \lambda > 0$) there is a need to find an improved estimator.

Even though the sample eigenvalues are not consistent anymore, the limiting distribution of eigenvalues of the sample covariance matrix and the connection it has with the limiting eigenvalue distribution of population covariance matrix are of importance. Determining the limiting spectral distribution can be very useful to test the underlying assumption of the model. These questions are addressed in random matrix theory. In the appendix we will give a very brief introduction of RMT, some of its basic tools and major results.

1.6 Contribution of the thesis

We show how asynchronicity can affect the statistical analysis of nonsynchronous intraday data. For the Elliptical copula a significant amount of underestimation of interstock dependence can occur if proper correction is not made. We propose a method to *pair* the nonsynchronous observations while keeping a *memory* of

their original arrival times. These *memories* help us to construct a consistent estimator of the copula parameter. We also determine a large sample distribution for the proposed corrected estimate. But mere consistency of the copula parameter does not imply consistency of the empirical copula function. We then show the consistency of copula function. We prove a theorem that determines the amount of loss of intraday data after applying the pairing algorithm when the original arrival processes are independent Poisson. Then we illustrate the effect of asynchronicity for a more general copula. To develop the appropriate correction factor, we focus on Kendall's Tau. Although establishing the exact relation between true and uncorrected Kendall's Tau is challenging, however, under some assumptions we prove some important theorems and through simulations and heuristic derivation, based on those theorems we propose a correction factor. We then show the superiority of this corrected estimate over the uncorrected one by simulation studies on some important Archimedean copulas.

Then we study and establish the limiting spectral distribution of the high-dimensional Hayashi-Yoshida covariance matrix, an estimate of the Integrated covariance matrix in presence of asynchronicity. We establish the limiting spectral distribution of this matrix in high-dimensional setup for both the cases when dimension to sample size ratio converges to a zero or a nonzero number.

Chapter 2

Effect of asynchronicity for Elliptical copula

2.1 Introduction

We have discussed the problems of asynchronicity in the first chapter. To avoid such problems, one can synchronize the data by considering two "close enough" (with respect to time of occurence) observations, one from each stock, as a pair and treat them as if they are synchronously observed. A common practice to avoid such problems is to set a predetermined sampling frequency and a synchronous grid [29]. The price at the time point- just previous to a grid point- is taken to be the price corresponding to that grid point. Details of such methods will be given in section 2.2. In this chapter, we first show that there can be a serious underestimation of the copula even after synchronizing the data with the known synchronization methods. Then we propose a method for the estimation of the copula parameter consistently. The rest of the chapter is

organized as follows. In section 2.2 we deal with the Elliptical copula parameter estimation for nonsynchronous data and prove the main theorem. In section 2.3 and 2.4, we show the results of simulation and real data analysis. Concluding remarks will be given in section 2.5. Section 2.6 contains the proofs of all the theorems presented in section 2.2. It is important to note that although this chapter aims to deal with copula estimation in presence of asynchronicity, its importance is not limited to copula. The results are also valuable when we are interested in estimation of correlation coefficient or Kendall's tau (discussed in the next chapter).

The name 'copula' comes from the precise role it plays. It is the idea of a function that 'couples' a multivariate distribution functions to its marginals. A formal definition is the following.

Definition 1. A d-dimensional distribution function $C(u_1, u_2, ..., u_d) : [0,1]^d \to [0,1]$, where the margin satisfy $C_j(u_j) = C(1, 1, ..., u_j, ...1) = u_j$ for all $u_j \in [0,1]$ and $j = 1, ..., d$, is called a copula.

It is clear from the definition that the copula is a distribution function with uniform margins. The immense importance of copula theory stems from a famous theorem proved by Sklar in 1959 which states that a multivariate distribution function can be uniquely decomposed into its marginals and the dependence structure captured by a copula. This means that the modelling problem can be tackled through two steps. In the first step, the marginal distributions can be modelled without taking the multivariate distribution into account. Then in the second step, the dependence structure can be modelled through the appropriate

choice of the copula.

2.2 Copula and asynchronicity

Suppose there are two stocks and their prices at time $t \in (0, 1)$ are denoted by S_t^1 and S_t^2. By X_t and Y_t we denote the corresponding log-returns. If the prices are assumed to follow the Black-Scholes model then the log returns have a Gaussian distribution and obey the independent increment property. As the stylized facts about financial data suggest that the Gaussian distribution is not an appropriate choice for modelling, a wider class of distributions need to be considered. That is where the true importance of copula lies.

In the Simulation section, the results of the simulation study are reported where the effect of asynchronicity on copula parameter estimation for the Black-Scholes model (and also other popular models commonly used) has been shown. The simulation results display serious underestimation. In this section, the cause of this problem is explained and a remedy is suggested. First, we will present the algorithm to pair the data to make it suitable for computation of the copula parameter estimator. The estimator is presented and consistency is proved in the subsequent section. Finally, an estimator of the copula function is obtained and its consistency is proved.

2.2.1 Pairing Method

The prices of the stocks are observed at random times when transactions take place. It is assumed that the observation times of the two stocks are independent Point processes. Therefore, if we have data of the first stock along with its time of occurrence as (x_i, t_i^1) and that of the second stock as (y_i, t_i^2), then t_i^1s and t_i^2s are independent.

Before fitting a copula model the data has to be paired such that they can be treated as synchronously observed. The conventional synchronizing methods such as *Generalized sampling times*, which includes previous tick sampling and refresh time sampling as special cases, aims to determine the sampling times or grid times with some desirable properties [30] . After determining the grid an observation pair is assigned at each time point in the grid. In this way no memory of the original time point is conserved. We will first introduce the conventional synchronizing method, known as *Generalized sampling times.*

Definition 2. Suppose we have M stocks. t_k^i is the kth arrival time of the ith asset. Then $\{\tau_j\colon\ 1 \leq j \leq n\}$, are called generalized sampling times [31] if

a) $0 = \tau_0 < \tau_1 < ... < \tau_n = T$.

b) $(\tau_{j-1}, \tau_j] \cap \{t_{i,k} : k = 1, ..., n_i\} \neq \varnothing$ for all $i = 1, ..., M$.

c) $\max_{1\leq j\leq n}\delta_j \to 0$ in probability, where $\delta_j = \tau_j - \tau_{j-1}$.

If $\delta_j = \delta$, then it is called *Previous tick sampling.* If $\tau_{j+1} = \max_{1\leq i\leq M} t^i_{N(i,t_j)+1}$, the sampling scheme becomes *refresh time sampling* scheme.

In contrast to the conventional method, we will refrain from uprooting an observation from its origin time point and assigning it to a particular grid point τ_j, for some j. Instead, we are taking into account the actual time of observations. In other words, instead of having a pair like (X_{τ_j}, Y_{τ_j}), we want to have a pair like $(X_{t^1_{k^1_i}}, Y_{t^2_{k^2_i}})$ where $t^1_{k^1_i}$and $t^2_{k^2_i}$ are the times at which the ith pair of stock prices where observed. To emphasise this we call it the 'pairing method' instead of 'synchronizing method'. This pairing method is nothing but a modification over the refresh-time sampling, we can say - it is refresh time sampling with the memory of the original arrival time.

The pairing method, to be followed throughout in this paper, is described

Figure 2.2.1: Two consecutive pairs represented by the arrows.

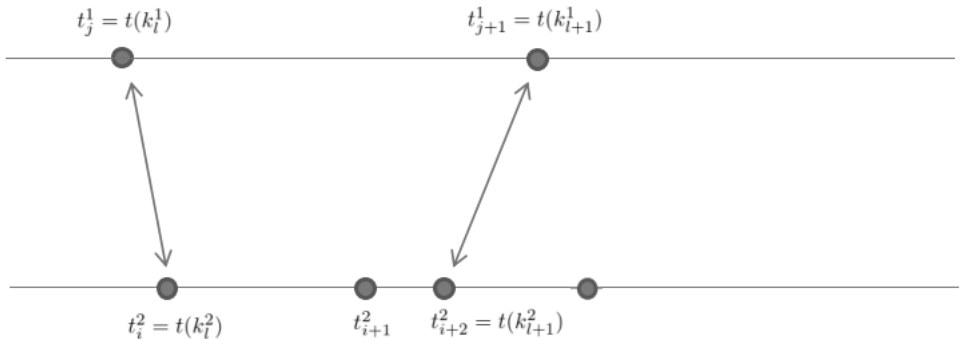


Figure 2.2.2: Two consecutive pairs represented by the arrows.

through the following algorithm:

Algorithm (A_0):

1. Take $i = 1$, $k^1_i = 1$ and $k^2_i = 1$

2. If $t^2_{k^2_i} > t^1_{k^1_i}$ then find $m = \max\{j : t^1_j < t^2_{k^2_i}\}$. The ith pair will be $(X_{t^1_m}, Y_{t^2_{k^2_i}})$. Modify $k^1_i = m$

3. If $t^2_{k^2_i} < t^1_{k^1_i}$ then find $m = \max\{j : t^2_j < t^1_{k^1_i}\}$. The ith pair will be $(X_{t^1_{k^1_i}}, Y_{t^2_m})$. Modify $k^2_i = m$

4. Modify $i = i + 1$. $k^1_i = k^1_i + 1$ and $k^2_i = k^2_i + 1$.

Repeat step 2,3,4.

So now we have paired data. Instead of writing $(X_{t^1_{k^1_i}}, Y_{t^2_{k^2_i}})$ we shall henceforth write $(X_{t(k^1_i)}, Y_{t(k^2_i)})$. Suppose total number of such pairs is n. In the next section, we will see that in order to obtain a consistent estimator for the copula parameter, we need the paired observations, as well as their respective locations $(t(k^1_i), t(k^2_i))$.

In Figure 2.2.1 and Figure 2.2.2, t^1_j and t^2_i are paired together- $(t(k^1_l), t(k^2_l))$.

The figures illustrate how the next pair is going to be chosen using the algorithm. In Figure 2.2.1, $t^1_{j+1} < t^2_{j+1}$. So $t^2_{j+1} = t(k^2_{l+1})$. $t(k^1_{l+1})$ is chosen to be the largest of the arrival times in the first stock that are less than t^2_{i+1}. In Figure 2.2.2, $t^1_{j+1} > t^2_{j+1}$. So $t^1_{j+1} = t(k^1_{l+1})$. $t(k^2_{l+1})$ is chosen to be the largest of the arrival times in the first stock that are less than t^1_{j+1}. The pairs are represented by the arrows.

2.2.2 Estimation of copula Parameter

Once we have paired observations for both the stocks, we can proceed to find out the correlation. But the correlation between the returns $X_{t(k^1_i)} - X_{t(k^1_{i-1})}$ and $Y_{t(k^2_i)} - Y_{t(k^2_{i-1})}$ is dependent on $(t(k^1_i), t(k^2_i))$ and $(t(k^1_{i-1}), t(k^2_{i-1}))$. To see this, first suppose $X_{t(k^1_i)} - X_{t(k^1_{i-1})} = \sum_{i=m}^{l}(X_{t_{i+1}} - X_{t_i})$ for some m and l. Here $\{t_i : i = 1(1)(n_1 + n_2)\}$ is the combined (ordered) time points at which a transaction (in any of the stocks) is noted. Then one of these four configurations is true:

$$\begin{bmatrix} 1. & Y_{t(k^2_i)} - Y_{t(k^2_{i-1})} & = & \sum_{i=m+1}^{l-1}(Y_{t_{i+1}} - Y_{t_i}) \\ 2. & Y_{t(k^2_i)} - Y_{t(k^2_{i-1})} & = & \sum_{i=m-1}^{l-1}(Y_{t_{i+1}} - Y_{t_i}) \\ 3. & Y_{t(k^2_i)} - Y_{t(k^2_{i-1})} & = & \sum_{i=m+1}^{l+1}(Y_{t_{i+1}} - Y_{t_i}) \\ 4. & Y_{t(k^2_i)} - Y_{t(k^2_{i-1})} & = & \sum_{i=m-1}^{l+1}(Y_{t_{i+1}} - Y_{t_i}) \end{bmatrix}. \tag{2.2.1}$$

See Figure 2.2.3 and Figure 2.2.4 for example of two such configurations.

We define a random variable I_i, denoting the overlapping time interval of ith interarrivals corresponding to $X_{t(k^1_i)} - X_{t(k^2_{i-1})}$ and $Y_{t(k^2_i)} - Y_{t(k^2_{i-1})}$

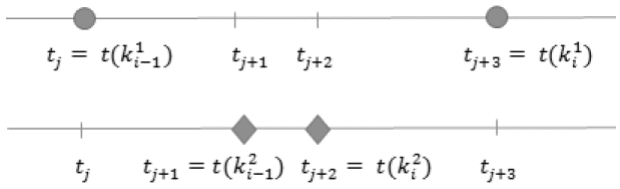


Figure 2.2.3: Two consecutive pairs of log-returns are $(X_{t(k^1_{i-1})}, Y_{t(k^2_{i-1})})$ and $(X_{t(k^1_i)}, Y_{t(k^2_i)})$ with their corresponding transaction times $(t(k^1_{i-1}), t(k^2_{i-1}))$, $(t(k^1_i), t(k^2_i))$

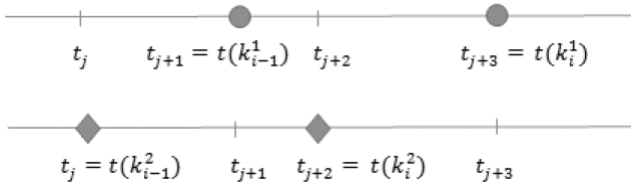


Figure 2.2.4: Two consecutive pairs of log-returns are $(X_{t(k^1_i)}, Y_{t(k^2_i)})$ and $(X_{t(k^1_{i-1})}, Y_{t(k^2_{i-1})})$ with their corresponding transaction times $(t(k^1_i), t(k^2_i))$, $(t(k^1_{i+1}), t(k^2_{i+1}))$

$$I_i = \begin{cases} t(k_i^2) - t(k_{i-1}^2) & \text{if } Y_{t(k_i^2)} - Y_{t(k_{i-1}^1)} = \sum_{i=m+1}^{l-1}(Y_{t_{i+1}} - Y_{t_i}) \\ t(k_i^2) - t(k_{i-1}^1) & \text{if } Y_{t(k_i^2)} - Y_{t(k_{i-1}^1)} = \sum_{i=m-1}^{l-1}(Y_{t_{i+1}} - Y_{t_i}) \\ t(k_i^1) - t(k_{i-1}^2) & \text{if } Y_{t(k_i^2)} - Y_{t(k_{i-1}^1)} = \sum_{i=m+1}^{l+1}(Y_{t_{i+1}} - Y_{t_i}) \\ t(k_i^1) - t(k_{i-1}^1) & \text{if } Y_{t(k_i^2)} - Y_{t(k_{i-1}^1)} = \sum_{i=m-1}^{l+1}(Y_{t_{i+1}} - Y_{t_i}). \end{cases}$$

For Figure 2.2.3, $I_i = t(k_i^2) - t(k_{i-1}^2)$ and for Figure 2.2.4, $I_i = t(k_i^2) - t(k_{i-1}^1)$. With the following example we will show that the correlation between $X_{t(k_i^1)} - X_{t(k_{i-1}^1)}$ and $Y_{t(k_i^2)} - Y_{t(k_{i-1}^2)}$ is dependent on the overlapping (I_i) and nonoverlapping region.

Suppose $(X_{t(k_{i-1}^1)}, Y_{t(k_{i-1}^2)})$ and $(X_{t(k_i^1)}, Y_{t(k_i^2)})$ are two consecutive pairs of log-prices with their corresponding transaction times $(t(k_{i-1}^1), t(k_{i-1}^2))$, $(t(k_i^1), t(k_i^2))$. We denote (X, Y) as a bivariate rv with mean 0 , variances 1 and correlation ρ. (X, Y) is independent of the arrival processes. Here we discuss two cases with example. Figure 2.2.3 and Figure 2.2.4 illustrate two situations where two consecutive pairs of log-prices are $(X_{t(k_i^1)}, Y_{t(k_{i-1}^2)})$ and $(X_{t(k_i^1)}, Y_{t(k_i^2)})$ with their corresponding transaction times- $(t(k_{i-1}^1), t(k_{i-1}^2))$, $(t(k_i^1), t(k_i^2))$. First consider Figure 2.2.3. Define $\mathbb{T} = \{t_i : i = 1(1)\bar{n}\}$ with $\bar{n} = n_1 + n_2$, where n_1, n_2 are the number of observations in the two stocks. The conditional expectation:

$$
\begin{aligned}
&E[(X_{t(k_i^1)} - X_{t(k_{i-1}^1)})(Y_{t(k_i^2)} - Y_{t(k_{i-1}^2)})|\mathbb{T}] \\
&= E[(X_{t(k_i^1)} - X_{t(k_{i-1}^1)})(Y_{t(k_i^2)} - Y_{t(k_{i-1}^2)})|\mathbb{T},\ (t(k_{i-1}^1), t(k_{i-1}^2), t(k_i^1), t(k_i^2)) = (t_j, t_{j+1}, t_{j+3}, t_{j+2})] \\
&= E[(X_{t_{j+3}} - X_{t_{j+2}} + X_{t_{j+2}} - X_{t_{j+1}} + X_{t_{j+1}} - X_{t_j})(Y_{t_{j+2}} - Y_{t_{j+1}})|(t_j, t_{j+1}, t_{j+3}, t_{j+2})] \\
&= E[(X_{t_{j+2}} - X_{t_{j+1}})(Y_{t_{j+2}} - Y_{t_{j+1}})|(t_j, t_{j+1}, t_{j+3}, t_{j+2})] \\
&= E(X\sqrt{t_{j+2} - t_{j+1}}.Y\sqrt{t_{j+2} - t_{j+1}}|(t_j, t_{j+1}, t_{j+3}, t_{j+2})) \\
&= (t_{j+2} - t_{j+1})E(XY) \\
&= (t_{j+2} - t_{j+1})\rho.
\end{aligned}
$$

Thus, $E[(X_{t(k_i^1)} - X_{t(k_{i-1}^1)})(Y_{t(k_i^2)} - Y_{t(k_{i-1}^2)})] = \rho E(t(k_i^2) - t(k_{i-1}^2))$.

$$
\begin{aligned}
\text{Also,}\quad E[(X_{t(k_i^1)} - X_{t(k_{i-1}^1)})^2|\mathbb{T}] &= E(X_{t_{i+3}} - X_{t_i}|t_j, t_{j+3})^2 \\
&\quad E(X_{t_{j+3}} - X_{t_j}|t_j, t_{j+3})^2 \\
&= (t_{j+3} - t_j)EX^2 \\
&= t_{j+3} - t_j,
\end{aligned}
$$

and hence, $E(X_{t(k_i^1)} - X_{t(k_{i-1}^1)})^2 = E(t(k_i^1) - t(k_{i-1}^1))$. Similarly, $E(Y_{t(k_i^2)} - Y_{t(k_{i-1}^2)})^2 = E(t(k_i^2) - t(k_{i-1}^2))$. Thus for the case of Figure 2.2.3,

$$
Cor[(X_{t(k_i^1)} - X_{t(k_{i-1}^1)}), (Y_{t(k_i^2)} - Y_{t(k_{i-1}^2)})] = \frac{\rho E(t(k_i^2) - t(k_{i-1}^2))}{\sqrt{E(t(k_i^1) - t(k_{i-1}^1))}.\sqrt{E(t(k_i^2) - t(k_{i-1}^2))}}.
$$

Taking intuition from these examples we will now state our first theorem.

Now we state our assumptions and the main theorem.

We take the following assumptions $\mathcal{A}$:

$\mathcal{A}_1$: The associated copula is an Elliptical copula.

$\mathcal{A}_2$: The log price process follows independent and stationary increment

property.

$\mathcal{A}_3$: The observation times (arrival process) of two stocks are independent Renewal processes and $n \to \infty$ as $n_1, n_2 \to \infty$ where n is number of paired observations.

$\mathcal{A}_4$: Estimation is based on paired data obtained by algorithm $\mathcal{A}_0$.

$\mathcal{A}_5 : P[(Y_{t(k_i^1)}, Y_{t(k_i^2)}) \in R(y,\delta)] = O(\frac{\delta}{n^\psi})$ with $|\delta| = |\max_i(y_{t(k_i^1)} - y_{t(k_i^2)})| < M$ where $y, \psi > 0$, M is a positive number and $R(a,b)$- is the rectangle in $\mathbb{R}^2$ formed by the interval $(b-a, b+a)$ i.e. $R(a,b) = (b-a, b+a) \times (b-a, b+a)$.

Assumption $\mathcal{A}_5$ is reasonable because of the high-frequency nature of the data, it also relies on continuity of the price processes.

Theorem 3. *Under the assumptions $\mathcal{A}_1 - \mathcal{A}_4$,$\hat{\theta}$ defined below is a consistent estimator of the true copula parameter*

$$\hat{\theta} = \hat{\rho}\frac{\sqrt{m_1.m_2}}{m(I)},$$

where for $l = 1, 2$ $m_l = \frac{1}{n}\sum_{i=1}^{n}(t(k_i^l) - t(k_{i-1}^l))$, $m(I) = \frac{1}{n}\sum_{i=1}^{n} I_i$ and $\hat{\rho}$ being the sample mean and sample correlation coefficient based on the pairs.

Also,

$$\sqrt{n}(f(\hat{\theta}) - f(\theta)) \xrightarrow{d} N(0,1),$$

where $f(\theta) = \frac{1}{2}[log(1+\frac{\theta}{\gamma}) - log(1-\frac{\theta}{\gamma})]$ and $f(\hat{\theta}) = \frac{1}{2}[log(1+\frac{\hat{\theta}}{\gamma}) - log(1-\frac{\hat{\theta}}{\gamma})]$ and

$$\gamma = \frac{\sqrt{E(t(k_i^1) - t(k_{i-1}^1))E(t(k_i^2) - t(k_{i-1}^2))}}{E(I_i)}.$$

Proof of Theorem 3 is given in Section 2.7.

According to the theorem in order to get a consistent estimator, we need to multiply the usual sample correlation coefficient, based on the paired observations (by algorithm A_0), by a correction factor. The correction factor is a

function of only the $t_{k_i^1}$ and $t_{k_i^2}$ for $i = 1(1)n$ i.e. it is only dependent on the arrival process and not on the copula.

As we know $F(x, y) = C(F_1(x), F_2(y); \theta)$, asynchronicity does not only affect the estimation of θ, but also estimation of the copula function because x and y are supposed to be observed synchronously. The convergence of $\hat{C}(\hat{F}_1(x), \hat{F}_2(y); \hat{\theta})$, where $\hat{F}_1(.)$ and $\hat{F}_2(.)$ are empirical distribution functions of $F_1(.)$ and $F_2(.)$, needs more than the convergence of $\hat{\theta}$. The next theorem tries to address this concern.

Theorem 4. *Under $\mathcal{A}_1 : \mathcal{A}_5$, the copula $C(\hat{F}_1(x_1), \tilde{F}_2(x_2); \hat{\theta})$,with $\hat{F}_1(.)$ and $\tilde{F}_2(.)$ being the empirical distribution functions of X and Y on the margins of paired data, and $\hat{\theta}$ defined as in Theorem 14, is uniformly convergent for the true copula.*

Proof of Theorem 4 is given in Section 2.7.

Remark: The assumption $\mathcal{A}_5$ can also be replaced by

$\mathcal{A}_5^* : P[(X_{t(k_i^1)}, X_{t(k_i^2)}) \in R(x, \delta)] = O(\frac{\delta}{n^\psi})$ with $|\delta| = |\max_i(x_{t(k_i^1)} - x_{t(k_i^2)})| < M$ where $\psi > 0$ and M is a positive number.

Now as we know that the correction factor in Theorem 16 is a function of the arrival process only, it is worthwhile to express γ in terms of the underlying parameters of the arrival process. In the next theorem, we will try to do so. But the implication of the theorem goes beyond this purpose. Remember that all the synchronization methods we have discussed, have one problem- it results in loss of data. If we look at Figure 1.1.1, it will be evident. The first observations of two stocks will form the first pair and the last observations of two stocks will form the second pairs. Then all four observations of the first stocks, lying in between the first and last observations, will be wasted. So one can ask the

question that with our pairing method (A_0), what proportion of observations (of each stock) will be wasted. This can be answered if we can compare average interarrival length in a stock (for example $E(t_i^1 - t_{i-1}^1)$ for the first stock) and average interarrival length formed by the pairs ($E(t_{k_i}^1 - t_{k_{i-1}}^1)$ for the first stock). Now if we try to express γ in terms of the arrival process then first we have to calculate $E(t_{k_i}^1 - t_{k_{i-1}}^1)$, $E(t_{k_i}^2 - t_{k_{i-1}}^2)$ and $E(I_i)$ separately. So as a by-product of the following theorem, we will get the answer to the previous question.

Theorem 5. *Suppose the two underlying point processes are Poisson processes with parameters λ_1 and λ_2 and $Z_k \sim Beta(1,k)$ for $k > 1$ then,*

$$\gamma = \frac{\sqrt{\eta_1\eta_2/\lambda_1\lambda_2}}{\frac{1}{2}\sum_{n=1}^{\infty} n[(\frac{1}{\lambda_1}+\frac{1}{\lambda_2})(p_n+q_n)]}$$

where

$$p_n = F_{Z_{n+1}}(\frac{\lambda_2}{\lambda_1+\lambda_2}) - F_{Z_n}(\frac{\lambda_2}{\lambda_1+\lambda_2}),$$

$$q_n = F_{Z_{n+1}}(\frac{\lambda_1}{\lambda_1+\lambda_2}) - F_{Z_n}(\frac{\lambda_1}{\lambda_1+\lambda_2})$$

and for i=1,2

$$\eta_i = \sum_{k=1}^{\infty}[\{F_{B(1,k+1)}(1-\frac{\lambda_i}{\lambda_1+\lambda_2}) - F_{B(1,k)}(1-\frac{\lambda_i}{\lambda_1+\lambda_2})\} + k\{F_{B(1,k+1)}(\frac{\lambda_i}{\lambda_1+\lambda_2}) - F_{B(1,k)}(\frac{\lambda_i}{\lambda_1+\lambda_2})\}].$$

Proof of Theorem 5 is given in Section 2.7.

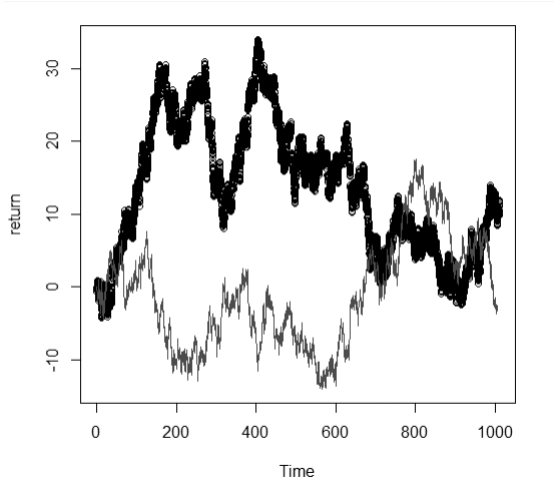


Figure 2.3.1: Realizations (stock price) from one simulation with copula parameter -0.4.

2.3 Simulation study

We simulated data of log returns of two stocks on a synchronized grid of $2n$ random time points (generated by a Poisson Process) by a specific copula function (Figure 2.3.1). Initially, we take a Gaussian copula with dependence parameter ρ. These $2n$ values, generated from the Gaussian copula, are then scaled properly to represent log returns for a particular interarrival. Now from the first stock, we randomly delete n time points and their corresponding returns. The remaining n data points constitute the data for 1st stock. For the 2nd stock, we will keep the time points which were deleted from 1st stock and will delete rest of the time points. These time points, which we decided to keep, along with their corresponding log returns constitute data for 2nd stock. So we are now left with nonsynchronous data for two stocks.

We then use the method, prescribed in the previous section, to estimate the copula parameter. We repeat the process 100 times. The mean and variance of the 100 estimates are reported in the Table 2.3 and 2.4. In Figure 2.3.2, we

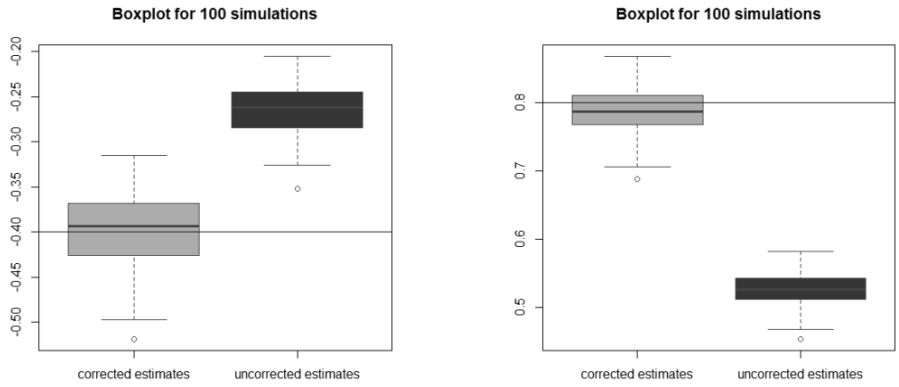


Figure 2.3.2: Box plots for two cases $\rho = -0.4$ (left) and $\rho = 0.8$ (right); the yellow plot is corresponding to the corrected estimate and the red plot corresponding to uncorrected estimates.

show the box plots for two cases- for $\rho = -0.4$ and 0.8 and same $n = 2000$. The boxplot on the left (yellow) is corresponding to the corrected estimate and on the right (red) corresponding to uncorrected estimates [1] (refresh time). The blue line suggests the true parameter.

As discussed earlier TSCV estimator [15], though aims to remove the bias due to microstructure noise, also tries to mitigate undesirable effects of asynchronicity. Therefore we have to compare our proposed estimator with the TSCV estimator. It turns out that although on average TSCV estimator performs better than the uncorrected (refresh time) estimate, it is also biased significantly. Moreover it has a huge variance compared to other estimates. In Figure 2.3.4 we showed the boxplot of the proposed and TSCV estimates.

From the table we see that both previous tick and refresh time sampling fails to capture the magnitude true dependence. In fact previous tick method is the worst choice for synchronization.

[1] Throughout this thesis we refer the estimate with simple refresh time sampling as "uncorrected estimates".

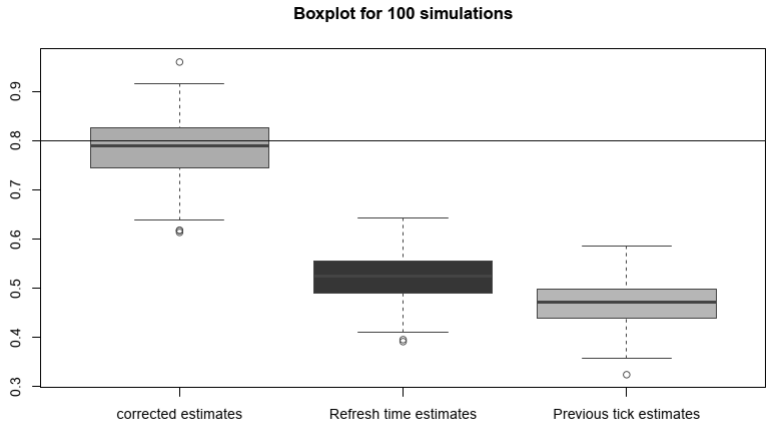


Figure 2.3.3: Box plot for $\rho = 0.8$; The yellow plot is for the corrected estimates of 100 simulations with sample size 500. The red and green plots correspond to the estimates from refresh time sampling and previous tick sampling respectively.

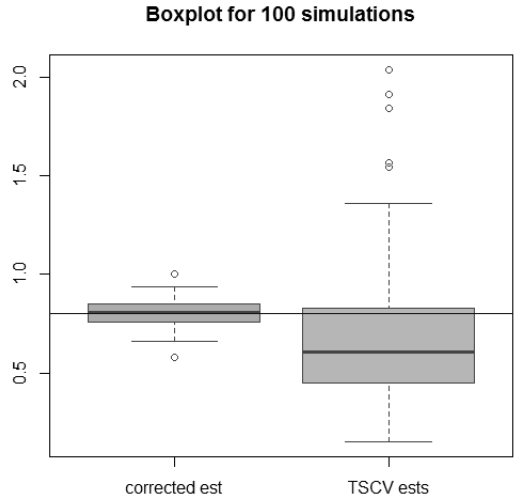


Figure 2.3.4: Boxplot for proposed and TSCV estimates.

ρ	n	Estimate from previous tick sampling	Estimates from refresh time sampling	Corrected estimate
-0.4	800	-0.2316	-0.2680	-0.4022
		(0.049)	(0.046)	(0.069)
0.2	800	0.1205	0.1237	0.1853
		(0.049)	(0.045)	(0.067)
0.8	800	0.4675	0.5255	0.7885
		(0.038)	(0.035)	(0.051)
-0.4	2000	-0.2325	-0.2682	-0.4022
		(0.034)	(0.025)	(0.039)
0.2	2000	0.1115	0.1274	0.1911
		(0.027)	(0.031)	(0.046)
0.8	2000	0.4613	0.5258	0.7888
		(0.028)	(0.019)	(0.029)
-0.4	5000	-0.2289	-0.2637	-0.3956
		(0.018)	(0.017)	(0.026)
0.2	5000	0.1165	0.1357	0.2036
		(0.019)	(0.015)	(0.023)
0.8	5000	0.4582	0.5224	0.7844
		(0.016)	(0.013)	(0.019)

Table 2.1: Mean and sd of estimators from 100 simulations for different ρ and sample size n. The standard (uncorrected) estimators (with previous tick and refresh time synchronization) and the corrected estimator, prescribed in this chapter, are reported.

ρ	n	MSE (previous tick)	MSE (refresh time)	MSE (corrected)
-0.4	800	0.0307	0.0195	0.0047
0.2	800	0.0087	0.0078	0.0047
0.8	800	0.112	0.0765	0.0027
-0.4	2000	0.029	0.0179	0.0015
0.2	2000	0.0085	0.0062	0.0021
0.8	2000	0.1155	0.0755	0.0009
-0.4	5000	0.0296	0.018	0.0006
0.2	5000	0.007	0.004	0.0003
0.8	5000	0.1117	0.077	0.0006

Table 2.2: Calculated mean-square-error for the results in Table2.3 .

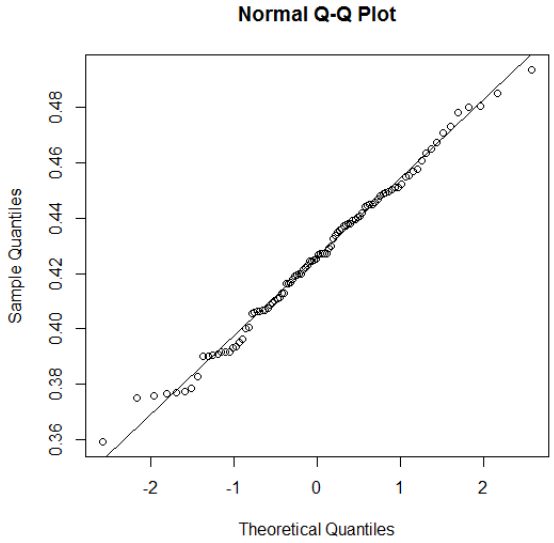


Figure 2.3.5: Normal QQ plot of $log(1 + \frac{\hat{\theta}}{\gamma})$ from 100 simulations.

The qqplot (see Figure 2.3.5) from 100 simulations of the estimates for parameter 0.8 suggests that the asymptotic distribution is Gaussian (as shown in Theorem 20).

We carried out the same analysis with the t copula, with different marginal distributions with different degrees of freedom, which is a more realistic scenario for intraday financial data. The result is similar i.e. not only that our prescribed correction gives a good estimate but also the uncorrected method (refresh time) returns a biased estimate and the bias is significant. The result of 100 simulations with parameter -0.4 is summarized in Table 2.3.

The relation between arrival process parameter and γ is illustrated in Figure 2.3.6. We keep λ_1 fixed at 5 and change the γ_2 to see how the γ is related with the parameter of poisson processes. We see that the value of γ is highest when both the λ's are same ($= 5$). The more λ_2 departs from λ_1 smaller the value of γ becomes.

t copula (df 8) marginals	uncorrected mean	estimate sd	corrected mean	estimate sd
$(t(5), t(7))$	-0.2623	0.036	-0.3932	0.054
$(N(0,2),\ N(0,4))$	-0.264	0.038	-0.3936	0.055
$(t(4),\ N(0,3))$	-0.2532	0.039	-0.3801	0.059

Table 2.3: Simulation for t copula for different marginals with $\rho = -0.4$. The standard estimates with refresh time sampling (uncorrected) and the corrected estimates are reported.

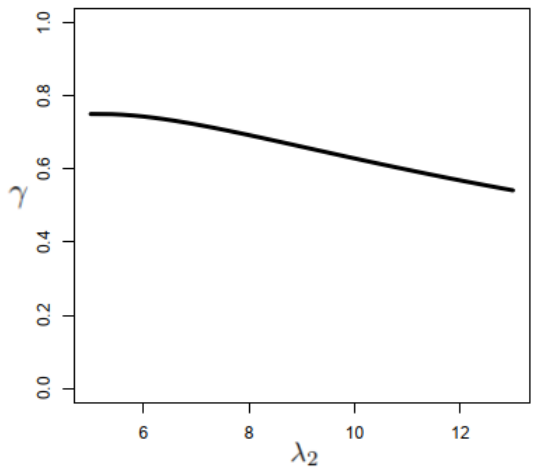


Figure 2.3.6: Figure illustrating how change in λ_2 affects γ. λ_1 is fixed at 5. Increasing value of $\lambda_2$is plotted at x axis. The y-axis represents values of $1/\gamma$.

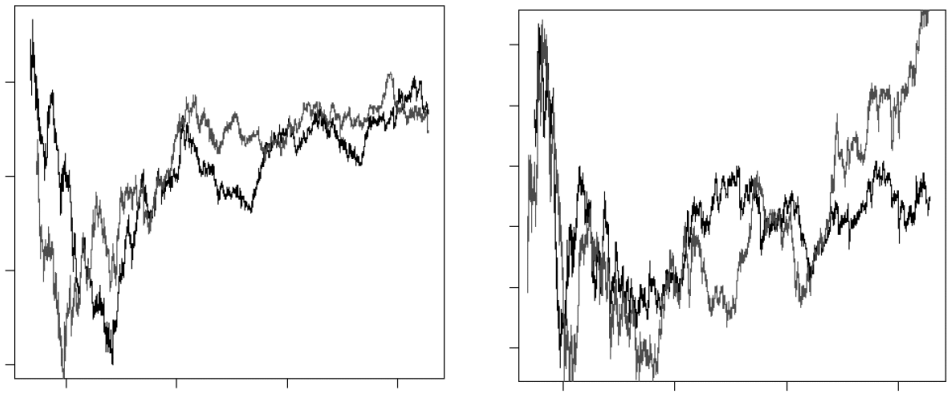


Figure 2.4.1: Facebook (black) and Apple (gray) data after subtracting the mean for two consecutive days 10.05.2017 and 11.05.2017.

2.4 Real data analysis

As shown in Section 2.3, the unbiased estimate can be very difficult to calculate depending on the choice of copula and the underlying process. So we have analyzed real financial intraday data to see which kind of copula is more likely to be encountered in practice. We use AIC to compare and select the best copula. In many of the analysis, we have found that the t-copula is a good choice to model bivariate intraday data.

To see the impact of asynchronicity for the real data we want to see the magnitude of correction to be undertaken. The intraday data for Apple and Facebook stocks (Figure 2.4.1) have been modeled by bivariate t copula for three consecutive days. For all three days both the uncorrected (biased) and the corrected (unbiased) estimates are evaluated (Table 2.4). The amount of correction to be undertaken is reported through the percentage change in values of uncorrected and corrected estimates. We have noticed almost 30 to 35% of data is lost or deleted in the process of pairing them.

We have also performed the same analysis for couple of other stocks and the results we got are very similar. For example, when we consider Amazon

	uncorrected biased estimate	corrected unbiased estimate	percentage change $(\frac{\text{unbiased est.}-\text{biased est}}{\text{biased est.}} \times 100)$
Day 1	0.098	0.141	43.87%
Day 2	0.129	0.186	44.18%
Day 3	0.111	0.159	43.24%

Table 2.4: Copula estimation for the joint distribution of Apple and Facebook data.

and Netflix on three nearly consecutive days, the percentage changes in copula parameter (t copula) are 41.75%, 39.84%, 42.76% respectively.

2.5 Conclusion

These values of both the simulations and real data analysis clearly show that the impact of asynchronicity can be very serious if not tackled properly. We have discussed some of the methods to avoid that. Careful preprocessing of intraday data is necessary in order to model or infer about the underlying realities. But we have to also bear in mind about the impact of the preprocessing undertaken and sometimes modifications of inference methods are required in order to avoid misleading conclusions.

2.6 Proofs of the theorems

2.6.1 Proof of Theorem 3

Proof. The conditional expectation:

$$E[(X_{t(k_i^1)}-X_{t(k_{i-1}^1)})(Y_{t(k_i^2)}-Y_{t(k_{i-1}^2)})|(t(k_i^1),t(k_i^2),t(k_{i-1}^1),t(k_{i-1}^2))] = I_i E(XY).$$

This is a consequence of the assumption A_2 (and illustrated in the examples).

$$E[(X_{t(k_i^1)} - X_{t(k_{i-1}^1)})^2 | t(k_i^1), t(k_{i-1}^1)] = (t(k_i^1) - t(k_{i-1}^1))E(X^2).$$

Similarly,

$$E[(Y_{t(k_i^2)} - Y_{t(k_{i-1}^2)})^2 | t(k_i^2), t(k_{i-1}^2)] = (t(k_i^2) - t(k_{i-1}^2))E(Y^2).$$

So $E(X_{t(k_i^1)} - X_{t(k_{i-1}^1)})^2 = E(t(k_i^1) - t(k_{i-1}^1))E(X^2)$.
and $E(Y_{t(k_i^2)} - Y_{t(k_{i-1}^2)})^2 = E(t(k_i^2) - t(k_{i-1}^2))E(Y^2)$. Therefore

$$\begin{aligned}\mathrm{Cor}((X_{t(k_i^1)} - X_{t(k_{i-1}^1)}), (Y_{t(k_i^2)} - Y_{t(k_{i-1}^2)})) &= \frac{E(XY)E(I_i)}{\sqrt{E(t(k_i^1) - t(k_{i-1}^1))}E(X^2)\sqrt{E(t(k_i^2) - t(k_{i-1}^2))}E(Y^2)} \\ &= \frac{\rho E(I_i)}{\sqrt{E(t(k_i^1) - t(k_{i-1}^1))}\sqrt{E(t(k_i^2) - t(k_{i-1}^2))}}\end{aligned}$$

Therefore the estimate $\hat{\theta}$, defined as

$$\hat{\theta} = \frac{\sqrt{m_1 m_2}}{m(I)}\hat{\rho},$$

where $\hat{\rho} = \hat{Cor}((X_{t(k_i^1)} - X_{t(k_{i-1}^1)}), (Y_{t(k_i^2)} - Y_{t(k_{i-1}^2)}))$, is also consistent estimator of θ because $\hat{\rho}$, $m(t(k_i^1) - t(k_{i-1}^1))$ and $m(I)$ almost surely converge to $Cor((X_{t(k_i^1)} - X_{t(k_{i-1}^1)}), (Y_{t(k_i^2)} - Y_{t(k_{i-1}^2)}))$, $E(t(k_i^1) - t(k_{i-1}^1))$ and $E(I_i)$ respectively.

Let us define the correction factor: $w = \frac{\sqrt{m_1 m_2}}{m(I)}$.

Now

$$m_1 \xrightarrow{P} E(t(k_i^1) - t(k_{i-1}^1))$$

and

$$m_2 \xrightarrow{P} E(t(k_i^1) - t(k_{i-1}^1))$$

and

$$m(I) \xrightarrow{P} E(I_i).$$

Therefore $w \xrightarrow{P} \gamma$, where

$$\gamma = \frac{\sqrt{E(t(k_i^1) - t(k_{i-1}^1))E(t(k_i^2) - t(k_{i-1}^2))}}{E(I_i)}$$

As $\hat{\rho}$ is sample correlation coefficient, we know

$$\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{d} N(0, (1-\rho^2)^2)$$

By Slutsky's theorem we get,

$$\sqrt{n}(w\hat{\rho} - \gamma\rho) \xrightarrow{d} N(0, \gamma^2(1-\rho^2)^2)$$

i.e. (note $\theta = \gamma\rho$)

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \gamma^2(1-\theta^2/\gamma^2)^2) = \gamma(1-\theta^2/\gamma^2)N(0,1).$$

Now we need to stabilize the variance.

Note that,

$$\frac{1}{\gamma(1-\theta^2/\gamma^2)} = \frac{1}{2\gamma}[\frac{1}{1-\theta/\gamma} + \frac{1}{1+\theta/\gamma}].$$

Define, $f(x) = \frac{1}{2}[log(1+\frac{x}{\gamma}) - log(1-\frac{x}{\gamma})]$.

So,

$$f'(x) = \frac{1}{2\gamma}[\frac{1}{1+x/\gamma} + \frac{1}{1-x/\gamma}].$$

Therefore a simple application of delta method implies

$$\sqrt{n}(f(\hat{\theta}) - f(\theta)) \xrightarrow{d} N(0,1).$$

This completes the proof.

□

2.6.2 Proof of Theorem 4

Proof. Note that $F(x,y) = C(F_1(x), F_2(y), \theta)$. So $\hat{C} = \hat{C}(\hat{F}_1(x_{t(k_i^1)}), \hat{F}_2(y_{t(k_i^1)}), \hat{\theta}; i = 1(1)n)$ is a consistent estimator for the copula C. But $y_{t(k_i^1)}$'s are unobserved, where $y_{t(k_i^2)}$'s are actually observed. Let us use the notation $\hat{F}_2(.)$ and $\tilde{F}_2(.)$ for the empirical distribution function of Y based on the observations $\{y_{t(k_i^1)} : i = 1(1)n\}$ and $\{y_{t(k_i^2)} : i = 1(1)n\}$ respectively. Therefore to claim that the estimated copula based on the paired data (observed) is consistent, we have to show that $|\hat{C}(\hat{F}_1, \hat{F}_2, \hat{\theta}) - \hat{C}(\hat{F}_1, \tilde{F}_2, \hat{\theta})| \to 0$ a.s.

Suppose $\delta_i = |y_{t(k_i^1)} - y_{t(k_i^2)}|$. Note that by Assumption 5, $\delta = \max(\delta_i) < M$. Now,

$$
\begin{aligned}
|\tilde{F}_2(y) - \hat{F}_2(y)| &= |\frac{1}{n}\sum_{i=1}^{n} I(Y_{t(k_i^2)} \le y) - \frac{1}{n}\sum_{i=1}^{n} I(Y_{t(k_i^1)} \le y)| \\
&= |\frac{1}{n}\sum_{i=1}^{n} \{I(Y_{t(k_i^2)} \le y) - I(Y_{t(k_i^1)} \le y)\}| \\
&\le \frac{1}{n}\sum_{i=1}^{n} |I(Y_{t(k_i^2)} \le y) - I(Y_{t(k_i^1)} \le y)| \\
&\le \frac{1}{n}\sum_{i=1}^{n} I(Y_{t_{k_i^2}} \in (y-\delta_i, y+\delta_i)) I(Y_{t(k_i^1)} \in (y-\delta_i, y+\delta_i)) \\
&\le \frac{1}{n}\sum_{i=1}^{n} I(Y_{t(k_i^2)} \in (y-\delta, y+\delta)) I(Y_{t(k_i^1)} \in (y-\delta, y+\delta)).
\end{aligned}
$$

We have to show $\frac{1}{n}\sum_{i=1}^{n} I(Y_{t(k_i^2)} \in (y-\delta, y+\delta)) I(Y_{t(k_i^1)} \in (y-\delta, y+\delta)) \to 0$

a.s. using-

$$
\begin{aligned}
&\sum_{n=1}^{\infty} P(|\tilde{F}_2(y) - \hat{F}_2(y)| > \eta) \\
&= \sum_{n=1}^{\infty} P(\frac{1}{n}\sum_{i=1}^{n} |I(Y_{t(k_i^1)} \in (y-\delta, y+\delta))I(Y_{t(k_i^2)} \in (y-\delta, y+\delta))| > \eta) \\
&\leq \sum_{n=1}^{\infty} \sum_{1\leq i,j\leq n} \frac{1}{n^2\eta^2} E(I(Y_{t(k_i^1)} \in (y-\delta, y+\delta)) \times I(Y_{t(k_i^2)} \in (y-\delta, y+\delta)) \times \\
&\qquad\qquad I(Y_{t(k_j^1)} \in (y-\delta, y+\delta)) \times I(Y_{t(k_j^2)} \in (y-\delta, y+\delta)) \\
&\leq \sum_{n=1}^{\infty} \frac{3}{\eta^2 n^2} \sum_{i=1}^{n} E(I(Y_{t(k_i^1)} \in (y-\delta, y+\delta)) \times I(Y_{t(k_i^2)} \in (y-\delta, y+\delta))) \\
&= \sum_{n=1}^{\infty} \frac{3}{\eta^2 n} P[(Y_{t(k_i^1)}, Y_{t(k_i^2)}) \in R(y,d)] \\
&= \sum_{n=1}^{\infty} \frac{3}{\eta^2} O(\frac{1}{n^{1+\psi}}) \\
&< \infty
\end{aligned}
$$

The first inequality is due to Chebyshev's inequality and the last equality is due to A_5. The second inequality is a consequence of asynchronicity as for each i there are at most two j's (the preceding and the next) for which $(X_{t(k_i^1)}, Y_{t(k_i^2)})$ and $(X_{t(k_j^1)}, Y_{t(k_j^2)})$ are dependent.

Hence by Borel Cantelli Lemma, $|\tilde{F}_2(y) - \hat{F}_2(y)| \overset{a.s}{\to} 0$. Again we know $|\hat{F}_2(y) - F_2(y)| \overset{a.s}{\to} 0$. So, $|\tilde{F}_2(y) - F_2(y)| \overset{a.s}{\to} 0$.

Now we have to show that uniform convergence will hold in this case. That is we want to show $\forall \epsilon > 0$, $\sup_y |\tilde{F}_2(y) - F(y)| < \epsilon$. For any given $\epsilon > 0$ we have a finite partition of the real line $-\infty = z_0 \leq z_1 \leq z_2 \leq \ddots \leq z_k = \infty$ such that $F(z_{i+1}^-) - F(z_i) \leq \epsilon$. This can be achieved by taking $z_0 = -\infty$ and $z_{j+1} = \sup\{z : \ F(z) \leq F(z_j) + \epsilon\}$. Then, $F(z_{j+1}) \geq F(z_j) + \epsilon$. Because if $F(z_{j+1}) < F(z_j) + \epsilon$ then by right continuity there exists a $\xi > 0$ such that $F(z_{j+1} + \xi) < F(z_j) + \epsilon$, hence contradicting the definition of z_{j+1}. So between

z_j and z_{j+1}, F jumps at least ϵ. This can happen at most finite number of times, so $k < \infty$. And by our definition of z_{j+1}, we have $F(z_{j+1}-\xi) \leq F(z_j)+\epsilon \; \forall \xi > 0$. Hence $F(z_{i+1}^-) - F(z_i) \leq \epsilon$.
If $y \in [z_i, z_{i+1})$ then

$$
\begin{aligned}
& \tilde{F}(z_i) \leq \tilde{F}(y) \leq \tilde{F}(z_{i+1}^-) \\
& F(z_i) \leq F(y) \leq F(z_{i+1}^-) \\
& \quad \implies \tilde{F}(z_i) - F(z_{i+1}^-) \leq \tilde{F}(y) - F(y) \leq \tilde{F}(z_{i+1}^-) - F(z_i) \\
& \quad \implies \tilde{F}(z_i) - F(z_i) + F(z_i) - F(z_{i+1}^-) \leq \tilde{F}(y) - F(y) \\
& \qquad \leq \tilde{F}(z_{i+1}^-) - F(z_{i+1}^-) + F(z_{i+1}^-) - F(z_i) \quad \forall n > n_i \\
\implies & -2\epsilon \leq \tilde{F}(z_i) - F(z_i) + F(z_i) - F(z_{i+1}^-) \leq \tilde{F}(y) - F(y) \\
& \qquad \leq \tilde{F}(z_{i+1}^-) - F(z_{i+1}^-) + F(z_{i+1}^-) - F(z_i) \\
& \qquad \leq 2\epsilon \quad \forall n > n_i \\
\implies & -2\epsilon \leq \tilde{F}(y) - F(y) \leq 2\epsilon \quad \forall y \in \mathbb{R}, \quad \forall n > \max(n_i) \\
& \quad \implies |\tilde{F}(y) - F(y)| \leq 2\epsilon.
\end{aligned}
$$

Now using properties of copula we can clearly see that

$$
\begin{aligned}
& |C(F_1(x), F_2(y)) - C(\hat{F}_1(x), \tilde{F}_2(y))| \\
& \leq |C(F_1(x), F_2(y)) - C(\hat{F}_1(x), F_2(y))| + |C(\hat{F}_1(x), F_2(y)) - C(\hat{F}_1(x), \tilde{F}_2(y))| \\
& \leq |\hat{F}_1(x) - F_1(x)| + |\tilde{F}_2(y) - F_2(y)|.
\end{aligned}
$$

As both $\hat{F}_1$ and $\tilde{F}_2$ are uniformly convergent to F_1 and F_2 respectively, the result follows.

□

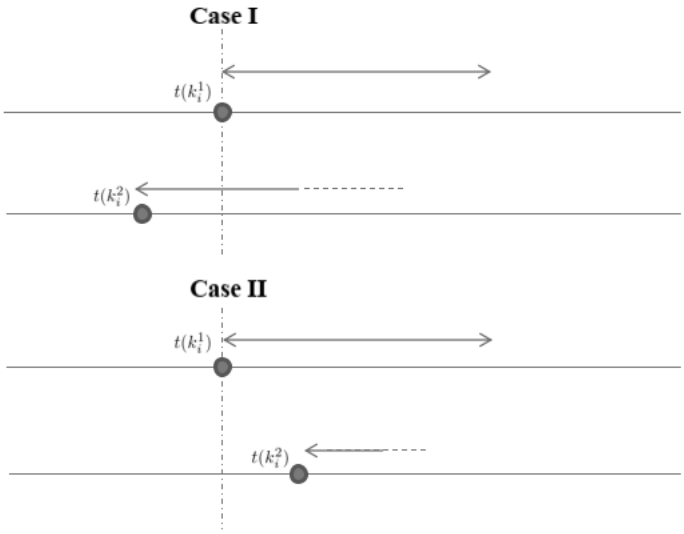


Figure 2.6.1: Two situations depending on the position of $t(k_i^2)$

.

2.6.3 Proof of Theorem 5

Proof. To determine γ, we need to determine $E(t(k_i^1) - t(k_{i-1}^1))$, $E(t(k_i^1) - t(k_{i-1}^1))$, $E(I_i)$.

Let us start with $E(I_i)$. If you fix the point $t(k_i^1)$ then there can two situations depending on the position of $t(k_i^2)$ as illustrated in Figure 2.6.1.

Consider Case 1 first. We are interested in the coming overlap. As illustrated in Figure 2.6.2, define S_1(the first interarrival of the first stock after $t(k_i^1)$) and T_1'(the first interarrival of the 2nd stock after $t(k_i^2)$) and T_1 (the first interarrival of the 2nd stock after $t(k_i^1)$ i.e. if we start observing the process only from $t(k_i^1)$ then the first arrival time for the second stock). As the arrival processes are Poisson processes, distributions of T_1 and T_1' are same (due to memoryless property). Denote all the subsequent interarrivals as $S_i : i = 2, 3, ...$ and

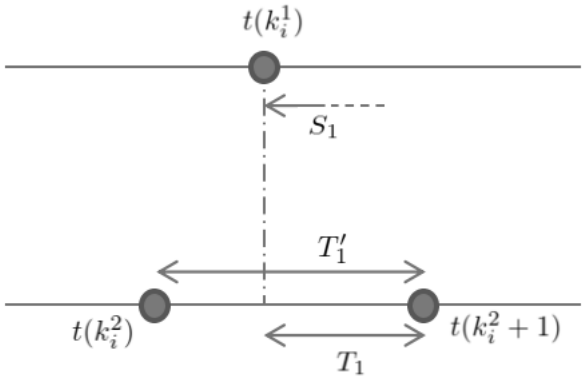


Figure 2.6.2: An illustration of the first case.

$T_i : i = 2, 3, ...$

Now (for case 1),

$$I = \begin{cases} \sum_{i=1}^{k} T_i & \text{if} \quad \text{for } k(\geq 1) \text{ such that } \sum_{i=1}^{k} T_i < S_1 < \sum_{i=1}^{k+1} T_i \\ \sum_{i=1}^{k} S_i & \text{if} \quad \text{for } k(\geq 1) \text{ such that } \sum_{i=1}^{k} S_i < T_1 < \sum_{i=1}^{k+1} S_i \end{cases} \tag{2.6.1}$$

Therefore (for case 1 i.e. actually conditional expectation),

$$E(I) = \sum_{n=1}^{\infty} [E(\sum_{i=1}^{n} T_i) P(\sum_{i=1}^{n} T_i < S_1 < \sum_{i=1}^{n+1} T_i) + E(\sum_{i=1}^{n} S_i) P(\sum_{i=1}^{n} S_i < T_1 < \sum_{i=1}^{n+1} S_i)].$$

$E(\sum_{i=1}^{n} T_i) = nE(T_1) = \frac{n}{\lambda_2}$ and $E(\sum_{i=1}^{n} S_i) = nE(S_1) = \frac{n}{\lambda_1}$.

Define $p_n = P(\sum_{i=1}^{n} T_i < S_1 < \sum_{i=1}^{n+1} T_i)$ and $q_n = P(\sum_{i=1}^{n} S_i < T_1 < \sum_{i=1}^{n+1} S_i)$.

So,

$$E(I) = \sum_{n=1}^{\infty} [\frac{n}{\lambda_1} p_n + \frac{n}{\lambda_2} q_n].$$

$$\begin{aligned} q_n &= P(\sum_{i=1}^{n} S_i < T_1 < \sum_{i=1}^{n+1} S_i) \\ &= P(T_1 < \sum_{i=1}^{n+1} S_i) - P(T_1 < \sum_{i=1}^{n} S_i). \end{aligned}$$

Note that $T_1 \sim exp(\lambda_2) \equiv Gamma(1, \lambda_2)$ and $\sum_{i=1}^{n} S_i \sim Gamma(n, \lambda_1)$.

So

$$\begin{aligned} P(\sum_{i=1}^{n} S_i > T_1) &= P(\frac{\sum_{i=1}^{n} S_i}{\lambda_1} > \frac{(\lambda_1+\lambda_2)T_1}{\lambda_1\lambda_2} - \frac{T_1}{\lambda_2}) \\ &= P(\frac{\sum_{i=1}^{n} S_i}{\lambda_1} + \frac{T_1}{\lambda_2} > \frac{(\lambda_1+\lambda_2)T_1}{\lambda_1\lambda_2}) \\ &= P(\frac{T}{\frac{\sum_{i=1}^{n} S_i}{\lambda_1} + \frac{T_1}{\lambda_2}} < \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}) \\ &= P(\frac{T/\lambda_2}{\frac{\sum_{i=1}^{n} S_i}{\lambda_1} + \frac{T_1}{\lambda_2}} < \frac{\lambda_1}{\lambda_1+\lambda_2}) \\ &= P(Z_n < \frac{\lambda_1}{\lambda_1+\lambda_2}), \end{aligned}$$

where $Z_n = \frac{T/\lambda_2}{\frac{\sum_{i=1}^{n} S_i}{\lambda_1} + \frac{T_1}{\lambda_2}} \sim Beta(1, n)$

So,

$$q_n = F_{Z_{n+1}}(\frac{\lambda_1}{\lambda_1+\lambda_2}) - F_{Z_n}(\frac{\lambda_1}{\lambda_1+\lambda_2}).$$

Similarly,

$$p_n = F_{Z_{n+1}}(\frac{\lambda_2}{\lambda_1+\lambda_2}) - F_{Z_n}(\frac{\lambda_2}{\lambda_1+\lambda_2}).$$

Case 2 is the same if the role of X and $Y$are interchanged. See Figure 2.6.3.

So for Case 2,

$$E(I) = \sum_{n=1}^{\infty}[\frac{n}{\lambda_1}p_n + \frac{n}{\lambda_1}q_n].$$

Now two cases are equally likely, $P(\text{Case 1}) = \frac{1}{2} = P(\text{Case 2})$.

Therefore both the cases combined, $E(I) = \frac{1}{2}\sum_{n=1}^{\infty} n[(\frac{1}{\lambda_1} + \frac{1}{\lambda_2})(p_n + q_n)]$.

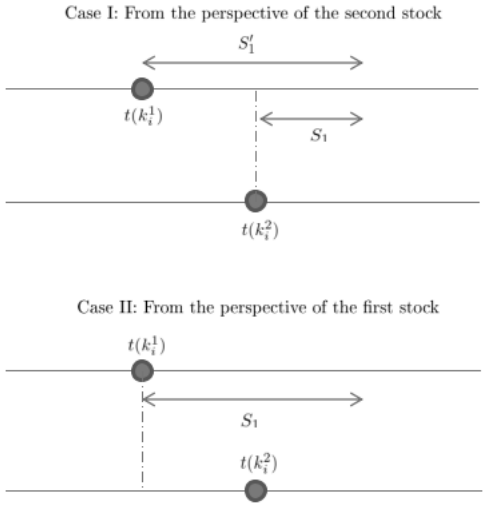


Figure 2.6.3: A case with two perspectives.

Now we have to calculate $E(t(k_i^1) - t(k_{i-1}^1))$.

For case 1:

Define,

$$N = \begin{cases} 1 & \text{for k}(\geq 1) \text{ such that } \sum_{i=1}^{k} T_i < S_1 < \sum_{i=1}^{k+1} T_i \\ \text{k} & \text{for k}(\geq 1) \text{ such that } \sum_{i=1}^{k} S_i < T_1 < \sum_{i=1}^{k+1} S_i \end{cases} \tag{2.6.2}$$

Then,

$E(t(k_i^1) - t(k_{i-1}^1)) = \frac{1}{\lambda_1}\eta_1$, where $E(N) = \eta_1$.

$$\begin{aligned} E(N) &= \sum_{k=1}^{\infty}\{P(\sum_{i=1}^{k} T_i < S_1 < \sum_{i=1}^{k+1} T_i) + kP(\sum_{i=1}^{k} S_i < T_1 < \sum_{i=1}^{k+1} S_i)\} \\ &= \sum_{k=1}^{\infty}[\{P(S_1 < \sum_{i=1}^{k+1} T_i) - P(S_1 < \sum_{i=1}^{k} T_i)\} + k\{P(T_1 < \sum_{i=1}^{k+1} S_i) - P(T_1 < \sum_{i=1}^{k} S_i)\}] \\ &= \sum_{k=1}^{\infty}[\{F_{B(1,k+1)}(\frac{\lambda_2}{\lambda_1+\lambda_2}) - F_{B(1,k)}(\frac{\lambda_2}{\lambda_1+\lambda_2})\} + k\{F_{B(1,k+1)}(\frac{\lambda_1}{\lambda_1+\lambda_2}) - F_{B(1,k)}(\frac{\lambda_1}{\lambda_1+\lambda_2})\}]. \end{aligned}$$

Similarly $E(t(k_i^2) - t(k_{i-1}^2)) = \frac{1}{\lambda_2}\eta_2$ where η_2 is the mean of N' defined below,

$$N' = \begin{cases} 1 \quad \text{if} \quad \text{for } k(\geq 1) \quad \sum_{i=1}^{k} S_i < T_1 < \sum_{i=1}^{k+1} S_i \\ k \quad \text{if} \quad \text{for } k(\geq 1) \quad \sum_{i=1}^{k} T_i < S_1 < \sum_{i=1}^{k+1} T_i. \end{cases}$$

$$\eta_2 = \sum_{k=1}^{\infty}[\{F_{B(1,k+1)}(\frac{\lambda_1}{\lambda_1+\lambda_2}) - F_{B(1,k)}(\frac{\lambda_1}{\lambda_1+\lambda_2})\} + k\{F_{B(1,k+1)}(\frac{\lambda_2}{\lambda_1+\lambda_2}) - F_{B(1,k)}(\frac{\lambda_2}{\lambda_1+\lambda_2})\}].$$

For case 2 the derivation is similar.

So ultimately we get

$$\gamma = \frac{\sqrt{\eta_1\eta_2/\lambda_1\lambda_2}}{\frac{1}{2}\sum_{n=1}^{\infty} n[(\frac{1}{\lambda_1}+\frac{1}{\lambda_2})(p_n+q_n)]}.$$

□

Chapter 3

Extension for more general copula

In this chapter, we will try to extend our result to a more general class of copulas. As the argument, in Chapter 2, is entirely based on the correlation coefficient, it cannot be directly extended to a larger class of copulas. One big drawback of the elliptical copula is that it does not provide the flexibility that copula is famous for. For example, we would find it problematic to model any of the marginal distributions by a non-elliptical distribution, otherwise, the copula parameter (correlation matrix) cannot be directly estimated from the data. The method, which we are going to prescribe here, is a solution to that problem.

Although for a general copula there is no direct relation between Pearson's correlation coefficient and the copula parameter, we saw that there is a relation between Kendall's tau and copula parameter. So we can study how Kendall's tau is affected by asynchronicity. The results in this chapter will enable us to estimate Kendall's tau for intraday financial data.

3.1 Dependence measures and relation with copula:

We discussed that copula completely captures the dependence structure of a multivariate distribution and therefore plays an important role in the study of measuring dependence. Some of the widely known association measures are directly related to copula only (and not to the marginals). Here we discuss some of them.

Kendall's τ, an important measure of association based on concordance, is defined as $\rho_\tau(X,Y) = E(\text{sign}(X-\tilde{X})(Y-\tilde{Y}))$, where $\tilde{X}$ and $\tilde{Y}$ are identical but independent copies of X and Y [32]. The relationship between copula and Kendall's tau is given by-

$$\rho_\tau(X,Y) = 4\int_0^1\int_0^1 C(u_1,u_2)dC(u_1,u_2) - 1. \tag{3.1.1}$$

For the elliptical copulas (with density), a simplified form can be derived [33],

$$\rho_\tau(X,Y) = \frac{2}{\pi}\text{arcsin}.\rho. \tag{3.1.2}$$

If X and Y are random variables with an Archimedean copula C generated by ϕ, then

$$\rho_\tau = 1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)}dt. \tag{3.1.3}$$

Spearman's Rho, another measure of association, is defined as

$$\rho_S(X_1,X_2) = 3[P(X_1-X_2)(Y_1-Y_3)>0] - P[(X_1-X_2)(Y_1-Y_3)<0]).$$

where (Y_1,Y_3) is independent (with same margins as (X_1,X_2)) of (X_1,X_2)

but Y_1 and Y_3 are independent. It can be shown that,

$$\rho_S(X,Y) = 12\int_0^1\int_0^1 C(u,v)dudv - 3.$$

For Gaussian copula $\rho_S = \frac{6}{\pi}\arcsin\frac{\rho}{2}$ [34].

A generalized form of measure of dependence [35] is given by

$$\mathscr{M}_C = \int\int_{I^2} f(u,v,C(u,v))dC(u,v),$$

where f is an appropriate smooth function (must be such that $|\mathscr{M}_C| \leq 1$ and $\mathscr{M}_\Pi = 0$). The cases $f(u,v,C) = 4C - 1$ and $f(u,v,C) = 12uv - 3$ give Kendall's tau and Spearman's Rho respectively.

Tail dependence is another extremely important measure of dependence, especially in context of finance, because it is linked to high simultaneous losses. Tail dependence looks at the concordance in the extreme values of X and Y. The upper tail dependence (λ_U) is defined by-

$$\lambda_U = \lim_{u\to 1} P(Y \leq F_Y^{-1}(u)|X \leq F_X^{-1}(u)).$$

Similarly the lower tail dependence (λ_L) is defined by-

$$\lambda_L = \lim_{u\to 0} P(Y > F_Y^{-1}(u)|X > F_X^{-1}(u)).$$

One of the reasons for the popularity of the copula in finance is that it has a nice and elegant relation with tail dependence as follows,

$$\lambda_U = \lim_{u\to 1}\frac{1-2u+C(u,u)}{1-u}$$

and

$$\lambda_L = \lim_{u\to 0} \frac{C(u,u)}{u}.$$

Thus, as we discussed, these dependence measures contain significant informations about the copula parameter. [1]. In this chapter we will particularly focus on the Kendall's tau. For a large class of copulas there is one-one relation between copula parameter and Kendall's tau. As it is difficult to measure the impact of asynchronicity on copula parameter directly, we will measure the impact on Kendall's tau. In other words, our intention is very much similar to that in the previous chapter- we will try to *correct* the sample Kendall's tau by a correction factor. If we obtain the corrected estimate of Kendall's tau, we can calculate copula parameter through the equation 3.1.1.

3.2 Asynchronicity and effect on Kendall's tau

The definition of Kendall's tau is $E(sign((X-\tilde{X})(Y-\tilde{Y})))$ where $\tilde{X}$ and $\tilde{Y}$ are identical but independent copies of X and Y. The problem with non-synchronous data is that any two independent pairs of returns cannot be taken as identical copies of each other. To see this, consider the Figure 3.2.1, where arrival times of the 1st stock is denoted by triangles and arrival times of the second stock is denoted by circles. After applying the pairing method, suppose the first circle and first triangle represent the location of first pair of prices. Similarly second circle and second triangle represent the next pair. From the figure it is evident that the returns, from these two pairs, are forming an example of second configuration. Similarly, the 3rd and 4th pair constitutes an example of the 4th configuration. So clearly the corresponding returns cannot

[1] Unfortunately correlation coefficient does not have any direct relationship with the copula parameter as it is not invariant under increasing transformation (and copula is invariant under increasing transformation).

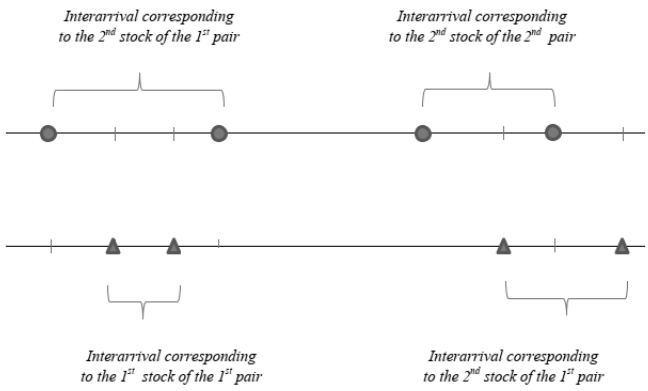


Figure 3.2.1: Two independent observations (pairs) of two different configurations.

be taken as identical. Instead, to have that we have to take two independent pairs of returns of the same configurations. Figure 3.2.2 represents the arrival times of two pairs of same configuration.

As illustrated in Figure 3.2.2, suppose we have two non-overlapping interarrivals u_1 and u_2 for the first stock, denoted by the triangles and two non-overlapping interarrivals $\epsilon_1 + u_1 + \eta_1$ and $\epsilon_2 + u_2 + \eta_2$ for the second stock, denoted by the circles. The log returns corresponding to those interarrivals of the 1st stock is given by $X_1 = X(u_1)$ and $X_2 = X(u_2)$. Similarly the log returns corresponding to the intervals of the 2nd stock is denoted by $Y_1 = Y_1(\epsilon_1 + u_1 + \eta_1) = Y_1(\epsilon_1) + Y_1(u_1) + Y_1(\eta_1)$ (due to independent increment property) and $Y_2 = Y_2(\epsilon_2 + u_2 + \eta_2) = Y_2(\epsilon_2) + Y_2(u_2) + Y_2(\eta_2)$.

In the following section, we will focus on two specific configurations (described in the previous chapter).

Define,

$$A = (X_1(I_1) - X_2(I_2))(Y_1(I_1) - Y_2(I_2))$$

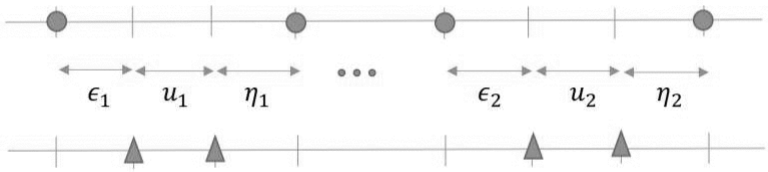


Figure 3.2.2: Two independent observations of a particular configuration.

and

$$B = \begin{cases} (X_1(I_1) - X_2(I_2))(Y_1(I_1^c) - Y_2(I_2^c)) & \text{for 4th configuration} \\ (X_1(I_1^c) - X_2(I_2^c))(Y_1(I_1) - Y_2(I_2)) & \text{for 1st configuration.} \end{cases}$$

where I_i and I_i^c are overlapping and non-overlapping regions respectively of the ith pair. Note that for first and fourth configurations, $E(sign(A))$ gives us true Kendall's tau. We cannot calculate $E(sign(A))$ because $Y_1(I_1)$ and $Y_2(I_2)$ are not observed. Instead, we observe $Y_1(I_1 \cup I_2)$ and $Y_2(I_2 \cup I_2^c)$. Therefore the observed Kendall's Tau is $E(sign(A+B))$. In this chapter, we will try to find out the relation between $E(sign(A))$ and $E(sign(A+B))$.

3.2.1 Assumptions

In order to establish our result, we need some assumptions. Here we will explain the ideas behind the assumptions and will state the assumptions. Suppose X and Y are positively associated random variables. Let (X_1, Y_1) and (X_2, Y_2) be two identical copies of (X, Y). Then, given the information that $Y_1 - Y_2 > 0$, we would expect that $X_1 - X_2$ is more likely to be positive than negative. Intuitively, positive association would also suggest that given the information $Y_1 - Y_2 \in S \subset \mathbb{R}^+$, $X_1 - X_2$ is more likely to be positive. This notion is not

in general captured by Kendall's tau. The first and second assumptions, stated below, try to capture the idea. For each of the following we define (X_1, Y_1) and (X_2, Y_2) are two identical copies of (X, Y), $U = X_1 - X_2$ and $V = Y_1 - Y_2$.

Assumption 1: If

$$P(UV > 0) - P(UV < 0) > 0 \text{ (or } < 0),$$

then for all $M > 0$, the following will also hold:

$$P(U > 0|0 < V < M) - P(U < 0|0 < V < M) \geq 0 \text{ (or } < 0)$$

and

$$P(V > 0|0 < U < M) - P(V < 0|0 < U < M) \geq 0 \text{ (or } < 0).$$

Assumption 2: If

$$P(UV > 0) - P(UV < 0) > 0 \text{ (}or\text{ } < 0),$$

then for all $M > 0$,

$$P(UV > 0| \ |V| > M) - P(UV < 0| \ |V| > M) > 0 \text{ (}or\text{ } > 0)$$

and,

$$P(UV > 0| \ |V| > M) - P(UV < 0| \ |U| > M) > 0 \text{ (}or\text{ } > 0).$$

We will call the next assumption as *Assumption 2a* because we will end up at the same conclusion by two different paths based on either *Assumption 2* or the following assumption (*Assumption 2a*).

Assumption 2a: If

$$P(UV > 0) - P(UV < 0) > 0,$$

then for $M > 0$,

$$\begin{aligned} &P(U > 0|V > M) - P(U < 0|V > M) > \\ &\max(P(U > 0|0 < V < M) - P(U < 0|0 < V < M), \\ &P(U < 0| - M < V < 0) - P(U > 0| - M < V < 0)) \end{aligned}$$

and ,

$$\begin{aligned} &P(V > 0|U > M) - P(V < 0|U > M) > \\ &\max(P(V > 0|0 < U < M) - P(V < 0|0 < U < M), \\ &P(V < 0| - M < U < 0) - P(V > 0| - M < U < 0)) \end{aligned}$$

where $U = X_1 - X_2$, $V = Y_1 - Y_2$ and (X_1, Y_1) and (X_2, Y_2) are two identical copies of (X, Y).

If

$$P(UV > 0) - P(UV) < 0) < 0,$$

then the inequalities are reversed.

3.2.2 Main Results:

Before stating the main theorems, we will first state some Lemmas which

will help us to prove the theorems.

Lemma 6. $E(\text{sign}(A)|\text{sign}(A) \neq \text{sign}(B)) = E(\text{sign}(A))$.

Proof of Lemma 6 is given in Section 3.4.

Lemma 7. $E(\text{sign}(A)|\text{sign}(A) \neq \text{sign}(B), |A| < |B|) = E(\text{sign}(A)||A| < |B|)$

This is a straightforward consequence of Lemma 6 and the independence of $\{\text{sign}(A) \neq \text{sign}(B)\}$ and $\{|A| < |B|\}$.

Lemma 8. *Suppose* $N = \{\text{sign}(A) \neq \text{sign}(B), |A| < |B|\}$. *Then under Assumption 1,*

(a) If $E(\text{sign}(A)) > 0$, *then*

$$E(\text{sign}(A)|N) > 0.$$

(b) If $E(\text{sign}(A)) < 0$, *then*

$$E(\text{sign}(A)|N) < 0.$$

Proof of Lemma 8 will be given in Section 3.4.

Lemma 9. ***(a)*** *For positive association, under Assumption 1 and 2',* $E(\text{sign}(A) \mid |A| > |B|) \geq E(\text{sign}(A) \mid |A| \leq |B|)$.

(b) *For positive association, under Assumption 1 and 2' ,* $E(\text{sign}(A) \mid |A| > |B|) \leq E(\text{sign}(A) \mid |A| \leq |B|)$.

Proof of Lemma 9 is trivial.

Theorem 10. *Under the Assumption 1, for the pairs with 1st and 4th configuration,*

$$|\tilde{\rho_\tau}| > |\rho_\tau|,$$

where ρ_τ *is the Kendall's tau calculated on the paired data with 1st and 4th configurations, i.e.* $\rho_\tau = E(\text{sign}(X_1 - X_2)(Y_1 - Y_2))$, *where* (X_1, Y_1) *and* (X_2, Y_2) *are independent pairs of the same configurations.*

Proof of Theorem 10 is given in Section 3.4.

Now we want to show $\text{sign}(\rho_\tau) = \text{sign}(\hat{\rho}_\tau)$.

We will show this through two alternative ways motivated by two different (but related) assumptions *Assumption 2* and *Assumption 2a.*

Theorem 11. *For the pairs with 1st and 4th configuration,*

$$\rho_\tau = E\text{sign}(A|\text{sign}(A) \neq \text{sign}(B) \& |A| > |B|)P(|A| > |B|),$$

where ρ_τ *is the Kendall's tau calculated on the paired data with 1st and 4th configurations, i.e.* $\rho_\tau = E(\text{sign}(X_1 - X_2)(Y_1 - Y_2))$, *where* (X_1, Y_1) *and* (X_2, Y_2) *are independent pairs of the same configurations.*

Proof of Theorem 11 is given in Section 3.4.

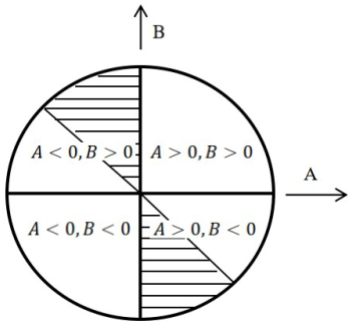


Figure 3.2.3: The region where signs of true and calculated Kendall's Tau are opposite.

Due to the Lemma 7 proved earlier,

$$\rho_\tau = E\text{sign}(A||A| > |B|)P(|A| > |B|).$$

This relation and Assumption 2 together imply that, $\text{sign}(\rho_\tau) = \text{sign}(\hat{\rho}_\tau)$.

An alternative way to show the same is through *Assumption 2a* and the following theorem.

Theorem 12. *Under the assumption 1 and 2a , for the pairs with 1st and 4th configuration,*

$$\text{sign}(\tilde{\rho_\tau}) = \text{sign}(\rho_\tau),$$

where ρ_τ is the Kendall's tau calculated on the paired data with 1st and 4th configurations, i.e. $\rho_\tau = E(\text{sign}(X_1 - X_2)(Y_1 - Y_2))$, where (X_1, Y_1) and (X_2, Y_2) are independent pairs of the same configurations.

Proof of Theorem 12 is in Section 3.4.

As a consequence of Theorem 10, we can write $\tilde{\rho_\tau} = k\rho_\tau$, for some $k > 1$. k cannot be estimated from the available (nonsynchronous) data. Based on our empirical study we recommend to take $k = 2$, (see Table 3.1). An intuitive reason for this value is given below. From the proof of Theorem 10 we have seen

$$\begin{aligned} E(\text{sign}(A)) - E(\text{sign}(A+B)) &= 2E(\text{sign}(A) \mid N)P(N) \\ &= 2[P(A>0, B<0, |A|<|B|) - P(A<0, B>0, |A|<|B|)] \\ &= 2[P(A>0)P(B<0, |A|<|B| \mid A>0) \\ &\quad - P(A<0)P(B>0, |A|<|B| \mid A<0)]. \end{aligned}$$

The region $\{A > 0\}$ and $\{A < 0\}$ are the right and left half circles in Figure 3.2.3. $\{B > 0, |A| < |B|\}$ is one of the four geometrically identical sections of $\{A < 0\}$. Similarly $\{B < 0, |A| < |B|\}$ is one of the four geometrically identical sections of $\{A > 0\}$. If all those sections/slices are probabilistically more or less equal then $P(B < 0, |A| < |B| \mid A > 0) \approx \frac{1}{4} \approx P(A < 0)P(B > 0, |A| < |B| \mid A < 0)$. Which leads to-

$$\begin{aligned} E(\text{sign}(A)) - E(\text{sign}(A+B)) &\approx 2[P(A>0)\frac{1}{4} - P(A<0)\frac{1}{4}] \\ &= \frac{1}{2}[P(A>0) - P(A<0)] \\ &= \frac{1}{2}E(\text{sign}(A)). \end{aligned}$$

With this intuition our corrected estimate is $\hat{\tilde{\rho_\tau}} = 2\hat{\rho_\tau}$ i.e. twice of the uncorrected estimate (with refresh time).

The assumptions we have taken are less artificial than it seems. Due to the Theorem 11, we can write $E(\text{sign}(A)| \, |A| > |B|) = k.E(\text{sign}(A))$, where k (> 1) is a constant. The assumptions ensure that smaller the value of $P(|A| > |B|)$ is the bigger the value of k becomes. To see this intuitively, suppose there is

positive association between two stocks. Now a very low value of $P(|A| > |B|)$ indicates the event with reasonably high values of $|A|$. That means possibly, either $|X_1(u_1) - X_2(u_2)|$ or $|Y_1(u_1) - Y_2(u_2)|$ is high. But as there is positive association if one of them is high with high chance the other one will also be high. Therefore both $X_1(u_1) - X_2(u_2)$ and $Y_1(u_1) - Y_2(u_2)$ would possibly having same sign resulting in $sign(A) = 1$ for vast majority of cases.

Due to this inverse relation between k and $P(|A| > |B|)$, the correction factor introduced above is equivalent of taking $\hat{k} = \frac{1/2}{P(|A|>|B|)}$.

3.2.3 Positive and negative connection

Suppose X and Y are positively associated. And (X_1, Y_1) and (X_2, Y_2) are two identical pairs. Then bigger the value of Y_1 from Y_2, bigger the probability of $\{X_1 > X_2\}$ would expected to be. This good expectation is formalized in the following definition.

Definition 13. X is said to be *positively connected* to Y if $\forall M > 0$,

$$\begin{aligned} P(U > 0|V > M) &\geq P(U > 0|0 < V < M) \\ &\geq P(U > 0| - M < V < 0) \\ &\geq P(U > 0|V < -M). \end{aligned}$$

X is said to be *negatively connected* to Y if $\forall M > 0$, the signs of the above inequalities are reversed.

Note that, the definition is not symmetric for X and Y. X is positively connected to Y does not mean that Y is positively connected to X.

Definition 14. If X is *positively (or negatively) connected* to Y and Y is *positively (or negatively) connected* to X, then we call- there is a *positive (or negative) connection* between X and Y.

It is easy to see that if X and Y have *positive (/negative) connection*, then *assumption 1*, *assumption 2* and *assumption 2'* are satisfied. This is because $X_1 - X_2|Y_1 - Y_2 \stackrel{d}{=} X_2 - X_1|Y_2 - Y_1$.

Therefore this stronger but reasonable assumption unifies the previous assumptions and enable us to state the following theorem.

Theorem 15. *Under the assumption that returns of two stocks have positive (/negative) connection, for the pairs with 1st and 4th configuration,*

$$|\tilde{\rho_\tau}| > |\rho_\tau|$$

and

$$\text{sign}(\tilde{\rho_\tau}) = \text{sign}(\rho_\tau),$$

where ρ_τ is the Kendall's tau calculated on the paired data with 1st and 4th configurations, i.e. $\rho_\tau = E(\text{sign}(X_1 - X_2)(Y_1 - Y_2))$, where (X_1, Y_1) and (X_2, Y_2) are independent pairs of the same configurations.

3.3 Simulation

For the computation of Kendall's tau, we have to simulate pairs with same configuration (otherwise two pairs won't have identical distribution). In Table 3.1 we illustrate the performance of our proposed estimator on simulated data

	Uncorrected	$\hat{\rho}_\tau$	Corrected	$\hat{\rho}_\tau$	Uncorrected	$\hat{\theta}$	Corrected	$\hat{\theta}$
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
CC(4)	0.327	0.016	0.655	0.033	0.977	0.074	3.86	0.574
GC(6)	0.375	0.019	0.751	0.038	1.60	0.049	4.12	0.659
NC(0.8)	0.305	0.021	0.611	0.042	0.461	0.029	0.817	0.039

Table 3.1: Simulation result for Clayton, Gumbel and Normal copula is presented. In all the cases the marginals taken to be Normal with mean 0 and variance 1. The uncorrected estimates correspond to refresh time synchronization. Sample size is 1000 in all the cases.

with 1st and 4th configuration. The performance of uncorrected and corrected estimate of Kendall's tau and the corresponding copula parameter is reported for Clayton copula (CC) with parameter $\theta = 4$ (Kendall's tau 0.67), Gumbel copula (GC) with parameter $\theta = 6$ (Kendall's tau 0.83) and Normal copula (NC) with parameter $\theta = 0.8$ (Kendall's tau 0.59). We note that the corrected estimate is, although, not completely satisfactory but performs much better than the uncorrected estimate. For example, for the Clayton case, the true parameter is 4. The uncorrected and corrected estimates are .977 and 3.86 respectively. For the Gumbel case, the true parameter is 6. The uncorrected and corrected estimates are 1.60 and 4.12 respectively.

We have also tested this against several other copulas with different margins, the results are very similar in nature.

3.4 Proofs of the Lemmas and Theorems

3.4.1 Proof of Lemma 6:

Proof.

$$\begin{aligned}E(\text{sign}(A)) &= E(\text{sign}(A)|\text{sign}(A) \neq \text{sign}(B))P(\text{sign}(A) \neq \text{sign}(B)) \\ &\quad + E(\text{sign}(A)|\text{sign}(A) = \text{sign}(B))P(\text{sign}(A) = \text{sign}(B)) \\ &= E(\text{sign}(A)|\text{sign}(A) \neq \text{sign}(B))\frac{1}{2} + E(\text{sign}(A)|\text{sign}(A) = \text{sign}(B))\frac{1}{2}.\end{aligned}$$

As $Y(\eta_1)+Y(\epsilon_1)-Y(\eta_2)-Y(\epsilon_2)$ is independent of $Y(u_1)-Y(u_2)$ and symmetric around 0, clearly-

$$\begin{aligned}&E(\text{sign}(A)|\text{sign}(A) \neq \text{sign}(B)) \\ &= E(\text{sign}(A)|\text{sign}(Y(u_1) - Y(u_2) \neq \text{sign}(Y(\eta_1) + Y(\epsilon_1) - Y(\eta_2) - Y(\epsilon_2))) \\ &= E(\text{sign}(A)|\text{sign}(Y(u_1) - Y(u_2) = \text{sign}(Y(\eta_1) + Y(\epsilon_1) - Y(\eta_2) - Y(\epsilon_2))) \\ &= E(\text{sign}(A)|\text{sign}(A) = \text{sign}(B)).\end{aligned}$$

So, $E(\text{sign}(A)|\text{sign}(A) \neq \text{sign}(B)) = E(\text{sign}(A))$.

□

3.4.2 Proof of Lemma 8:

Proof. We will only prove the first part. The second part can be proved similarly.

By Lemma 19, $E(\text{sign}(A)|N) = E(\text{sign}(A)|\ |A| < |B|)$.

Now,

$$
\begin{aligned}
E(\text{sign}(A)|\ |A|<|B|) &= EE(\text{sign}(A)|\ |Y(u_1)-Y(u_2)|<M) \qquad \text{where } M>0\\
&= EE(\text{sign}(A)|\ -M<Y(u_1)-Y(u_2)<M)\\
&= EE(\text{sign}(A)|\ -M<Y(u_1)-Y(u_2)<0 \text{ or } 0<Y(u_1)-Y(u_2)<M)\\
&= E[\frac{E(\text{sign}(A)|\ -M<Y(u_1)-Y(u_2)<0)P(-M<Y(u_1)-Y(u_2)<0)}{P(-M<Y(u_1)-Y(u_2)<M)}+\\
&\qquad \frac{E(\text{sign}(A)|\ 0<Y(u_1)-Y(u_2)<M)P(0<Y(u_1)-Y(u_2)<M)}{P(-M<Y(u_1)-Y(u_2)<M)}]\\
&= E[\frac{T_1P(-M<Y(u_1)-Y(u_2)<0)+T_2P(0<Y(u_1)-Y(u_2)<M)}{P(-M<Y(u_1)-Y(u_2)<M)}],
\end{aligned}
$$

where $T_1 = E(\text{sign}(A)|\ -M<Y(u_1)-Y(u_2)<0)$ and $T_2 = E(\text{sign}(A)|\ 0<Y(u_1)-Y(u_2)<M)$. By Assumption 1,

$$
\begin{aligned}
T_2 &= E(\text{sign}(A)|\ 0<Y(u_1)-Y(u_2)<M)\\
&= P(X(u_1)-X(u_2)>0|0<Y(u_1)-Y(u_2)<M)\\
&- P(X(u_1)-X(u_2)<0|0<Y(u_1)-Y(u_2)<M)\\
&> 0.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
T_1 &= E(\text{sign}(A)|\ -M<Y(u_1)-Y(u_2)<0)\\
&= P(X(u_1)-X(u_2)<0|-M<Y(u_1)-Y(u_2)<0)\\
&- P(X(u_1)-X(u_2)>0|-M<Y(u_1)-Y(u_2)<0)\\
&> 0.
\end{aligned}
$$

This implies $E(\text{sign}(A)|\ |A|<|B|)>0$.

□

3.4.3 Proof of Theorem 10:

Proof. Due to symmetry of two configurations, it is enough to prove the theorem for one configuration. We consider the case of Figure 3.2.2 (4th configuration). Here $I_1 = u_1$, $I_2 = u_2$, $I_1^c = \epsilon_1 + \eta_1$ and $I_2^c = \epsilon_2 + \eta_2$.

The Kendall's tau for this nonsynchronous configuration, as defined in section 1, is:

$$\begin{aligned}
\rho_\tau &= E(\text{sign}(X_1 - X_2)(Y_1 - Y_2)) \\
&= E(\text{sign}(X_1(u_1) - X_2(u_2))(Y_1(\epsilon_1) + Y_1(u_1) + Y_1(\eta_1) - Y_2(\epsilon_2) - Y_2(u_2) - Y_2(\eta_2))) \\
&= E(\text{sign}\{(X_1(u_1) - X_2(u_2))(Y_1(u_1) - Y_2(u_2)) \\
&\quad + (X_1(u_1) - X_2(u_2))(Y_1(\epsilon_1) + Y_1(\eta_1) - Y_2(\epsilon_2) - Y_2(\eta_2))\}).
\end{aligned}$$

According to our notation, $A = (X_1(u_1) - X_2(u_2))(Y_1(u_1) - Y_2(u_2))$ and $B = (X_1(u_1) - X_2(u_2))(Y_1(\epsilon_1) + Y_1(\eta_1) - Y_2(\epsilon_2) - Y_2(\eta_2))$. So, $\rho_\tau = E(\text{sign}(A + B))$ and $\tilde{\rho_\tau} = E(\text{sign}(A))$.

In Figure 3.2.3 the shaded region is the range where $\text{sign}(A) \neq \text{sign}(A + B)$ because A and B are of opposite sign and $|B| > |A|$. Let us denote the region by N i.e. $N = \{\text{sign}(A) \neq \text{sign}(B)\ \&|B| > |A|\}$.

$$\begin{aligned}
E(\text{sign}(A + B)) &= E(\text{sign}(A + B)|N^c)P(N^c) + E(\text{sign}(A + B)|N)P(N) \\
&= E(\text{sign}(A)|N^c)P(N^c) - E(\text{sign}(A)|N)P(N).
\end{aligned}$$

But

$$E(\text{sign}(A)) = E(\text{sign}(A)|N^c)P(N^c) + E(\text{sign}(A)|N)P(N).$$

By Lemma 20, $|E(\text{sign}(A)|N)| > 0$. This implies $|E(\text{sign}(A+B))| < |E(\text{sign}(A))|$.

□

3.4.4 Proof of Theorem 11:

Due to symmetry of first and fourth configurations it is enough to prove the results, concerning first and fourth configurations, for only 4th configuration. Note that for 4th configuration (Figure 3.2.2), $I_1 = u_1$, $I_2 = u_2$, $I_1^c = \epsilon_1 + \eta_1$ and $I_2^c = \epsilon_2 + \eta_2$. According to our notation, $A = (X_1(u_1) - X_2(u_2))(Y_1(u_1) - Y_2(u_2))$ and $B = (X_1(u_1) - X_2(u_2))(Y_1(\epsilon_1) + Y_1(\eta_1) - Y_2(\epsilon_2) - Y_2(\eta_2))$.

Proof. Due to symmetry of two configurations, it is enough to prove the theorem for one configuration. We consider the case of Figure 3.2.2 (4th configuration). Here $I_1 = u_1$, $I_2 = u_2$, $I_1^c = \epsilon_1 + \eta_1$ and $I_2^c = \epsilon_2 + \eta_2$.

The Kendall's tau for this nonsynchronous configuration, as defined in section 1, is:

$$\begin{aligned}
\rho_\tau &= E(\text{sign}(X_1 - X_2)(Y_1 - Y_2)) \\
&= E(\text{sign}(X_1(u_1) - X_2(u_2))(Y_1(\epsilon_1) + Y_1(u_1) + Y_1(\eta_1) - Y_2(\epsilon_2) - Y_2(u_2) - Y_2(\eta_2))) \\
&= E(\text{sign}\{(X_1(u_1) - X_2(u_2))(Y_1(u_1) - Y_2(u_2)) \\
&\quad + (X_1(u_1) - X_2(u_2))(Y_1(\epsilon_1) + Y_1(\eta_1) - Y_2(\epsilon_2) - Y_2(\eta_2))\}).
\end{aligned}$$

According to our notation, $A = (X_1(u_1) - X_2(u_2))(Y_1(u_1) - Y_2(u_2))$ and $B = (X_1(u_1) - X_2(u_2))(Y_1(\epsilon_1) + Y_1(\eta_1) - Y_2(\epsilon_2) - Y_2(\eta_2))$. So,

$$\begin{aligned}
\rho_\tau &= E(\text{sign}(A+B)) \\
&= E(\text{sign}(A)|\text{sign}(A+B)=\text{sign}(A))P(\text{sign}(A+B)=\text{sign}(A)) \\
&- E(\text{sign}(A)|\text{sign}(A+B)\neq\text{sign}(A))P(\text{sign}(A+B)\neq\text{sign}(A)) \\
&= E(\text{sign}(A)|\text{sign}(A+B)=sign(A))p - E(\text{sign}(A)|\text{sign}(A+B)\neq\text{sign}(A))(1-p) \\
&= p(E(\text{sign}(A)|\text{sign}(A+B)=\text{sign}(A)) + E(\text{sign}(A)|\text{sign}(A+B)\neq\text{sign}(A))) \\
&- E(\text{sign}(A)|\text{sign}(A+B)\neq\text{sign}(A))
\end{aligned}$$

and

$$\begin{aligned}
\tilde{\rho}_\tau &= E(\text{sign}(A)|\text{sign}(A+B)=\text{sign}(A))P(\text{sign}(A+B)=\text{sign}(A)) \\
&+ E(\text{sign}(A)|\text{sign}(A+B)\neq\text{sign}(A))P(\text{sign}(A+B)\neq\text{sign}(A)) \\
&= p(E(\text{sign}(A)|\text{sign}(A+B)=\text{sign}(A)) - E(\text{sign}(A)|\text{sign}(A+B)\neq\text{sign}(A))) \\
&+ E(\text{sign}(A)|\text{sign}(A+B)\neq\text{sign}(A)),
\end{aligned}$$

where $\tilde{\rho}_\tau$ is the true Kendall's tau and $p = P(\text{sign}(A+B)=\text{sign}(A))$.

Therefore,

$$\rho_\tau + \tilde{\rho}_\tau = 2p(E(\text{sign}(A)|\text{sign}(A+B)=\text{sign}(A))). \tag{3.4.1}$$

Now we have to calculate the probability p.

$$
\begin{aligned}
p &= P(\text{sign}(A+B) = \text{sign}(A)) \\
&= P(\{\text{sign}(A) = \text{sign}(B)\} \textit{ or } \{\text{sign}(A) \neq \text{sign}(B) \ \& \ |A| > |B|\}) \\
&= P(\{\text{sign}(A) = \text{sign}(B)\}) + P(\{\text{sign}(A) \neq \text{sign}(B) \ \& \ |A| > |B|\}) \\
&= P(\{\text{sign}(A) = \text{sign}(B)\}) + P(\{\text{sign}(A) \neq \text{sign}(B)\})P(|A| > |B|) \\
&= \frac{1}{2} + \frac{1}{2}P(|A| > |B|).
\end{aligned}
$$

The 3rd step of the above derivation is justified because $\{\text{sign}(A) = \text{sign}(B)\}$ and $\{\text{sign}(A) \neq \text{sign}(B) \ \& \ |A| > |B|\}$ are clearly independent. Also independence of $\{\text{sign}(A) \neq \text{sign}(B)$ and $\{|A| > |B|\}$ is self evident which results in step 4. Note that $\text{sign}(A)$ and $\text{sign}(B)$ will depend on $\text{sign}(Y_1(u_1) - Y_2(u_2))$ and $\text{sign}(Y_1(\epsilon_1) + Y_1(\eta_1) - Y_2(\epsilon_2) - Y_2(\eta_2))$. Due to independent increment property, $\text{sign}(Y_1(u_1) - Y_2(u_2))$ and $\text{sign}(Y_1(\epsilon_1) + Y_1(\eta_1) - Y_2(\epsilon_2) - Y_2(\eta_2))$ are independent. Therefore $P(\text{sign}(A) = \text{sign}(B)) = P(\text{sign}(A) \neq \text{sign}(B)) = \frac{1}{2}$.

Let us denote the events $C = \{\text{sign}(A) = \text{sign}(B)\}$ and $D = \{\text{sign}(A) \neq \text{sign}(B) \& |A| > |B|\}$. Note that $\tilde{\rho}_\tau = E(\text{sign}(A)) = E(\text{sign}(A)|C)$ as condition on the event C does not influence the expected value of $\text{sign}(A)$.

$$
\begin{aligned}
E(\text{sign}(A)|\text{sign}(A+B) = \text{sign}(A)) & \\
&= \frac{E(\text{sign}(A)|C)P(C) + E(\text{sign}(A)|D)P(D)}{P(C \cup D)} \\
&= \frac{\tilde{\rho}_\tau/2 + E(\text{sign}(A)|D)(P(|A| > |B|)/2)}{p}.
\end{aligned}
$$

Inserting this to 5.2.1 we get

$$\rho_\tau + \tilde{\rho}_\tau = 2p\frac{\tilde{\rho}_\tau/2 + E(\text{sign}(A)|D)(P(|A| > |B|)/2)}{p}$$
$$\implies \rho_\tau = E(\text{sign}(A)|D)P(|A| > |B|).$$

Hence we proved the result.

□

3.4.5 Proof of Theorem 12

Proof. To do so, it is enough to show that given N^c is nonempty, $E(\text{sign}(A)|N^c) > E(\text{sign}(A)|N)$. Note that, $N^c = N_1 + N_2 + N_3$, where $N_1 = \{\text{sign}(A) = \text{sign}(B)\} \cap \{|A| > |B|\}$, $N_2 = \{\text{sign}(A) \neq \text{sign}(B)\} \cap \{|A| > |B|\}$, $N_3 = \{\text{sign}(A) = \text{sign}(B)\} \cap \{|A| < |B|\}$.

Therefore,

$$E(\text{sign}(A)|N^c) = \frac{E(\text{sign}(A)|N_1)P(N_1) + E(\text{sign}(A)|N_2)P(N_2) + E(\text{sign}(A)|N_3)P(N_3)}{P(N)}.$$

By Lemma 19, $E(\text{sign}(A)|N_3) = E(\text{sign}(A)|N)$ and $E(\text{sign}(A)|N_2) = E(\text{sign}(A)|N_1) = E(\text{sign}(A)||A| > |B|)$.

$$
\begin{aligned}
E(\text{sign}(A) \mid |A| > |B|) &= EE(\text{sign}(A) \mid |A| > M) \\
&= E[\frac{1}{P(|A| > M)}\{E(\text{sign}(A) \mid Y(u_1) - Y(u_2) > M)P(Y(u_1) - Y(u_2) > M) + \\
&\qquad E(\text{sign}(A) \mid Y(u_1) - Y(u_2) < -M)P(Y(u_1) - Y(u_2) < -M)\}] \\
&\geq E[\frac{1}{P(|A| > M)}\{E(\text{sign}(A) \mid 0 < Y(u_1) - Y(u_2) < M)P(Y(u_1) - Y(u_2) > M) \\
&\qquad + E(\text{sign}(A) \mid 0 < Y(u_1) - Y(u_2) < M)P(Y(u_1) - Y(u_2) < -M)\}] \\
&= EE(\text{sign}(A) \mid 0 < Y(u_1) - Y(u_2) < M).
\end{aligned}
$$

The inequality is due to Assumption 2a. Similarly,

$$
\begin{aligned}
E(\text{sign}(A) \mid |A| > |B|) &= EE(\text{sign}(A) \mid |A| > M) \\
&= E[\frac{1}{P(|A| > M)}\{E(\text{sign}(A) \mid Y(u_1) - Y(u_2) > M)P(Y(u_1) - Y(u_2) > M) + \\
&\qquad E(\text{sign}(A) \mid Y(u_1) - Y(u_2) < -M)P(Y(u_1) - Y(u_2) < -M)\}] \\
&\geq E[\frac{1}{P(|A| > M)}\{E(\text{sign}(A) \mid -M < Y(u_1) - Y(u_2) < 0)P(Y(u_1) - Y(u_2) > M) \\
&\qquad + E(\text{sign}(A) \mid -M < Y(u_1) - Y(u_2) < 0)P(Y(u_1) - Y(u_2) < -M)\}] \\
&= EE(\text{sign}(A) \mid -M < Y(u_1) - Y(u_2) < 0).
\end{aligned}
$$

Therefore

$$
\begin{aligned}
E(\text{sign}(A) \mid |A| > M) &\geq \max(E(\text{sign}(A) \mid 0 < Y(u_1) - Y(u_2) < M), \\
&\qquad E(\text{sign}(A) \mid -M < Y(u_1) - Y(u_2) < 0)) \\
&\geq \frac{1}{P(|A| < M)}\{E(\text{sign}(A) \mid -M < Y(u_1) - Y(u_2) < 0)P(-M < Y(u_1) - Y(u_2) < 0) \\
&\qquad + E(\text{sign}(A) \mid 0 < Y(u_1) - Y(u_2) < M)P(0 < Y(u_1) - Y(u_2) < M)\}] \\
&\geq E(\text{sign}(A) \mid |A| < M)
\end{aligned}
$$

and

$$E(\text{sign}(A) \mid |A| > |B|) \geq E(\text{sign}(A) \mid |A| < |B|).$$

With this, we get that

$$\begin{aligned} E(\text{sign}(A)|N^c) &= \frac{E(\text{sign}(A)|N_1)P(N_1) + E(\text{sign}(A)|N_2)P(N_2) + E(\text{sign}(A)|N_3)P(N_3)}{P(N)} \\ &\geq \frac{E(\text{sign}(A)|N)P(N_1) + E(\text{sign}(A)|N)P(N_2) + E(\text{sign}(A)|N)P(N_3)}{P(N)} \\ &= E(\text{sign}(A)|N), \end{aligned}$$

which implies $E(\text{sign}(A)|N^c) > E(\text{sign}(A)|N)$. This completes the proof.

□

,

Chapter 4

Spectral Distribution of Hayashi-Yoshida Estimator

4.1 Introduction

One of the important quantities in financial markets to look at is the volatility which is a measure of dispersion of a given security. For intraday data of asset returns the volatility we are concerned with is called *Integrated volatility* (IV). It is the variance of daily log-return.

Often univariate modeling is not sufficient, it is also important to model the correlation dynamics between several assets. When we are concerned with multiple stocks and their interrelation we analogously capture it by variance-covariance matrix and call it *integrated co-volatility*. The question is how to accurately estimate it. For single stock price data Merton (1980) showed that the variance over a fixed interval can be estimated accurately as sum of squared realization as long as the data are available at sufficiently high sampling fre-

quency [17]. This estimator is known as *Realized volatility*. Analogously, for multiple stocks, we can define a *realized volatility matrix/realized covariance matrix*. But realized volatility matrix relies upon synchronous observations and ignores inherent asynchronicity of the stock price processes. Therefore in order to exercise the tool at our disposal we have to first force the data to be synchronized. There are two well known methods of synchronizing- fixed clock time and refresh time [36]. But the estimator, we get after synchronizing, is biased. Hayashi and Yoshida (2005) proposed an alternative estimator of Integrated covolatility, which is unbiased, without force-synchronization of the data [37]. Although in presence of microstructure noise Hayashi-Yoshida estimator is biased, a bias-corrected version can be developed [38]. But as the dimension of the data increases the estimator becomes more and more unstable (no more a consistent estimator of ICV matrix). Developing good estimate of high-dimensional covolatility is challening unless we impose some structure. A consistent and positive definite estimator is proposed based on blocking and regularization techniques [39]. The central idea is to obtain one large covariance matrix from a series of smaller covariance matrices, each based on different sampling frequency. Shrinkage estimator of covariance matrix with optimal shrinkage intensity also reduces the estimation error significantly [40]. The shrinkage estimator is also useful for the purpose of portfolio optimization. The choice of sampling frequency can be important in presence of microstructure noise [41] [31]. Sub-sampling can be used to reduce the variance of the estimator [14]. The mixed frequency factor models, that uses high-frequency data to estimate factor covariance and low frequency data to estimate the factor loadings, are also good alternative to estimate high-dimensional covariance matrices [42]. The composite realized Kernel approach that estimates each entry of ICV matrix optimally (in terms of bandwidth and data loss) has been proposed and

asymptotic properties are established [43].

In absence of the structure an alternative approach is to study the spectral distribution of the estimator. Zheng, Li (2011) explored the limiting spectral distribution of realized covariance matrix [44]. Recently an asymptotic relationship has been established between the limiting spectral distribution of sample covariance matrix and sample covariance matrix based on noisy data [45]. Limiting spectral distribution of the covariance matrices of time lagged processes have been established [46]. When modeled as VARMA(p, q) the limiting spectral distribution of sample covariance matrix can be derived [47]. We intend to establish limiting spectral distribution for Hayashi's estimator which has not yet been studied. In Section 4.2 we discuss the background of the problem. Section 4.3 deals with a very brief introduction to Random matrix theory. In Section 4.4 we determine the limiting spectral distribution for high-dimensional stock price process. Simulation and real data analysis results are presented in Section 4.5.

4.2 Integrated Covariance Matrix and Asynchronicity

Suppose that we have multiple stocks, say, p stocks, whose price processes are denoted by S_t^j for $j = 1, ..., p.$ and $X_t^j := \log S_t^j$ are the log price processes. Let $X_t = (X_t^1, ..., X_t^p)^T$. Then we can model X_t as a p-dimensional diffusion process described as

$$dX_t = \mu_t dt + \sigma_t dW_t, \tag{4.2.1}$$

where μ_t is p dimensional drift process and σ_t is a pxp matrix, called instantaneous covolatility process. W_t is a p dimensional standard Brownian motion.

The Integrated covariance (ICV) matrix is defined by

$$\Sigma_p = \int_0^1 \sigma_t \sigma_t^T dt. \tag{4.2.2}$$

This is our parameter of interest.

The observations occur in high-frequency at random time points for each stock. Let n_i be the number of observations for ith stock. Throughout this paper we are assuming all n_j's are equal, say n. Suppose that lth observation of the ith stock is observed at random time point t_l^i . As discussed, a very commonly used estimator (when the observations are assumed to be synchronous i.e. $t_l^i = t_l$ for $\forall i$) of ICV matrix is the Realized Covariance (RCV) matrix:

$$\Sigma_p^{RCV} = \sum_{l=1}^{n} \Delta X_l \Delta X_l^T, \tag{4.2.3}$$

where $\Delta X_l = \begin{pmatrix} \Delta X_l^1 \\ \Delta X_l^2 \\ . \\ . \\ \Delta X_l^p \end{pmatrix} = \begin{pmatrix} X_{t_l}^1 - X_{t_{l-1}}^1 \\ . \\ . \\ . \\ X_{t_l}^p - X_{t_{l-1}}^p \end{pmatrix}$.

Formation of RCV matrix assumes that the prices are observed synchronously for all the stocks, which is not a realistic assumption. Force synchronization can cause bias. In order to avoid force synchronization Hayashi and Yoshida proposed an alternative estimator for ICV matrix. It is defined as:

$$\Sigma_p^{HY} = \sum_{k,l} \Delta X_k \Delta X_l^T \circ I(k,l), \tag{4.2.4}$$

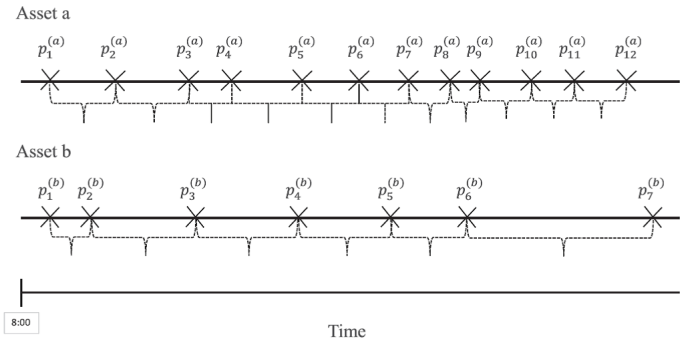


Figure 4.2.1: Positions of arrivals and the interarrivals of two assets- Asset a and Asset b.

where $\Delta X_l = \begin{pmatrix} \Delta X_l^1 \\ \Delta X_l^2 \\ . \\ . \\ \Delta X_l^p \end{pmatrix} = \begin{pmatrix} X^1_{t_l^1} - X^1_{t_{l-1}^1} \\ . \\ . \\ . \\ X^p_{t_l^p} - X^p_{t_{l-1}^p} \end{pmatrix}$, '$\circ$' is Hadamard product and $I(k,l)$ is a $p \times p$ matrix with $(i,j)^{th}$element is the indicator function involving k^{th}interarrival of i^{th}stock and l^{th}interarrival of j^{th}stock: $I(I_k^i \cap I_l^j \neq \phi)$, where $I_k^i = (t_{k-1}^i, t_k^i)$. In other words if two interarrivals intersect then product $\Delta X_k^i \Delta X_l^{j\,T}$ will be included in the sum. In Figure 4.2.1, $I_k^a = (t_{k-1}^a, t_k^a) = (p_{k-1}^a, p_k^a)$ and $I_k^b = (t_{k-1}^b, t_k^b) = (p_{k-1}^b, p_k^b)$ are shown.

For low-dimensions Hayashi's estimator is consistent estimator of ICV matrix. But for high-dimensional stock price process, when the dimension grows at same rate as the number of observations, neither RCV [44] nor Hayashi's estimator is consistent anymore. Here we want to study the asymptotic behavior of the distribution of eigenvalues of Hayashi's matrix.

4.3 Spectral Distribution

The *empirical spectral distribution (ESD)* function of a symmetric (more generally Hermitian) matrix $A_{p\times p}$ is defined as

$$F_p^A(x) = \frac{1}{p}\#\{j \leq p : \ \lambda_j \leq x\},$$

where $\lambda_j' s$ are the eigenvalues of the matrix A and $\#E$ denotes the cardinality of set E. In this work we are going to investigate the convergence of ESD of a sequence of random matrices A_n. The limit distribution (F) is called *limiting spectral distribution (LSD)*. One commonly used method of finding LSD is through Stieltjes transform. Let $A_{p\times p}$is a Hermitian matrix and F_p^A be its ESD. Then the Stieltjes transform of F_p^A is defined as

$$\begin{aligned} s_p(z) &= \int \frac{1}{x-z} dF_p^A(x) \\ &= \frac{1}{p} tr(A - zI)^{-1}, \end{aligned}$$

where $z \in D = \{z \in \mathbb{C} | \mathbb{I}(z) > 0\}$, $\mathbb{I}(z)$ being the imaginary part of z. The role Stieltjes transformation play in Random Matrix Theory is due to the following theorems. For reference see [48].

Theorem: If G is a function of bounded variation of real line then for any continuity points $a < b$ of G, we have

$$G\{[a,b]\} = lim_{\epsilon \to 0^+} \frac{1}{\pi} \int_a^b \mathbb{I}(s_G(x + i\epsilon))dx,$$

where $s_G(z) = \int \frac{1}{\lambda - z} dG(\lambda)$.

Theorem: If G_n is a sequence of functions of bounded variation of real line

and $G_n(-\infty) = 0 \, \forall n$ then

$$lim_{n\to\infty} s_{G_n}(z) = s(z) \;\; \forall z \in D,$$

if and only if there is a function G of bounded variation with $G(-\infty) = 0$ and Stieltjes transform $s(z)$ such that $G_n \to G$ vaguely.

In the next section we are going to derive the limiting spectral distribution of Hayashi's estimator.

4.4 Spectral Analysis of High-Dimensional Hayashi's Estimator

We are now going to discuss very briefly what kind of assumption we would need in order to achieve our goal. As we are taking into account the time variability of the covolatility process we have to assume something on the nature of time dependence or to consider a restricted class of time-varying co-volatility processes. Here we assume that $\sigma_t = \gamma_t \Sigma^{\frac{1}{2}}$ where $\Sigma^{\frac{1}{2}}$ a $p \times p$ symmetric matrix (not time varying) and γ_t is a time varying *cadlag* process. This is also the assumption taken by Zeng, Li while investigating the spectral distribution of Realized volatility matrix [44]. If we assume this, then ΔX_l has the same distribution as the following-

$$\Delta X_l = \Sigma^{\frac{1}{2}} \begin{pmatrix} \int_{t_{l-1}^1}^{t_l^1} \gamma_s dW_1 \\ \int_{t_{l-1}^2}^{t_l^2} \gamma_s dW_2 \\ . \\ . \\ \int_{t_{l-1}^p}^{t_l^p} \gamma_s dW_p \end{pmatrix} \stackrel{d}{=} (\Sigma^{1/2} \circ A) diag(\sqrt{\int_{t_{l-1}^1}^{t_l^1} \gamma_s^2 ds}, \sqrt{\int_{t_{l-1}^2}^{t_l^2} \gamma_s^2 ds}, ..., \sqrt{\int_{t_{l-1}^p}^{t_l^p} \gamma_s^2 ds}) x_l,$$

where $x_l^p = (x_l^j)_{1\le j\le p}$;$Z_l^j$'s are iid normal with mean 0 and variance 1 for $p = 1, 2, ...$ and $1 \le l \le n$ and $A = (a_{ij})$ when $a_{ij} = \int_{t_{l-1}^i \vee t_{l-1}^j}^{t_l^i \wedge t_l^j} \gamma_s^2 I((t_{l-1}^i, t_l^i) \cap I(t_{l-1}^j, t_l^j))ds$.

Let us write $D_l = diag(\sqrt{\int_{t_{l-1}^1}^{t_l^1} \gamma_s^2 ds}, \sqrt{\int_{t_{l-1}^2}^{t_l^2} \gamma_s^2 ds}, ..., \sqrt{\int_{t_{l-1}^p}^{t_l^p} \gamma_s^2 ds})$.

Therefore we are interested in the spectral distribution of

$$S_0 = \sum_k \sum_l \{(\Sigma^{1/2} \circ A) D_k x_k x_l^T D_l (\Sigma^{1/2} \circ A) \circ I(k, l)\}$$

as both S_0 and Σ_p^{HY} have the same LSD. However in order to determine the LSD we are going to use some already known lemmas and some of them demands the condition of nonnegative-definiteness of the matrix. Therefore it is difficult and challenging to work with S_0. Instead, we will work with a nnd matrix which is in some sense close to S_0 and then we need to show that LSD of S_0 and S are same. We will return to this point again in assumption 5 of the main theorem.

Secondly, we are going to assume dimension of the matrix is increasing with number of observations and the ratio of two will converge to a nonzero finite quantity. Thirdly, we assume spectral distribution of Σ , $F^{\Sigma} \Rightarrow H$ and H has finite second order moment. Fourthly, some assumption regarding the time varying component γ_t would be helpful. Finally, assumptions on relation between γ_t and x_i's and also between x_i's and Σ are necessary. Apart from these we would need additional assumptions along the way of establishing theoretical result.

We summarize all the assumptions and denote it by $\mathcal{A}$ described below. Condition $\mathcal{A}5$ is a bit technical where the sequence τ_l are carefully chosen to construct S such that the $tr(S_0 - S)^2$ is relatively samll compared to p.

Assumption $\mathcal{A}$:

1. $y_n = {}^p/_n \to y > 0$ as $p \to \infty$

2. $F^{\Sigma} \stackrel{d}{\Rightarrow} H$ ($\neq$ delta measure at 0) a.s. and H has a finite second order moment.

3. x_l^j's ($j = 1(1)p$) are iid with mean 0, variance 1 and finite moments of all orders.

4. $\int_{t_{l-1}^j}^{t_l^j} \gamma_s^2 ds$) are independent of Z_l.

5. $\exists\ \tau_l \in \mathbb{R}^+$ and let $S = \sum_l \tau_l \Sigma^{1/2} x_l (x_l)^T \Sigma^{1/2}$ such that ${}^{tr(S_0-S)^2}/_p \to 0$ with probability 1 and $tr([(S_0 - zI)^{-1}(S - zI)^{-1}]^2)/p$ is bounded with probability 1 where $(S_0 - zI)^{-1}(S - zI)^{-1}$ being positive semi definite and $S_0 = \sum_k \sum_l \{\Sigma^{1/2} D_k x_k x_l^T D_l \Sigma^{1/2} \circ I(k,l)\}$ and $z = i.v$ with $v > 0$ and sufficiently large.

6. There exists κ, such that $\max_{\mathrm{n}} \max_{\mathrm{l}} (n\tau_l) \leq \kappa$. Also there exists a nonnegative cadlag process τ_s' such that

$$\lim_{\mathrm{n}} \sum_{1 \leq l \leq n} \int_{\frac{l-1}{n}}^{\frac{l}{n}} |n\tau_l - \tau_s'| ds = 0$$

7. $\exists$a $K < \infty$ and $\delta < 1/6$ such that for all p, $\|\Sigma\| \leq Kp^{\delta}$almost surely.

Now we are ready to state the main theorem.

Theorem. *Under the assumptions ($\mathcal{A}$), almost surely, ESD of S_0 converges in distribution to a probability distribution with Stieltjes transform*

$$s(z) = -\frac{1}{z} \int \frac{1}{1 - \frac{\lambda}{z} \int_0^1 \frac{\tau_t'}{1 + y\tau_t' \tilde{s}(z)} dt} dH(\lambda)$$

where $\tilde{s}(z)$ can be solved by the equation: $\tilde{s}(z) = -\frac{1}{z} \int_{\lambda} \frac{\lambda}{1 - \frac{\lambda}{z} \int_0^1 \frac{\tau_t'}{1 + y\tau_t' \tilde{s}(z)} dt} dH(\lambda)$.

This theorem establishes the Limiting spectral distribution of Hayashi-Yoshida estimator by determining its spectral distribution through the limiting spectral distribution of Σ. In other words the theorem establishes the link between the limiting spectral distributions of Σ and Hayashi-Yoshida estimator through its Stieltjes transform.

Before proving the theorem let us discuss the following Lemmas, which will be used in the proof (everywhere the norm being used, is spectral norm)

Lemma 16. *Let $w_{1,}w_2 \in \mathbb{C}$, with $Re(w_1) \geq 0$ and $Re(w_2) \geq 0$, A be a $p \times p$ Hermitian nnd matrix B any $p \times p$ matrix matrix, and $q \in \mathbb{C}^{\prime}$, then*

$$|q^*B(w_1A + I)^{-1}q - q^*B(w_2A + I)^{-1}q| \leq |w_1 - w_2|.|q|^2||B||.||A||$$

Proof: See [49]

Lemma 17. *Let $z \in \mathbb{C}$, with $v = I(z) > 0$, A and B be $p \times p$ with B Hermitian, and $q \in \mathbb{C}^p$. Then*

$$|tr(((B - zI)^{-1} - (B + \tau qq^* - zI)^{-1}A| \leq \frac{||A||}{v}$$

for all $\tau \in \mathbb{R}$.

Proof: See [49]

Lemma 18. *For any Hermitian matrix A and $z \in \mathbb{C}$, with $Im(z) = v > 0$*

$$||(A - zI)^{-1}|| \leq 1/v.$$

Proof: See [48]

Lemma 19. *For $X = (X_1, X_2, ..., X_p)^T$ where X_j's are iid random variables such that $\mathbb{E}(X_1) = 0$, $\mathbb{E}|X_1|^2 = 1$, and $\mathbb{E}|X_1|^{2k} < \infty$ for some $2 \leq k \in \mathbb{N}$, there exists $C_k \geq 0$, depending only on k, $\mathbb{E}|X_1|^4$and $\mathbb{E}|X_1|^{2k}$, such that for any $p \times p$ nonrandom matrix A,*

$$\mathbb{E}|X^*AX - tr(A)|^{2k} \leq C_k(tr(AA^*))^k \leq C_k p^k \|A\|^{2k} \tag{4.4.1}$$

Proof: See [44]

Lemma 20. *Suppose S is a matrix defined as assumption $\mathcal{A}_5$ and $\frac{1}{p}tr(S_{n_k} - zI)^{-1} \to s(z)$, then $\sum_{l=1}^{n_k} \frac{\tau_l}{1+\tau_l a_l} \to \int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt \neq 0$ where $\tilde{s}(z)$ is the unique solution in Q_1 to the following equation*

$$\int_0^1 \frac{\tau_t'}{1 + y\tau_t'\tilde{s}(z)}dt = 1 - y(1 + zs(z))$$

Proof: [44]

Lemma 21. *Let $z = iv \in \mathbb{C}$,with $v > 0$, A be any $p \times p$ matrix, and B be a*

$p \times p$ Hermitian nonnegative definite matrix. Then $tr(A(B - zI)^{-1}A^*) \in Q_1$.

Proof: See [44]

Lemma 22. *Let $z = iv \in \mathbb{C}$ with $v > 0$, A be a $p \times p$ Hermitian nonnegative definite matrix, $q \in \mathbb{C}^p, a > 0$. Then*

$$-\frac{1}{z}.\frac{1}{1 + a.q^*(A - zI)^{-1}q} \in Q_1 = \{z \in \mathbb{C} : Re(z) \geq 0, Im(z) \geq 0\}. \quad (4.4.2)$$

Proof: See [44]

Lemma 23. *Suppose that P_n are real probability measures with Stieltjes transforms $s_n(z)$. Let $K \subset \mathbb{C}_+$ be an infinite set with limit points in $\mathbb{C}_+$. If $lim s_n(z) = s(z)$ exists for all $z \in K$, then there exists a Probability measure P with Stieltjes transform $m(z)$ if and only if*

$$lim_{v\to\infty} iv.s(iv) = -1$$

in which case $P_n \to P$.

Proof: See [50].

Lemma 24. *For x_l, Σ, τ_l described as Assumption $\mathcal{A}$ and $z = iv$ with $v > 0$, define,*

$$M_1 = x_l'\Sigma^{\frac{1}{2}}(S_l - zI)^{-1}(\Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} - zI)^{-1}\Sigma^{\frac{1}{2}}x_l$$

and

$$M_2 = x_l' \Sigma^{\frac{1}{2}} (S_l - zI)^{-1} (\Sigma \sum_{j \neq l} \frac{\tau_j}{1 + \tau_j b_j^l} - zI)^{-1} \Sigma^{\frac{1}{2}} x_l$$

Then,

$$\frac{1}{p} \max_l (M_1 - M_2) \to 0 \ a.s.$$

Proof: See Appendix.

Lemma 25. *For x_l, Σ, τ_l described as Assumption $\mathcal{A}$ and $z = iv$ with $v > 0$, define,*

$$M_3 = x_l' \Sigma^{\frac{1}{2}} (S_l - zI)^{-1} (\Sigma \sum_{j \neq l} \frac{\tau_j}{1 + \tau_j b_j^l} - zI)^{-1} \Sigma^{\frac{1}{2}} x_l$$

and

$$M_4 = tr[\Sigma^{\frac{1}{2}} (S_l - zI)^{-1} (\Sigma \sum_{j \neq l} \frac{\tau_j}{1 + \tau_j b_j^l} - zI)^{-1} \Sigma^{\frac{1}{2}}],$$

then

$$\frac{1}{p} max_l |M_3 - M_4| \to 0 \ a.s.$$

Proof: See Appendix.

Now we are ready to prove the theorem. But note that it is enough to show that the theorem holds for $z = iv$, where $v > 0$ is suficiently large (See Appendix for justification).

Proof of Theorem:

Define $\bar{S} = \Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l}$ where $a_l = x_l' \Sigma^{\frac{1}{2}} (S_l - zI)^{-1} \Sigma^{\frac{1}{2}} x_l$ and $S_l = \sum_{j \neq l} \Sigma^{\frac{1}{2}} x_j x_j' \Sigma^{\frac{1}{2}}$.

We denote the Stieltjes transform of F^S by s_n,

$$s_n := s_n(z) = \frac{tr((S - zI)^{-1})}{p}.$$

Similarly, Stieltjes transform of $F^{S_0}$and Stieltjes transform of $F^{\bar{S}}$ are denoted by s_n^0 and $\bar{s}_n$ respectively.

In order to show the convergence of $s_n(z)$ it is enough to show for all $z = iv$ with $v > 0$ sufficiently large and the condition, in Lemma 32, is satisfied.

We will start by showing that s_n^0 and $\bar{s}_n$ converge to the same limit.

$$
\begin{aligned}
&\bar{s}_n - s_n^0 \\
&= \frac{1}{p} tr[(\bar{S} - zI)^{-1} - (S_0 - zI)^{-1}] \\
&= \frac{1}{p} tr[(\bar{S} - zI)^{-1} - (S - zI)^{-1} + (S - zI)^{-1} - (S_0 - zI)^{-1}] \\
&= \frac{1}{p} tr[(\bar{S} - zI)^{-1}[S - \bar{S}](S - zI)^{-1}] + tr[\frac{(S - zI)^{-1} - (S_0 - zI)^{-1}}{p}] \\
&= \frac{1}{p} tr[(\Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} - zI)^{-1}[\sum_{l=1}^{n} \tau_l \Sigma^{\frac{1}{2}} x_l x_l' \Sigma^{\frac{1}{2}} - \Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l}](S - zI)^{-1}] \\
&+ tr[\frac{(S - zI)^{-1}(S_0 - zI - S + zI)(S_0 - zI)^{-1}}{p}] \\
&= \frac{1}{p} tr[(\Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} - zI)^{-1}[\sum_{l=1}^{n} \tau_l \Sigma^{\frac{1}{2}} x_l x_l' \Sigma^{\frac{1}{2}}] \frac{(S_l - zI)^{-1}}{1 + \tau_l a_l}] \\
&- \frac{1}{p} tr[(\Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} - zI)^{-1}[\Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l}](S - zI)^{-1}] \\
&+ tr[\frac{(S - zI)^{-1}(S_0 - S)(S_0 - zI)^{-1}}{p}] \\
&= \frac{1}{p} tr[\sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} (\Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} - zI)^{-1} \Sigma^{\frac{1}{2}} x_l x_l' \Sigma^{\frac{1}{2}} (S_l - zI)^{-1}] \\
&- \frac{1}{p} tr[\sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} (\Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} - zI)^{-1} \Sigma (S - zI)^{-1}] \\
&+ tr[\frac{(S_0 - S)(S_0 - zI)^{-1}(S - zI)^{-1}}{p}] \\
&= \frac{1}{p} \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} tr[(\Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} - zI)^{-1} \Sigma^{\frac{1}{2}} x_l x_l' \Sigma^{\frac{1}{2}} (S_l - zI)^{-1}] \\
&- \frac{1}{p} \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} tr[(\Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} - zI)^{-1} \Sigma (S - zI)^{-1}] \\
&+ tr[\frac{(S_0 - S)(S_0 - zI)^{-1}(S - zI)^{-1}}{p}] \\
&= \frac{1}{p} \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} tr[x_l' \Sigma^{\frac{1}{2}} (S_l - zI)^{-1} (\Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} - zI)^{-1} \Sigma^{\frac{1}{2}} x_l] \\
&- \frac{1}{p} \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} tr[\Sigma^{\frac{1}{2}} (S - zI)^{-1} (\Sigma \sum_{l=1}^{n} \frac{\tau_l}{1 + \tau_l a_l} - zI)^{-1} \Sigma^{\frac{1}{2}}] \\
&+ tr[\frac{(S_0 - S)(S_0 - zI)^{-1}(S - zI)^{-1}}{p}]
\end{aligned}
$$

$$\leq \frac{1}{p}\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}(x_l'\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}-zI)^{-1}\Sigma^{\frac{1}{2}}x_l$$
$$-tr[\Sigma^{\frac{1}{2}}(S-zI)^{-1}(\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}-zI)^{-1}\Sigma^{\frac{1}{2}}])$$
$$+\sqrt{\frac{tr(S_0-S)^2}{p}\times\frac{tr([(S-zI)^{-1}(S_0-zI)^{-1}]^2)}{p}}$$

According to our assumption 6, $\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}\leq n\ \max(\tau_l)\leq\kappa$.

It is enough to show that

$$\frac{1}{p}\max_l(x_l'\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}-zI)^{-1}\Sigma^{\frac{1}{2}}x_l$$
$$-tr[\Sigma^{\frac{1}{2}}(S-zI)^{-1}(\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}-zI)^{-1}\Sigma^{\frac{1}{2}}])\to 0 \text{ a.s.}$$
$$i.e.\frac{1}{p}\max_l(x_l'\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}-zI)^{-1}\Sigma^{\frac{1}{2}}x_l$$
$$-tr[\Sigma^{\frac{1}{2}}(S-zI)^{-1}(\Sigma\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l}-zI)^{-1}\Sigma^{\frac{1}{2}}])$$
$$+\frac{1}{p}\max_l(tr[\Sigma^{\frac{1}{2}}(S-zI)^{-1}(\Sigma\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l}-zI)^{-1}\Sigma^{\frac{1}{2}}]$$
$$-tr[\Sigma^{\frac{1}{2}}(S-zI)^{-1}(\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}-zI)^{-1}\Sigma^{\frac{1}{2}}])\to 0 \text{ a.s.}$$
$$i.e.\frac{1}{p}\max_l(x_l'\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}-zI)^{-1}\Sigma^{\frac{1}{2}}x_l$$
$$-x_l'\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(\Sigma\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l}-zI)^{-1}\Sigma^{\frac{1}{2}}x_l)$$
$$+\frac{1}{p}\max_l(x_l'\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(\Sigma\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l}-zI)^{-1}\Sigma^{\frac{1}{2}}x_l$$
$$-tr[\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(\Sigma\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l}-zI)^{-1}\Sigma^{\frac{1}{2}}])$$
$$+\frac{1}{p}\max_l(tr[\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(\Sigma\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l}-zI)^{-1}\Sigma^{\frac{1}{2}}]$$

$$
\begin{aligned}
&- tr[\Sigma^{\frac{1}{2}}(S_l - zI)^{-1}(\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l} - zI)^{-1}\Sigma^{\frac{1}{2}}])\\
&+ \frac{1}{p}\max_l(tr[\Sigma^{\frac{1}{2}}(S_l - zI)^{-1}(\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l} - zI)^{-1}\Sigma^{\frac{1}{2}}]\\
&- tr[\Sigma^{\frac{1}{2}}(S - zI)^{-1}(\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l} - zI)^{-1}\Sigma^{\frac{1}{2}}]) \to 0 \text{ a.s.}\\
=&I + II + III + IV
\end{aligned}
$$

where $b_j^l = x_j'\Sigma^{\frac{1}{2}}(S_{j,l} - zI)^{-1}\Sigma^{\frac{1}{2}}x_j$ with $S_{j,l} = \sum_{i\neq j,l}\Sigma^{\frac{1}{2}}x_i x_i'\Sigma^{\frac{1}{2}}$ and I, II, III, IV are four terms in the last line of the equation.

It is sufficient to show that $\text{I}, \text{II}, \text{III}, \text{IV} \to 0$ a.s.

Convergence ofI and II are followed by Lemma 33 and Lemma 34. Convergence of III and IV can be proved similarly.

Now due to Lemma 29 and Assumption 2, we have

$$\frac{1}{p}tr(\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}\Sigma - zI)^{-1} \to -\frac{1}{z}\int\frac{1}{1-\frac{\lambda}{z}\int_0^1\frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt}dH(\lambda)$$

As $\frac{1}{p}tr(\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}\Sigma - zI)^{-1} - \frac{1}{p}tr(S - zI)^{-1} \to 0$, we get

$$s(z) = -\frac{1}{z}\int\frac{1}{1-\frac{\lambda}{z}\int_0^1\frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt}dH(\lambda)$$

The fact that $\int_0^1\frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt \neq 0$ and $\text{Real}(\int_0^1\frac{\tau_t'}{1+y\tilde{s}(z)}dt) \geq 0$ implies $|s(z)| < \frac{1}{|z|}$ and therefore $1 + zs(z) \neq 0$. So from Lemma 29, it is clear that $\tilde{s}(z) \neq 0$.

Now

$$\begin{aligned}\int_0^1 \frac{\tau_t'}{1+y\tilde{s}(z)\tau_t'}dt &= \frac{1}{y\tilde{s}(z)}(1-\int_0^1 \frac{1}{1+y\tilde{s}(z)\tau_t'}dt)\\ &= \frac{1}{y\tilde{s}(z)}(1-(1-y(1+zs(z))))\\ &= \frac{1+zs(z)}{\tilde{s}(z)}.\end{aligned}$$

Also,

$$\begin{aligned}s(z) &= -\frac{1}{z}\int \frac{1}{1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt}dH(\lambda)\\ \implies 1+zs(z) &= 1-\int_{\lambda\in\mathbb{R}} \frac{1}{1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt}dH(z)\\ \implies 1+zs(z) &= -\frac{1}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt\int_\lambda \frac{\lambda}{1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt}dH(\lambda)\\ \implies \frac{1+zs(z)}{\tilde{s}(z)} &= -\frac{1}{z\tilde{s}(z)}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt\int_\lambda \frac{\lambda}{1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt}dH(\lambda)\\ \implies \int_0^1 \frac{\tau_t'}{1+y\tilde{s}(z)\tau_t'}dt &= -\frac{1}{z\tilde{s}(z)}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt\int_\lambda \frac{\lambda}{1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt}dH(\lambda)\\ \implies \tilde{s}(z) &= -\frac{1}{z}\int_\lambda \frac{\lambda}{1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt}dH(\lambda).\end{aligned}$$

Due to Lemma 30 and Lemma 31, $-\frac{1}{z}\sum_{l=1}^n \frac{\tau_l}{1+a_l\tau_l} \in Q_1$ and $\frac{1}{p}tr(S-zI)^{-1} \in Q_1$. Same will be true for their limits. Now to show that $s(z)$ is not unique it is enough to show that if there are two solutions $s_1(z)$ and $s_2(z)$ (and therefore $\tilde{s}_1(z), \tilde{s}_2(z)$) then $s_1 = s_2$. If possible let there are two limiting spectral distributions s_1 and s_2 such that $s_1 \neq s_2$. To show the contradiction, it is enough to show $\tilde{s}_1(z) = \tilde{s}_2(z)$.

Note that,

$$\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_1(z)}dt - \int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_2(z)}dt$$

$$= \int_0^1 \frac{\tau_t'((1+y\tau_t'\tilde{s}_2(z)) - (1+y\tau_t'\tilde{s}_1(z))}{(1+y\tau_t'\tilde{s}_1(z))(1+y\tau_t'\tilde{s}_2(z))}dt$$

$$= \int_0^1 \frac{y(\tau_t')^2(\tilde{s}_2(z) - \tilde{s}_1(z))}{(1+y\tau_t'\tilde{s}_1(z))(1+y\tau_t'\tilde{s}_2(z))}dt$$

But,

$$\tilde{s}_1(z) - \tilde{s}_2(z)$$

$$= -\frac{1}{z}\int_{\lambda\in\mathbb{R}} [\frac{\lambda}{1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_1(z)}dt} - \frac{\lambda}{1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_2(z)}dt}]dH(\lambda)$$

$$= -\frac{1}{z^2}\int_{\lambda\in\mathbb{R}} \frac{\lambda^2[\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_1(z)}dt - \int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_2(z)}dt]}{(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_1(z)}dt)(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_2(z)}dt)}dH(\lambda)$$

$$= -\frac{1}{z^2}[\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_1(z)}dt - \int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_2(z)}dt]\int_{\lambda\in\mathbb{R}} \frac{\lambda^2}{(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_1(z)}dt)(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_2(z)}dt)}dH(\lambda)$$

$$= -\frac{1}{z^2}\int_0^1 \frac{y\tau_t'(\tilde{s}_2(z) - \tilde{s}_1(z))}{(1+y\tau_t'\tilde{s}_1(z))(1+y\tau_t'\tilde{s}_2(z))}dt\int_{\lambda\in\mathbb{R}} \frac{\lambda^2}{(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_1(z)}dt)(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_2(z)}dt)}dH(\lambda)$$

$$= (\tilde{s}_1(z)-\tilde{s}_2(z))\frac{1}{z^2}\int_0^1 \frac{y\tau_t'}{(1+y\tau_t'\tilde{s}_1(z))(1+y\tau_t'\tilde{s}_2(z))}dt\int_{\lambda\in\mathbb{R}} \frac{\lambda^2}{(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_1(z)}dt)(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_2(z)}dt)}dH(\lambda)$$

Therefore

$$1 = \frac{y}{z^2}\int_0^1 \frac{(\tau_t')^2}{(1+y\tau_t'\tilde{s}_1(z))(1+y\tau_t'\tilde{s}_2(z))}dt \times \int_0^1 \frac{\lambda^2}{(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_1(z)}dt)(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_2(z)}dt)}dH(\lambda) \tag{4.4.3}$$

As $\tilde{s}_1, \tilde{s}_2 \in Q_1$, $|\int_0^1 \frac{(\tau_t')^2}{(1+y\tau_t'\tilde{s}_1(z))(1+y\tau_t'\tilde{s}_2(z))}dt| \leq \int_0^1 (\tau_t')^2 dt < \infty$.

And $-\frac{1}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_i(z)}dt$, $i = 1, 2$ implies $|\int_0^1 \frac{\lambda^2}{(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_1(z)}dt)(1-\frac{\lambda}{z}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}_2(z)}dt)}dH(\lambda)| \leq \int \lambda^2 dH(\lambda) < \infty$.

So for $z = iv$, with v sufficiently large, 4.4.3 cannot be true. So $s(z)$ is unique.

4.5 Simulation

Although we are working with high-dimensional set up, the computational complexity of the Hayashi's estimator is worth paying attention to. Due to the fact that- time to compute Hayashi's estimtor is much greater compared to Realized covariance estimator, we have not taken a huge dimension along with high sampling frequency to illustrate the theory. We simulate 30 stocks with each 500 observations where spot volatility matrix is I. The empirical and theoretical cdf's are shown in Figure 4.5.1. The difference between theoretical and empirical cdf's of the spectral distribution will be reduced once we increase the dimension and sampling frequency appropriately. One important point to note is that in both our theory and simulation, we have assumed that all the stocks have same number of observation in a day. This is of course not a practical assumption and the difference in sample sizes between stocks can be significant particularly as we are considering high-dimensional set up. In figure 4.5.2, we have plotted the histograms for 60 stocks and sample size 600 and 1000

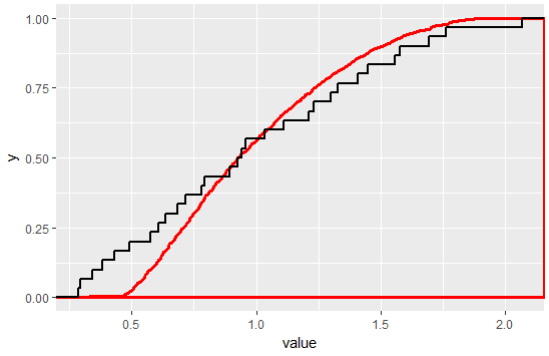


Figure 4.5.1: Plot for empirical cumulative distribution functions and theoretical cumulative distribution functions for Hayashi Yoshida estimator.

respectively (spot volatility matrix is I). Although the dimension and sample size is low, the histogram shows similarity with the Marcenko-Pastur distribution. In Figure 4.5.3, we plot the kernel density plot.

4.6 Conclusion and further direction

In this work we have determined the limiting spectral distribution of high-dimensional Hayashi-Yoshida estimator for nonsynchronous data. We considered a reasonable class of time varying covolatility processes. The limiting spectral distribution is particularly useful to test for deviation from null model, for example, whether the covolatility process is I or not. In this work we have only considered asynchronicity but not the presence of microstructure noise. In presence of noise the spectral distribution may deviate from the ideal situation in significant ways. We have restricted ourselves to the simple Black-Scholes setup. Black-Scholes models are not always very realistic models to describe financial data. One can try to go beyond that and investigate into the spectral analysis of Hayashi's estimator for advanced model. One can also try to extend the results for more general class of time varying covolatility processes. Changes

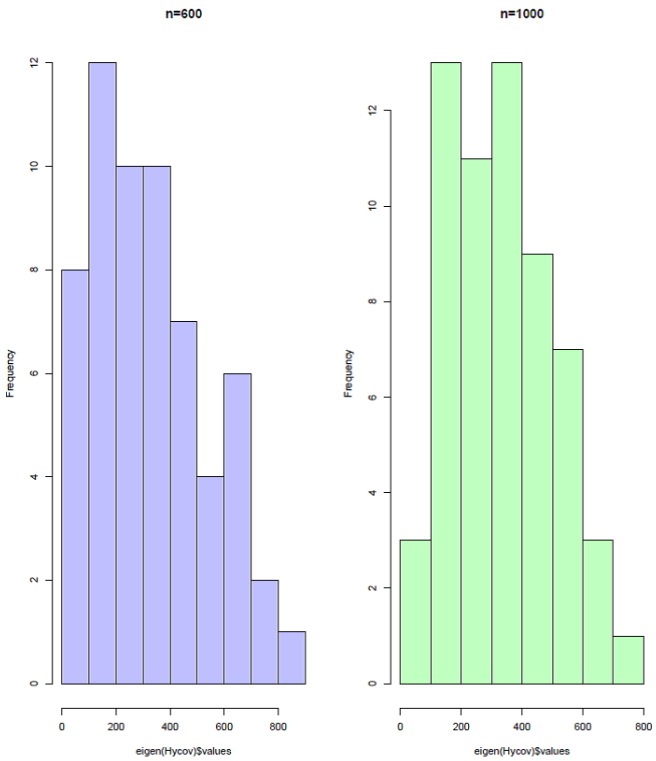


Figure 4.5.2: The histogram of eigenvalues for 60 stocks and sample size 600 (Left) and 1000 (Right).

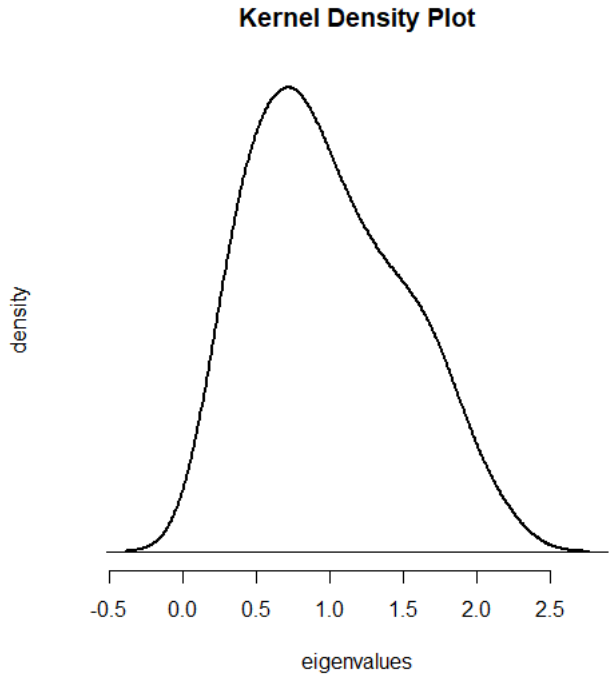


Figure 4.5.3: Kernel density plot of the empirical cumulative distribution function of eigenvalues for 50 stocks with sample size 1000.

due to leverage effects can also be quite serious and so worth looking into.

Chapter 5

SPECTRAL DISTRIBUTION FOR $p/n \to 0$

5.1 Introduction

We have earlier discussed that if $\lim \frac{p}{n} \in (0, \infty)$ then the Hayashi's estimator as well as its eigenvalues fail to be consistent. But when $\lim p/n = 0$, the Hayashi's estimator becomes consistent: $||S^{HY} - \bar{\Sigma}|| \overset{a.s.}{\to} 0$. So the sample eigenvalues will also be consistent for their population counterpart. In this section we are mainly interested in the nontrivial asymptotic distribution of the sample eigenvalues of the Hayashi's estimator. When we have synchronously observed samples from Gaussian distribution with mean 0 and covariance matrix I then the spectral distribution of $\sqrt{\frac{n}{p}}(\frac{1}{n}(S^{RCV} - \bar{\Sigma}))$ is similar to a $p \times p$ Wigner matrix . The first paper exploring spectral distribution of sample covariance

matrix for the case $p/n \to 0$, was by Bai and Yin [51]. Later a series of works have extended and modified the result [52], [53], [54], [55]. Many of the results are proved under the 4th moment condition and the limiting spectral distribution is derived by studying a system of equations determining the LSD [53]. These asymptotic results are also important to derive a test, based on the empirical spectral distribution function [54], for hypothesis testing of cross sectional independence and also to show asymptotic consistency of the test, although a complete discussion on the power properties of many of the resulting tests are yet to be done. Moreover fluctuations of eigenvalues of sample covariance matrix is studied where the asymptotic distribution of jth largest eigenvalue is shown to be convergent to a mixture of rescaled semi-circle law [55]. As discussed in Chapter 1, all these developments were for separable covariance matrix as it is hard to find a general result for non-separable matrix. As Hayashi's estimator is not a separable matrix we have to take help of a separable matrix close enough to the Hayashi's matrix.

In the first chapter we discussed that one strategy of determining the limiting Stieltjes transform is to express the Stieltjes transformation of F_n as

$$s_n(z) = \frac{1}{f(z, s_n(z)) + o(1)},$$

for some function f. Then the the Stieltjes transform of the limiting spectral distribution will be the solution to the equation $s = \frac{1}{f(z,s)}$.

In this chapter, we will take a modified version of the same strategy. We will show that under some condition, for some function f and g.

$$s_n(z) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{f(z, \lambda_i, \beta_n(z)) + o(1)}$$

and

$$\beta_n(z) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{g(z,\lambda_i,\beta_n(z)+o(1))}.$$

($\beta_n(.)$ is a function of z and will be defined later in the proof). So instead of one equation, the Stieltjes transform can be expressed by a system of two equations [1]. .

5.2 Spectral Analysis of High-dimensional Hayashi's Estimator for $p/n \to 0$

Due to the fact that Hayashi's estimator is unbiased we will be concerned about the following matrix (let us call $\tilde{S}$),

$$\tilde{S} = \sqrt{\frac{n}{p}}(S_0/n - \Sigma_p/n).$$

Like the previous chapter here also we have to take some assumptions, we will call it $\mathcal{A}$.

A1: $\sigma_t = \gamma_t \Sigma^{\frac{1}{2}}$ where $\Sigma^{\frac{1}{2}}$ a $p \times p$ symmetric matrix (not time varying) with $\Sigma(= \Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}})$ being a non-nnegative definite matrix and γ_t is a **one-dimensional** time varying *cadlag* process.

A2: $\lim p/n \to 0$ as $p \to \infty$ and $n \to \infty$.

A3: $x'_{ij}s$ $(1 \leq i \leq p,\ 1 \leq j \leq n)$ are iid Gaussian random variables with $E(X_{ij}) = 0$, $E|X_{ij}|^2 = 1$.

A4: $F^{\Sigma_p} \xrightarrow{L} F^H (\neq \delta_{\{0\}})$ as $p \to \infty$ where F^H is a distribution function.

A5: $\exists \tau_l,\ l = 1(1)n$ such that $\mathrm{tr}(\bar{S}_n - \tilde{S})/p \to 0$ with probability 1 and $\mathrm{tr}([(\bar{S}_n - zI)^{-1}(\tilde{S} - zI)^{-1}])/p$ is almost surely bounded where $\bar{S}_n = \sqrt{\frac{n}{p}}(\sum_{l=1}^{n}\tau_l \Sigma_p^{\frac{1}{2}} x_l x_l' \Sigma_p^{\frac{1}{2}} - (\sum_{l=1}^{n}\tau_l)\Sigma)$.

[1] For more details see [48] and [55]. This chapter is going to be a straightforward application of [55]

A6: $||\text{diag}(\tau_1, \tau_2, ..., \tau_n)||$ is bounded above.

A7: $\frac{1}{n}(\sum_{l=1}^{n} \tau_l) \to \tau > 0$ and $\frac{1}{n}(\sum_{l=1}^{n} \tau_l^2) \to \bar{\tau} > 0$ as $n \to \infty$

Now we are ready to state the main theorem.

Theorem 26. *If the above assumptions ($\mathcal{A}$) are true then the empirical spectral distribution of $\sqrt{\frac{n}{p}}(\frac{1}{n}S_0 - \frac{1}{n}\Sigma_p)$ almost surely converges weakly to a nonrandom distribution F as $n \to \infty$, whose Stieltjes transform $s(z)$ is determined by the following system of equations:*

$$\begin{cases} s(z) = & -\int \frac{dF(\lambda)}{z+\bar{\tau}\lambda\beta(z)} \\ \beta(z) = & -\int \frac{\lambda dF(\lambda)}{z+\bar{\tau}\lambda\beta(z)} \end{cases}$$

for any $z \in \mathbb{C}_+$.

Before proving the theorem we will define some quantities and make some observations.

Let $\Sigma = U^*\Lambda U$ be the spectral decomposition of Σ. Now define,

$$W = U\Sigma^{1/2}\{\sum_{l} \tau_l^{\frac{1}{2}} x_l e_l^*\}.$$

Let w_k be the kth row of W and W_k be the matrix after deleting the kth row. Define,

$$
\begin{aligned}
h_k =& \sqrt{\frac{n}{p}} W_k w_k, \\
M =& \sqrt{\frac{n}{p}} (WW^* - (\sum_l \tau_l)\Lambda), \\
M_k =& \sqrt{\frac{n}{p}} (W_k W^* - (\sum_l \tau_l)\Lambda_k), \\
\bar{M}_k =& \sqrt{\frac{n}{p}} (W_k W_k^* - (\sum_l \tau_l)\Lambda_k), \\
t_{kk} =& \sqrt{\frac{n}{p}} (w_k w_k^* - (\sum_l \tau_l)\lambda_k),
\end{aligned}
$$

where Λ_k is the matrix obtained by deleting kth diagonal element of Λ.

Now we will make some remarks. Justifications of the remarks are given in the appendix.

Remark 1: With the mentioned assumption and definition,

$$M = \sum_{k=1}^{p} e_k (h_k + t_{kk} e_k)^*$$

and

$$M = M_k + e_k (h_k + t_{kk} e_k)^*.$$

Remark 2:

$$tr(I + z(M - zI)^{-1}) = \sum_{k=1}^{p} \frac{(h_k + t_{kk} e_k)^* (M_k - zI)^{-1} e_k}{1 + (h_k + t_{kk} e_k)^* (M_k - zI)^{-1} e_k}.$$

Remark 3:

$$(h_k + t_{kk} e_k)^* (M_k - zI)^{-1} e_k = \frac{h_k^* (\bar{M}_k - zI)^{-1} h_k - t_{kk}}{z}.$$

Remark 4.

$$E(h_k^*(\bar{M}_k - zI)^{-1}h_k - (\sum_{l=1}^{n} \tau_l^2)\lambda_k tr((\bar{M}_k - zI)^{-1}\Lambda_k)/np) = 0.$$

The next Lemma is known as McDiarmid Inequality [56].

Lemma 27. *Let $Y_1, Y_2, ..., Y_m$ be independent random vectors taking values in $\mathcal{X}$. Suppose that $f : \mathcal{X}^k \to \mathbb{R}$ is a function of $Y_1, Y_2, ..., Y_m$ satisfying $\forall y_1, y_2, .., y_m, y_i'$,*

$$|f(y_1, y_2, .., y_i, .., y_m) - f(y_1, y_2, .., y_i', .., y_m)| \leq c_i,$$

Then for all $\epsilon > 0$,

$$P(|f(y_1, y_2, ..., y_m) - f(y_1, y_2, ..., y_m)| > \epsilon) \leq 2exp(-\frac{2\epsilon^2}{\sum_{i=1}^{m} c_i^2}).$$

Now we are ready to prove Theorem 26.

Proof. Suppose the Stiletjes transfor of $\bar{S}_n$ is $\bar{s}_n(z)$.

So $\bar{s}_n(z) = \frac{1}{p}\text{tr}(\bar{S}_n - zI)^{-1}$

Observe that,

$$\begin{aligned} z^{-1}I + (\bar{S}_n - zI)^{-1} &= z^{-1}I(\bar{S}_n - zI + zI)(\bar{S}_n - zI)^{-1} \\ &= z^{-1}\bar{S}_n(\bar{S}_n - zI)^{-1}. \end{aligned}$$

$$\begin{aligned}\implies (\bar{S}_n - zI)^{-1} =& z^{-1}\bar{S}_n(\bar{S}_n - zI)^{-1} - z^{-1}I \\ \implies \bar{s}_n(z) =& -z^{-1} + \frac{z^{-1}}{p} tr(\bar{S}_n(\bar{S}_n - zI)^{-1}). \end{aligned} \tag{5.2.1}$$

Notice that,

$$\begin{aligned}\bar{s}_n(z) - \tilde{s}_n(z) &= \frac{1}{p}[tr(\bar{S}_n - zI)^{-1} - tr(\tilde{S}_n - zI)^{-1}] \\ &= \frac{1}{p} tr[(\bar{S}_n - zI)^{-1} - (\tilde{S}_n - zI)^{-1}] \\ &= \frac{1}{p} tr[(\bar{S}_n - zI)^{-1}[\tilde{S}_n - zI - \bar{S}_n + zI](\tilde{S}_n - zI)^{-1}] \\ &= \frac{1}{p} tr[(\tilde{S}_n - \bar{S}_n)(\tilde{S}_n - zI)^{-1}(\bar{S}_n - zI)^{-1}] \\ &\to 0 \text{ a.s.}\end{aligned}$$

The last line of the above derivation is a consequence of Assumption 5 and Cauchy Schwartz inequality (as in Page 91 in the proof of the Theorem in Chapter 4). This means that we can derive valuable information about $\tilde{s}_n(z)$ by studying the spectral distribution of $\bar{S}_n(z)$.

Note that, according to our definitions

$$\begin{aligned}tr(\bar{S}_n - zI)^{-1} &= tr[\sqrt{\frac{n}{p}}(\Sigma^{1/2}\{\sum_l \tau_l^{\frac{1}{2}} x_l e_l^*\})(\Sigma^{1/2}\{\sum_l \tau_l^{\frac{1}{2}} x_l e_l^*\}^* - (\sum_l \tau_l)\Sigma)) - zI]^{-1} \\ &= tr[\sqrt{\frac{n}{p}}(\Sigma^{1/2}\{\sum_l \tau_l^{\frac{1}{2}} x_l e_l^*\})(\{\sum_l \tau_l^{\frac{1}{2}} x_l e_l^*\}^*\Sigma^{1/2} - (\sum_l \tau_l)\Sigma)) - zI]^{-1} \\ &= tr[\sqrt{\frac{n}{p}}U^*(U\Sigma^{\frac{1}{2}}\{\sum_l \tau_l^{\frac{1}{2}} x_l e_l^*\})(\{\sum_l \tau_l^{\frac{1}{2}} x_l e_l^*\}^*\Sigma^{1/2}U^*U - (\sum_l \tau_l)U^*\Lambda U)) - zI]^{-1} \\ &= tr[U^*\sqrt{\frac{n}{p}}\{(WW^* - (\sum_l \tau_l)\Lambda) - zI\}U]^{-1} \\ &= tr[\sqrt{\frac{n}{p}}(WW^* - (\sum_l \tau_l)\Lambda) - zI]^{-1}. \end{aligned} \tag{5.2.2}$$

Now from 5.2.1 we have,

$$\begin{aligned}
\bar{s}_n(z) &= -z^{-1} + \frac{z^{-1}}{p} tr(\bar{S}_n(\bar{S}_n - zI)^{-1}) \\
&= -z^{-1} + \frac{z^{-1}}{p} tr(\sqrt{\frac{n}{p}}(WW^* - (\sum_l \tau_l)\Lambda)(\sqrt{\frac{n}{p}}((WW^* - (\sum_l \tau_l)\Lambda) - zI)^{-1}) \\
&= -z^{-1} + \frac{z^{-1}}{p} tr\{M(M - zI)^{-1}) \\
&= -z^{-1} + \frac{z^{-1}}{p} \sum_{k=1}^{p} (1 - \frac{1}{1 + (h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k}) \qquad \text{by Lemma 36} \\
&= -\frac{z^{-1}}{p} \sum_{k=1}^{p} \frac{1}{1 + (h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k} \\
&= -\frac{z^{-1}}{p} \sum_{k=1}^{p} \frac{1}{1 + \frac{h_k^*(\bar{M}_k - zI)^{-1}h_k - t_{kk}}{z}} \\
&= -\frac{1}{p} \sum_{k=1}^{p} \frac{1}{z + h_k^*(\bar{M}_k - zI)^{-1}h_k - t_{kk}} \\
&= -\frac{1}{p} \sum_{k=1}^{p} \frac{1}{z - t_{kk} + \frac{n\lambda_k(\sum_l \tau_n^2)}{p} tr(\Lambda_k(\bar{M}_k - zI)^{-1}) + \epsilon_{1k}},
\end{aligned}$$

where the last equality is due to Lemma 38 with $E(\epsilon_{1k}) = 0$.

Define $\beta_n = \frac{1}{p} tr(\bar{S}_n - zI)\Sigma)^{-1}$, then similar derivation will lead to

$$\beta_n(z) = -\frac{1}{p} \sum_{k=1}^{p} \frac{\lambda_k}{z + h_k^*(\bar{M}_k - zI)^{-1}h_k - t_{kk}}. \tag{5.2.3}$$

But again,

$$\begin{aligned}
\beta_n(z) &= \frac{1}{p} tr(\bar{S}_n - zI)\Sigma)^{-1} \\
\implies \beta_n(z) &= \frac{1}{p} tr(\bar{S}_n - zI)U^*\Lambda U)^{-1} \\
\implies \beta_n(z) &= \frac{1}{p} tr(M - zI)^{-1}\Lambda) \qquad \text{by 5.2.2} \\
\implies \lambda_k n(\sum_{l=1}^{n} \tau_l^2)\beta_n(z) &= \lambda_k n(\sum_{l=1}^{n} \tau_l^2)\frac{1}{p} tr(\Lambda(M - zI)^{-1}).
\end{aligned}$$

By observation 4, we know that $h_k^*(\bar{M}_k - zI)^{-1}h_k = n(\sum_{l=1}^n \tau_l^2)\lambda_k tr(\Lambda_k(\bar{M}_k - zI)^{-1})/p + \epsilon_{1k}$.

Define, $\epsilon_{2k} = \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)) - h_k^*(\bar{M}_k - zI)^{-1}h_k$.

Therefore from 5.2.3,

$$\begin{aligned}
\beta_n(z) &= -\frac{1}{p}\sum_{k=1}^{p}\frac{\lambda_k}{z + h_k^*(\bar{M}_k - zI)^{-1}h_k - t_{kk}} \\
&= -\frac{1}{p}\sum_{k=1}^{p}\frac{\lambda_k}{z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)) - (t_{kk} + \epsilon_{2k})} \\
&= -\frac{1}{p}\sum_{k=1}^{p}\frac{\lambda_k}{z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z))} \\
&\{-\frac{1}{p}\sum_{k=1}^{p}\frac{\lambda_k}{z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)) - t_{kk} - \epsilon_{2k}} + \frac{1}{p}\sum_{k=1}^{p}\frac{\lambda_k}{z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z))}\}.
\end{aligned}$$

Let us consider the second part of the last term,

$$\begin{aligned}
&\frac{1}{p}\sum_{k=1}^{p}\frac{\lambda_k}{z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z))} - \frac{1}{p}\sum_{k=1}^{p}\frac{\lambda_k}{z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)) - t_{kk} - \epsilon_{2k}} \\
&= \frac{1}{p}\sum_{k=1}^{p}\frac{-\lambda_k(t_{kk} + \epsilon_{2k})}{(z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)))(z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)) - t_{kk} - \epsilon_{2k})} \\
&= \frac{1}{p}\sum_{k=1}^{p}\frac{-\lambda_k(t_{kk} + \epsilon_{2k})}{(z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)))^2} \\
&+ \frac{1}{p}\sum_{k=1}^{p}\frac{-\lambda_k(t_{kk} + \epsilon_{2k})}{(z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)))(z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)) - t_{kk} - \epsilon_{2k})} \\
&- \frac{1}{p}\sum_{k=1}^{p}\frac{-\lambda_k(t_{kk} + \epsilon_{2k})}{(z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)))^2} \\
&= -\frac{1}{p}\sum_{k=1}^{p}\frac{\lambda_k(t_{kk} + \epsilon_{2k})}{(z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)))^2} \\
&+ \frac{1}{p}\sum_{k=1}^{p}\lambda_k(\epsilon_{2k} + t_{kk})\frac{-(t_{kk} + \epsilon_{2k})}{(z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)))^2(z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z)) - t_{kk} - \epsilon_{2k})} \\
&= \epsilon_{3k} + \epsilon_{4k}.
\end{aligned}$$

We will try to argue that $|E(\epsilon_{2k})| \to 0$ as $p \to \infty$,

$$
\begin{aligned}
|E(\epsilon_{2k})| &= |E(\lambda_k \frac{(\sum_{l=1}^n \tau_l^2)}{n} E(\beta_n(z)) - h_k^* \bar{M}_k(z)^{-1} h_k)| \\
&= |E(\lambda_k \frac{(\sum_{l=1}^n \tau_l^2)}{n} E(\beta_n(z)) - E((\sum_{l=1}^n \tau_l^2)\lambda_k tr(\Lambda_k \bar{M}_k(z)^{-1})/np - \epsilon_{1k})| \\
&= |\lambda_k \frac{\sum_{l=1}^n \tau_l^2}{np} E\{tr(\Lambda M(z)^{-1})\} - \lambda_k \frac{\sum_{l=1}^n \tau_l^2}{np} E\{tr(\Lambda_k \bar{M}_k(z)^{-1}) - \epsilon_{1k}| \\
&= |\lambda_k \frac{\sum_{l=1}^n \tau_l^2}{np} E\{tr((M(z)^{-1}\Lambda) - tr(\bar{M}_k(z)^{-1}\Lambda_k)\} - \epsilon_{1k}| \\
&= |\lambda_k \frac{\sum_{l=1}^n \tau_l^2}{np} E\{tr((M(z)^{-1}\Lambda) - tr(\bar{M}_k(z)^{-1}(\Lambda - \lambda_k e_k e_k^*))\} - \epsilon_{1k}| \\
&= |\lambda_k \frac{\sum_{l=1}^n \tau_l^2}{np} E\{tr((M(z)^{-1} - \bar{M}_k(z)^{-1})\Lambda) + \lambda_k E(\bar{M}_k(z)^{-1} e_k e_k^*)\} - \epsilon_{1k}| \\
&= |\lambda_k \frac{\sum_{l=1}^n \tau_l^2}{np} E\{tr((M(z)^{-1} - \bar{M}_k(z)^{-1})\Lambda) + \lambda_k e_k^* \bar{M}_k(z)^{-1} e_k)\} - \epsilon_{1k}| \\
&= |\lambda_k \frac{\sum_{l=1}^n \tau_l^2}{np} E\{tr((M(z)^{-1} - \bar{M}_k(z)^{-1})\Lambda) - \frac{\lambda_k}{z})\} - \epsilon_{1k}| \\
&\leq \lambda_k \frac{\sum_{l=1}^n \tau_l^2}{np} |Etr((M(z)^{-1} - \bar{M}_k(z)^{-1})\Lambda)| + |\frac{\lambda_k^2 (\sum_{l=1}^n \tau_l^2)}{znp}| + \lambda_k \frac{\sum_{l=1}^n \tau_l^2}{np} |\epsilon_{1k}|.
\end{aligned}
$$

It can be shown that the above expression is $o(p)$ (see Appendix).

$$
\begin{aligned}
E(|\epsilon_{2k}+t_{kk}|^2) &= E|\epsilon_{2k}+t_{kk}-E(\epsilon_{2k})|^2+|E(\epsilon_{2k})|^2 \\
&= E|\lambda_k\frac{(\sum_{l=1}^n\tau_l^2)}{n}E(\beta_n(z))-h_k^*\bar{M}_k(z)^{-1}h_k+t_{kk} \\
&- E(\lambda_k\frac{(\sum_{l=1}^n\tau_l^2)}{n}E(\beta_n(z))-h_k^*\bar{M}_k(z)^{-1}h_k)|^2+|E(\epsilon_{2k})|^2 \\
&= E|-h_k^*\bar{M}_k(z)^{-1}h_k+t_{kk}+h_k^*\bar{M}_k(z)^{-1}h_k)|^2+|E(\epsilon_{2k})|^2 \\
&= E|-h_k^*\bar{M}_k(z)^{-1}h_k+t_{kk}+(\sum_{l=1}^n\tau_l^2)\lambda_k trE(\Lambda_k\bar{M}_k(z)^{-1})/np|^2+|E(\epsilon_{2k})|^2 \\
&= E|-h_k^*\bar{M}_k(z)^{-1}h_k+t_{kk}+(\sum_{l=1}^n\tau_l^2)\lambda_k trE(\Lambda_k\bar{M}_k(z)^{-1})/np|^2 \\
&+ E|(\sum_{l=1}^n\tau_l^2)\lambda_k trE(\Lambda_k\bar{M}_k(z)^{-1})/np-(\sum_{l=1}^n\tau_l^2)\lambda_k tr(\Lambda_k\bar{M}_k(z)^{-1})/np|^2+|E(\epsilon_{2k})|^2.
\end{aligned}
$$

It can be shown that all three terms is $o(p)$ (See Appendix).

We further observe that

$$
|z+\lambda_k\frac{(\sum_{l=1}^n\tau_l^2)}{n}E(\beta_n(z))|\geq Im(z+\lambda_k\frac{(\sum_{l=1}^n\tau_l^2)}{n}E(\beta_n(z)))\geq Im(z)=v>0
$$

and

$$
|z+\lambda_k\frac{(\sum_{l=1}^n\tau_l^2)}{n}E(\beta_n(z))-t_{kk}-\epsilon_{2k}|\geq|z+h_k^*\bar{M}_k^{-1}h_k-t_{kk}|\geq v.
$$

This implies $\epsilon_{3k}\to 0$ a.s.

With this we showed that

$$
E(\beta_n(z))=-\frac{1}{p}\sum_{k=1}^p\frac{\lambda_k}{z+\lambda_k\frac{(\sum_{l=1}^n\tau_l^2)}{n}E(\beta_n(z))}+d_n,
$$

where $d_n\to 0$ as $n\to\infty$.

If we replace $(x_{1i},...,x_{pi})$ by $(x'_{1i},...,x'_{pi})$ and call the resulting new $\bar{s}_n$ by s'_n

(similarly β_n by $\beta_n'(z)$) then it is easy to show that

$$|\bar{s}_n(z) - s_n'(z)| \leq \frac{c_1}{pv}$$

and

$$|\bar{\beta}_n(z) - \beta_n'(z)| \leq \frac{c_1}{pv}.$$

Hence by Lemma 36,

$$\sum_n P(|\bar{s}_n(z) - s_n'(z)| > \delta) < \infty$$

and

$$\sum_n P(|\bar{\beta}_n(z) - \beta_n'(z)| > \delta) < \infty.$$

By Borel-Cantelli lemma, $\bar{s}_n(z) - E(\bar{s}_n) \to 0$ a.s. and $\beta_n(z) - E(\beta_n(z)) \to 0$ a.s. and thus we also have $\tilde{s}_n(z) - E(\bar{s}_n) \to 0$.

Now as $max_k|\lambda_k| \leq a_0$ and $|z + \lambda_k n(\sum_{l=1}^n \tau_l^2)E(\beta_n(z))| \geq v$ is bounded (see appendix), $E(\beta_n(z))$ is bounded. So by dominated convergence theorem $E(\beta_n(z))$ converges to $\beta(z)$. But as $\beta_n(z) - E(\beta_n(z)) \to 0$ a.s., we have $\beta_n(z) \to \beta(z)$ a.s. Similarly for $s_n(z) \to s(z)$ a.s. So $s_n(z)$ can be evaluated by the following two equations,

$$s(z) = -\int \frac{dF^{\Sigma}(\lambda)}{z + n(\sum_l \tau_l^2)\lambda E(\beta(z))}$$

$$\beta(z) = -\int \frac{\lambda dF^{\Sigma}(\lambda)}{z + n(\sum_l \tau_l^2)\lambda E(\beta(z))}.$$

□

Chapter 6

Conclusion

In this thesis, we have seen how asynchronicity can affect the statistical analysis of nonsynchronous intraday data. We showed that for the Elliptical copula a significant amount of underestimation of interstock dependence can occur if proper correction is not made. In chapter 2 we have proposed a method to pair the nonsynchronous observations while keeping a memory of their original arrival times. These memories helped us to construct a consistent estimator of the copula parameter. Through simulation, we have shown the amount of gain we have made in terms of mean-square-error. We also have determined a large sample distribution for the proposed corrected estimate. But as discussed mere consistency of the copula parameter does not imply consistency of the empirical copula function. We have shown the consistency of copula function. One important question to address is how much data is getting lost or deleted in the process of paring. We proved a theorem that determines the amount of loss of intraday data after applying the pairing algorithm when the original arrival processes are independent Poisson. Then we investigated the effect of asynchronicity for a more general copula. To develop the appropriate correction

factor, we looked into how Kendall's Tau is affected by asynchronicity. If we can measure the exact amount of this influence, we can then use the relation between copula and Kendall's tau to correct the copula parameter appropriately. Unfortunately, we couldn't get the exact relation between true and uncorrected Kendall's Tau. However under some assumptions, we proved some important theorem and through simulation and heuristic derivation, we proposed a correction factor. We then showed the superiority of this corrected estimate over the uncorrected one by simulation studies on some important Archimedean copula.

Then we looked at the high-dimensional case. As discussed, one of the unbiased estimators of Integrated covariance matrix is the estimator developed by Hayashi and Yoshida. When the dimension is high, the good properties of this estimator is no longer preserved. We have studied the limiting spectral distribution of this matrix for both the cases when dimension to sample size ratio converges to a zero or a nonzero number. When it converges to 0, the sample eigenvalues are consistent to their population counterpart (unlike for the other case). So we had to establish a nontrivial large sample distribution for this case. Limiting spectral distribution help us to test significant statistical hypotheses, some are discussed in detail in [57]. Some other financial application of the theory, such as in portfolio optimization, can be found in [58]. It is also used for high-dimensional clustering in the context of genetics but also can be adopted for investigating relevant aspects of financial data [59].

There are several scopes of future research. It is worthwhile to investigate how the theorems of chapter 4 and 5 are affected by unequal sample sizes of the stocks especially when stocks with different liquidity are present. One can study the spectral distribution for more general time varying covolatility process. The effect of asynchronicity on tail probabilities and other parameters (such as Spearman's rho) can be studied as well. In the thesis only the impact of

asynchronicity is studied, the presence of microstructure noise is not assumed. One can extend all these works in presence of both microstructure noise and asynchronicity.

Chapter 7

Appendix

7.1 Introduction to copula:

Theorem 28. *Let $F_{1,2,...,d}$ be a d dimensional distribution function with marginal distribution functions $F_1,\ F_2...,\ F_d$. Then there exists a copula such that for all $x \in \mathbb{R}^d$,*

$$F_{1,2,...,d}(x_1, x_2, ..., x_d) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d)).$$

If $F_1, ..., F_d$ are all continuous functions then C is unique. Otherwise, C is uniquely determined on the Cartesian product of the ranges of the marginal distribution functions $\mathrm{Ran}(F_1)\times\mathrm{Ran}(F_2)\times....\times\mathrm{Ran}(F_d)$. *Conversely, if $F_1.F_2, ..., F_d$ are distribution functions and C is a copula then the function $F_{1,2,...,d}$ defined above is a d dimensional multivariate distribution function with margins $F_1, F_2, ..., F_d$.*

The density of a copula (denoted by c) C is defined as

$$c(u_1, ..., u_d) = \frac{\partial C(u_1, u_2, ..., u_d)}{\partial u_1 \partial u_2 ... \partial u_d}.$$

Then the d dimensional density of the joint distribution is given by

$$f(x_1, x_2, ..., x_d) = c(F_1(x_1), F_2(x_2), ..., F_d(x_d)) \prod_{i=1}^{d} f_n(x_i),$$

where f_n is the density of margin F_n.

7.1.1 Some examples of copula

Two most well known copula families are the Elliptical copula family, Archimedean copula family.

7.1.1.1 Elliptical copula family

Definition 29. A d-dimensional random vector z is said to have an elliptical distribution ($z \sim EC_d(\mu, \Sigma, g)$) with parameters μ ($d \times 1$) and Σ ($d \times d$) if it has a stochastic representation

$$z \stackrel{d}{=} \mu + rAu,$$

where $r \geq 0$ is a random variable and u is uniformly distributed on a unit sphere in $\mathbb{R}$. A is a constant matrix such that $\Sigma = AA'$.

If r has a density then density of z is of the following form

$$|\Sigma|^{-\frac{1}{2}} g((z-\mu)\Sigma^{-1}(z-\mu)),$$

where g is a scale function uniquely determined by distribution of r.

For our purpose without loss of generality we can consider $z \sim EC_d(0, R, g)$, where R is the correlation matrix. In this case all the marginal distributions of z are identical with pdf

$$q_g(x) = \frac{\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d-1}{2})} \int_{x^2}^{\infty} (y - x^2)^{\frac{d-1}{2}-1} g(y) d(y).$$

Elliptical copula has the following form

$$C_R(u_1, u_2, ..., u_d) = F_R(F^{-1}(u_1), F^{-1}(u_2), ..., F^{-1}(u_d)),$$

where $F_R(x_1, x_{2,}...., x_d)$ is a d-dimensional elliptical distribution function with correlation matrix R and $F(.)$ is the marginal distribution of F_R[60].

Two important and widely used elliptical copulas are Gaussian copula and Student's t copula. In Gaussian copula,

$$C_R(u) = \Phi_R^d(\Phi^{-1}(u_1), \Phi^{-1}(u_2), ..., \Phi^{-1}(u_d)),$$

where $\Phi_R^d(.)$ is a d-dimensional Gaussian distribution function with correlation coefficient R and $\Phi(.)$ is univariate Gaussian distribution function.

For Student's t copula,

$$C_{\nu,R}(u) = t_{\nu,R}^d(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_d)),$$

where $t_{\nu,R}^d$ is the joint distribution of d dimensional t distribution with correlation matrix R with degrees of freedom ν and t_ν is univariate t distribution with same degrees of freedom. ν controls the heaviness of the tails and as $\nu \to \infty$ it behaves like a Gaussian copula.

It is important to note that it is not necessary that a multivariate distribution with Gaussian (or t) copula should have Gaussian (or t) margins. For example, it is straight forward to see that (from Sklar's theorem) a $d-$dimensional multivariate distribution with distribution function F can have different margins F_1,F_2,....,F_d and a Gaussian copula with dependence R i.e.

$$\begin{aligned} F(x_1, x_2, ..., x_d) &= C_R(F_1(x_1), ..., F_d(x_d)) \\ &= \Phi_R^d(\Phi^{-1}(F_1(x_1)), \Phi^{-1}(F_2(x_2)), ..., \Phi^{-1}(F_d(x_d))). \end{aligned}$$

Sometimes it is also called meta-elliptical copula family [33].

7.1.1.2 Archimedean copula family

Definition 30. An Archimedean copula [61] is defined as the following-

$$C(u_1, u_2, ..., u_d) = \begin{cases} \phi^{-1}(\phi(u_1) + ... + \phi(u_d)) & \text{if } \sum_{i=1}^{d} \phi(u_i) \leq \phi(0) \\ 0 & 0 \text{ otherwise,} \end{cases}$$

where $\phi(u) : [0, 1] \to [0, \infty)$ function with $\phi(1) = 0$, $\phi'(u) < 0$ and $\phi''(u) > 0$ for all $0 \leq u \leq 1$. $\phi(u)$ is called the generator of a copula.

Archimedian copula has many desirable properties, for example it is symmetric and associative. Some examples of Archimedean copulas are- Gumbel copula, for which $\phi(u) = (-ln(u))^{\alpha}$ and Joe copula, for which $\phi(u) = (-ln1 - (1-u)^{\alpha})$.

Copula	$C(u,v)$	α
Clayton	$max\{[u^{-\alpha}+v^{-\alpha}-1]^{-\frac{1}{\alpha}},0\}$	$\alpha\in[-1,\infty)-\{0\}$
Gumbel	$e^{-((-log(u))^{\alpha}+(-log(v))^{\alpha})^{\frac{1}{\alpha}}}$	$\alpha\in[1,\alpha)$
Frank	$-\frac{1}{\alpha}log(1+\frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{e^{-\alpha}-1})$	$\alpha\in\mathbb{R}-\{0\}$
Ali-Mikhail-Haq	$\frac{uv}{1-\alpha(1-u)(1-v)}$	$\alpha\in[-1,1]$
Joe	$1-[(1-u)^{\alpha}+(1-v)^{\alpha}-((1-u)(1-v))^{\alpha}]^{\frac{1}{\alpha}}$	$\alpha\in[1,\infty)$
Independent	uv	

Table 7.1: Different Archimedean copulas with the corresponding parameter range.

Copula	$\phi(t)$	$\phi^{-1}(t)$
Clayton	$\frac{1}{\alpha}(t^{-\alpha}-1)$	$(1+\alpha t)^{-\frac{1}{\alpha}}$
Gumbel	$(-log(t))^{\alpha}$	$e^{-t^{1/\alpha}}$
Frank	$-log(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1})$	$-\frac{1}{\alpha}log(1+e^{-t}(e^{-\alpha}-1))$
Ali-Mikhail-Haq	$log(\frac{1-(1-t)\alpha}{t})$	$\frac{1-\alpha}{e^{t}-\alpha}$
Joe	$-log(1-(t-1)^{\alpha})$	$1-(1-e^{-t})^{\frac{1}{\alpha}}$
Independent	uv	

Table 7.2: Different Archimedean copulas with the generators.

7.1.2 Estimation

The precision of copula estimation can be greatly influenced by the specifications of marginals. Well-specified marginals can greatly improve the inference. On the other hand, misspecification can cause a great deal of harm. If we have no prior information a nonparametric approach for estimation would be preferable. The probabilities $P(X > F_X^{-1}(p), Y > F_Y^{-1}(p))$, which is an important quantity in the context of stock market dynamics, can be greatly underestimated if either the marginals or the dependence structure is misspecified.

But if we have a good prior idea about the marginals parametric models are easy to deal with. If we have reasons to believe that the margins and the copula are absolutely continuous with corresponding densities, maximum likelihood estimation is the first method of estimation available to us. But

maximum likelihood estimation of copula to find the optimal set of parameters can be computationally very expensive. Joe and Xu proposed a method, called IFM (Inference for the margins) [62], to reduce the complexity of the maximum likelihood estimate. In this approach, first, the marginal parameter is calculated and then the copula parameter but not simultaneously. If the parameter of marginal distributions is θ_1 and copula parameter is θ_2 then in IFM we estimate θ_1 and θ_2 in the following manner:

$$\hat{\theta}_1 = \text{argmax}_{\theta_1} \sum_{i=1}^{n} \sum_{d=1}^{p} \log\ f_d(x_{id}; \theta_1)$$
$$\hat{\theta}_2 = \text{argmax}_{\theta_2} \sum_{i=1}^{n} \log\ c(F_1(x_{1i}), F_2(x_{2i}), ..., F_d(x_{di})),$$

where p is the number of variables and n is the number of observations. Although the IFM estimate is different from MLE, the estimator is also a consistent estimator.

An alternative to IFM estimation is MPLE or maximum pseudo likelihood estimation [63], where the jth margin is estimated nonparametrically by its empirical distribution function $\hat{F}_{n,j}, j \in \{1, .., p\}$. So the copula parameters are estimated based on the pseudo observations

$$\hat{U}_{ij} = \frac{n}{n+1} \hat{F}_{n,j}(X_{ij}) = \frac{R_{ij}}{n+1}.$$

Estimations can also be accomplished using Kernel methods [64].

7.2 Limiting Spectral Analysis of Large Dimensional Random Matrices

The study of the spectra of large dimensional random matrices first started in nuclear physics during the 1950s where the unobservable energy levels of quantums were characterized by the eigenvalues of a matrix of observable quantities. The first pioneering work is Wigner's semicircle law for a Gaussian matrix [65], where the limiting spectral distribution of a large dimensional Wigner matrix tends to the semi-circle law.

To see what Random matrix theory tries to achieve let us start with some basic definitions of Random Matrix Theory (RMT) [1]:

Definition 31. (Empirical Spectral Distribution) For any square matrix A, the probability distribution P which puts equal mass on each eigenvalue of A is called the empirical spectral distribution (ESD) of A and is denoted as F^A.

If λ is an eigenvalue of a $p \times p$ matrix A_p of multiplicity m, then ESD puts mass m/p on λ.

Definition 32. (Limiting Spectral Distribution) Let $\{A_n\}$ be a sequence of random matrices with corresponding ESD $\{P_n\}$. The limiting spectral distribution (LSD) of the sequence is defined as the limit of $\{P_n\}$, if it exists.

If A_n's are random then the limit is understood to be in some probabilistic sense (almost surely/ in probability). As mentioned, the first breakthrough in

[1] For comprehensive background on RMT, see [48], [66]

Random matrix theory came by establishing the limiting spectral distribution of the Wigner matrix. Let us formally define and state the related results.

Definition 33. (Wigner Matrix) A Wigner matrix of order p and scale parameter σ is a Hermitian matrix of order p, whose entries above the diagonal are iid complex random variables with 0 mean and variance σ^2 and the diagonals are iid real random variables [2].

The Wigner matrix also has a statistical meaning. Suppose $(x_1, x_2, .., x_n)$ are iid samples drawn from $N_p(\mu, I_p)$. Then, the sample covariance matrix is

$$S_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})',$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. As $n \to \infty$, $S_n \to I_p$ and $\sqrt{n}(S_n - I_p) \xrightarrow{d} \sqrt{p}W_p$.

The limiting spectral distribution of $p^{-\frac{1}{2}}W_p$ is a semi circle law (with scale parameter σ) with the density,

$$f_\sigma(s) = \begin{cases} \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - s^2}, & |\mathrm{s}| \le 2\sigma, \\ 0, & \text{otherwise.} \end{cases}$$

The result has been generalized by Arnold [67] and Grenander [68][69].

A generalization of the above result for non iid case is the following:

Theorem 34. *Suppose that $W_p = \frac{1}{p}X_p$ is a Wigner matrix and the entries above or on the diagonal of X_p are independent but may depend on p and not necessarily identically distributed. Assume that all the entries of X_p (x_{jk}^p) are*

[2] Variance of a complex random variable Z is defined as $E|Z - E[Z]|^2$

of mean 0 *and variance* 1 *satisfying the condition that, for any constant* $\eta > 0$,

$$\lim_{p\to\infty} \frac{1}{p^2} \sum_{j,k} E|x_{jk}^{(p)}|^2 I(|x_{jk}^{(p)}| > \eta\sqrt{p}) = 0,$$

then the ESD of W_p *converges to the semicircle law almost surely.*

The condition in Theorem 4 is also necessary for the conclusion [70].

From a statistical point of view, the most important random matrix to look at is the sample covariance matrix

$$\tilde{S} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(x_k - \bar{x})^*.$$

As $\bar{x}\bar{x}^*$ is a rank 1 matrix, we actually can remove $\bar{x}$ from the expression. This is due to a straightforward consequence of the following theorem (taking $A = 0$, $B = I$, $S = C$, $\tilde{S} = D$).

Theorem 35. *Let* A *and* B *are two Hermitian matrices of orders* $p \times p$ *and* $n \times n$ *and* C *and* D *are two* $p \times n$ *complex matrices. Then*

$$||F^{A+DBD^*} - F^{A+CBC^*}|| \leq \frac{1}{p}\text{rank}(D - C).$$

Therefore it is enough to study the spectral distribution of

$$S = \frac{1}{n-1} \sum_{k=1}^{n} x_k x_k^*.$$

Here n denotes the number of observations and p denotes the number of variables. In order to establish the limiting spectral distribution of the sample

covariance matrix, we need to take into account both the size of p and n or more importantly the relative size of n and p.

At this point, we should emphasize that all the sample covariance matrices that have been studied so far are of separable form. Here, below, we will give the precise definition of separable covariance matrix.

Definition 36. For a $p \times p$ positive definite matrix Σ_a and a $n \times n$ positive definite matrix Σ_b, consider a random $p \times n$ matrix X , where X_{ij}'s are centered *i.i.d* random variables such that $E(X_{ij}) = 0$, $E|X_{ij}|^2 = \frac{1}{n}$ and $\sqrt{n}X_{ij}$ are bounded in any L_p space, uniformly in n, i, j, then we call $\Sigma_a^{\frac{1}{2}} X \Sigma_b X^* \Sigma_a^{\frac{1}{2}}$ a separable sample covariance matrix.

It means when a random matrix can be rearranged using the *vec* operator (vectorizing elements of a matrix) that transforms a $n_1 \times n_2$ matrix into a column vector of size $n_1 n_2$, the variance-covariance matrix of the vector is expected to be expressed through a Kronecker product of two variance-covariance matrices-each capturing the variability in one dimension (over time and over cross section) [71]. For example, if we want to measure prices of two stocks at three time points by a matrix:

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{pmatrix},$$

then the variance-covariance matrix of the vectorized matrix will be $\Sigma^{6\times 6} = \Sigma_1^{2\times 2} \otimes \Sigma_2^{3\times 3}$ where Σ_1 captures cross sectional variation and $\Sigma_2$captures variation due to time.

Inference in separability of covariance matrix [3]is well studied in statistics and

[3]The problem with asynchronicity is that it does not naturally allow us to take separability assumption (in chapter 4, we will see that our matrix of concern does not lead to a separable covariance matrix). To avoid that problem we will take some assumptions to be introduced later.

in most of the cases the test statistics needs special structural or distributional assumptions [72] [73] [74] [75] [76] [72] [77] [78].

The first result on the LSD of sample covariance matrix was established by Marcenko and Pastur [79] [80] [81] for the case $p/n \to c > 0$ as $p \to \infty$, $n \to \infty$.

Theorem 37. *Suppose that $\{x_{ij}\}$ are iid random variables with mean 0 and variance σ^2. Assume that $p/n \to c \in (0, \infty)$ and define $a = \sigma^2(1 - \sqrt{c})^2$ and $b = \sigma^2(1 + \sqrt{c})^2$. Then with probability 1, F^S tends to the Marcenko Pastur Law (M-P Law) with density-*

$$p_c(x) = \begin{cases} \frac{1}{c} \times \frac{1}{2\pi x\sigma^2}\sqrt{(b-x)(x-a)} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

and a point mass $1 - \frac{1}{c}$ at the origin if $c > 1$.

If $\sigma^2 = 1$ then the distribution is called the standard Marcenko Pastur law.

It is clear from the result that the existence of the second order moment of the entries is necessary for the result to hold.

The result can be extended to the appropriate form for the case where the entries $x_{ij}'s$ depend on n and for each n they are independent but not necessarily identically distributed and for the case where $x_{ij}'s$ are indeed not dependent[66].

It is also worthwhile to look at the spectral distribution when p is large ($p \to \infty$) but small compared to n ($\frac{p}{n} \to 0$). The following theorem [51] addressed this concern and determines the nontrivial limiting spectral distribution:

Theorem 38. *If $p/n \to 0$ then the empirical spectral distribution of $W_n = \sqrt{\frac{n}{p}}(S_n - \sigma^2 I_p)$ converges almost surely to the semicircle law with scale index*

σ^2.

This result was also extended to a more general setting. Like Theorem 7, we need some extra assumptions to be taken in order to extend it for non-iid case.

Theorem 39. *Suppose that for each n, the entries of X_n are independent and complex random variables with a common mean and variance σ^2. Assume that for each constant $\delta > 0$, as $p \to \infty$ with $p/n \to 0$,*

$$\frac{1}{p\delta^2}\sqrt{np}\sum_{jk}((x_{jk}^{(n)})^2 I(|x_{jk}^{(n)}| \geq \delta(np)^{\frac{1}{4}})) = o(1)$$

and

$$\frac{1}{np^2}\sum_{jk}((x_{jk}^{(n)})^4 I(|x_{jk}^{(n)}| \geq \delta(np)^{\frac{1}{4}})) = o(1),$$

then with probability 1, the ESD of $W_n = \sqrt{\frac{n}{p}}(S_n - \sigma^2 I_p)$ tends to the semi-circlular law with scale index σ^2.

After establishing the limiting spectral distribution of a particular matrix, the next big problems are the convergence rate and bounds of extreme eigenvalues. It is proved that the largest eigenvalue of a sample covariance matrix converges almost surely to a limit under the existence of fourth moment [82] (which is also a necessary condition [83]). Almost sure limit of the smallest eigenvalue of the sample covariance matrix is studied in [84], [85]. The problem of convergence rate is relatively more challenging. Even for Wigner's matrix, the problem was open for decades up until 1993 [48]. In this thesis, we will not discuss these two problems. We will concentrate on finding the limiting spectral distribution. The main interest to determine the limiting spectral distribution

is that it helps us to develop a test for special structure of covariance matrix.

7.2.1 Stieltjes transform method

The eigenvalues of a matrix are a function of its entries. But as there is no closed form expression of the function (particularly when the dimension is more than 4) we need some other methods to look into the eigenvalues. There are three methods that can be used to find the distribution of eigenvalues.

1. The moment method
2. Stieltjes Transform method
3. Orthogonal Polynomial decomposition.

In this thesis, we will work with Stieltjes transform only. So now we will give a brief outline of the Stieltjes transform method.

Definition 40. If $G(x)$ is a function of bounded variation on real line, then its Stieltjes transform is defined by

$$s_G(z) = \int \frac{1}{\lambda - z} dG(\lambda), \quad z \in \{z \in \mathbb{C} : Im(z) > 0\}.$$

Stieltjes transform is an important tool in RMT because of its one-one relationship with the distribution function (ESD in our context). Particularly in RMT, the significance of Stieltjes transform relies on the fact that to determine the limiting spectral distribution one only needs to show the convergence

of corresponding Stieltjes transform. These are captured in the following two theorems [48].

Theorem 41. *For any continuity points $a < b$ of G, we have*

$$G\{[a,b]\} = \lim_{\epsilon \to 0^+} \frac{1}{\pi} \int_a^b Im(s_G(x + i\epsilon))dx.$$

Theorem 42. *Assume that $\{G_n\}$ is a sequence of functions of bounded variations and $G_n(-\infty) = 0$ for all n. Then,*

$$\lim_{n\to\infty} s_{G_n}(z) = s(z) \quad \forall z \in \{z \in \mathbb{C} : Im(z) > 0\},$$

if and only if there is a function of bounded variation G with $G(-\infty) = 0$ and Stieltjes transform $s(z)$ and such that $G_n \to G$ vaguely.

If C be a $p \times p$ Hermitian matrix and $F_n$be its ESD, then elementary calculation shows that the Stieltjes transform of G_n is given by

$$\begin{aligned} s_n(z) &= \int \frac{1}{x - z} dG_n(x) \\ &= \frac{1}{p} tr(C - zI)^{-1} \\ &= \frac{1}{p} \sum_{i=1}^{p} \frac{1}{c_{ii} - z - c_i^*(C_i - zI)^{-1} c_i}, \end{aligned}$$

where C_i is the $(p-1) \times (p-1)$ matrix obtained by removing ith row and column from C and c_i is the ith column vector of C after removing the ith

element.

One standard technique of establishing limit of Stieltjes transform is to show $c_{ii} - z - c_i^*(C_i - zI)^{-1}c_i$ to be $f(z, s_n(z)) + o(1)$ for some function f. Then the LSD exists and is the solution to the equation

$$s = \frac{1}{f(z, s)}.$$

The limiting spectral distribution is particularly useful for testing of statistical hypothesis for high-dimensional data. Sometimes it is necessary to test for a specific structure of correlation matrix. The limiting spectral distribution helps us to construct a test statistics in such situation [55]. Another example of use of limiting spectral distribution is the following [86]. Consider the testing problem of the means of two multivariate populations with dimension p. When p is large the classical Hotelling test is no more perform well. An alternative test has been proposed and the limiting spectral distribution is used to show that the test is more powerful than the Hotelling test [87].

The concepts described in this section will be used in chapters 4 and 5 to obtain the limiting spectral distribution of the high-dimensional covariance matrix obtained from non-synchronous financial data.

7.3 Chapter 5

In chapter 5, we claimed that it is enough to prove the theorem for $z = iv$, where $v > 0$ is suficiently large. We will provide the justification here below.

We have seen that

$$1.\ \mathrm{Re}(-\frac{1}{\mathrm{iv}}\int_0^1 \frac{\tau_\mathrm{t}'}{1+\mathrm{y}\tau_\mathrm{t}'\tilde{\mathrm{s}}(\mathrm{z})}\mathrm{dt}) \geq 0$$

$$2.\ |\frac{1}{1-\frac{\lambda}{iv}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt}| \leq 1 \quad \forall \lambda \geq 0$$

$$3.\ |\frac{1}{iv}\int_0^1 \frac{\tau_t'}{1+y\tau_t'\tilde{s}(z)}dt| \leq \frac{1}{v}\int_0^1 \tau_t' dt.$$

Applying dominated convergence theorem, we see that $ivs(iv) \to -1$ as $v \to \infty$.

Now from Lemma 36, we can conclude the justification.

Proof of Lemma 33:

$$\begin{aligned}
&\frac{1}{p}\mathrm{max_l}|x_l'\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}-zI)^{-1}\Sigma^{\frac{1}{2}}x_l - x_l'\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(\Sigma\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l}-zI)^{-1}\Sigma^{\frac{1}{2}}x_l| \\
&= \frac{1}{p}\mathrm{max_l}|x_l'\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(-\frac{1}{z})(-\frac{1}{z}\Sigma\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}+I)^{-1}\Sigma^{\frac{1}{2}}x_l \\
&- x_l'\Sigma^{\frac{1}{2}}(S_l-zI)^{-1}(-\frac{1}{z})(-\frac{1}{z}\Sigma\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l}+I)^{-1}\Sigma^{\frac{1}{2}}x_l| \\
&\leq \frac{1}{p}\mathrm{max_l}|-\frac{1}{z}\sum_l\frac{\tau_l}{1+\tau_l a_l}+\frac{1}{z}\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l}||\Sigma^{\frac{1}{2}}x_l|^2 Cp^{\delta}.
\end{aligned}$$

The last inequality is due to Lemma 1 (where $q=\Sigma^{\frac{1}{2}}x_l$, $B=(S_l-zI)^{-1}$,$A=\Sigma$, $w_1=-\frac{1}{z}\sum_{l=1}^{n}\frac{\tau_l}{1+\tau_l a_l}$, $w_2=-\frac{1}{z}\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l}$), and Assumption 7.

Applying Markov Inequality and Lemma 28,

$$\begin{aligned}
P(|x_l'\Sigma x_l - tr(\Sigma)| \geq p\epsilon) &\leq \frac{E|x_l'\Sigma x_l - \mathrm{tr}(\Sigma)|^{2k}}{(p\epsilon)^{2k}} \\
&\leq \frac{C_k p^k ||\Sigma||^{2k}}{(p\epsilon)^{2k}} \\
&\leq \frac{CC_k p^k p^{2\delta k}}{(p\epsilon)^{2k}}.
\end{aligned}$$

Choosing $k > \frac{2}{1-2\delta}$, and using Borel Cantelli lemma (try to include the sum

in the above equation), we get $\max_l |x_l'\Sigma x_l - tr(\Sigma)|/p \to 0$ a.s. As a consequence $\max_l |\Sigma^{\frac{1}{2}} x_l|^2/p < M \forall n > n_0$ for some n_0.

Define, $c = tr(\Sigma^{\frac{1}{2}}(S - zI)^{-1}\Sigma^{\frac{1}{2}})$, and consider

$$\max_l p^\delta |\sum_l \frac{\tau_l}{1+\tau_l a_l} - \sum_{j\neq l} \frac{\tau_j}{1+\tau_j b_j^l}|.$$

$$= \max_l p^\delta |\sum_l \frac{\tau_l}{1+\tau_l a_l} - \sum_l \frac{\tau_l}{1+\tau_l c} + \sum_l \frac{\tau_l}{1+\tau_l c} - \sum_{j\neq l} \frac{\tau_j}{1+\tau_j c} + \sum_{j\neq l} \frac{\tau_j}{1+\tau_j c} - \sum_{j\neq l} \frac{\tau_j}{1+\tau_j b_j^l}|$$

$$\leq \max_l p^\delta |\sum_l \frac{\tau_l}{1+\tau_l a_l} - \sum_l \frac{\tau_l}{1+\tau_l c}| + \max_l p^\delta |\sum_l \frac{\tau_l}{1+\tau_l c} - \sum_{j\neq l} \frac{\tau_j}{1+\tau_j c}| + \max_l |\sum_{j\neq l} \frac{\tau_j}{1+\tau_j c} - \sum_{j\neq l} \frac{\tau_j}{1+\tau_j b_j^l}|$$

$$= \max_l p^\delta |\sum_l \frac{\tau_l}{1+\tau_l a_l} - \sum_l \frac{\tau_l}{1+\tau_l c}| + \max_l p^\delta |\frac{\tau_l}{1+\tau_l c}| + \max_l |\sum_{j\neq l} \frac{\tau_j}{1+\tau_j c} - \sum_{j\neq l} \frac{\tau_j}{1+\tau_j b_j^l}|.$$

$\max_l p^\delta |\frac{\tau_l}{1+\tau_l c}| \leq \max_l p^\delta |\frac{n\tau_l}{n(1+\tau_l c)}| \leq \frac{\kappa p^\delta}{n} \to 0.$

Now we will consider the third part of the above equation,

$$\max_l p^\delta |\sum_{j\neq l} \frac{\tau_j}{1+\tau_j b_j^l} - \sum_{j\neq l} \frac{\tau_j}{1+\tau_j c}|$$
$$= \max_l p^\delta |\sum_{j\neq l} \frac{\tau_j^2 (c - b_j^l)}{(1+\tau_j b_j^l)(1+\tau_j c)}|$$
$$= \max_l p^\delta |\sum_{j\neq l} \frac{\tau_j^2 p(c - b_j^l)/p}{(1+\tau_j b_j^l)(1+\tau_j c)}|.$$

$\max_l |\sum_{j\neq l} \frac{\tau_j^2 p}{(1+\tau_j b_j^l)(1+\tau_j c)}| \leq \frac{\kappa^2 p}{n^2}$ by Assumption 1 and 6,

$$\begin{aligned}
&\max_l \max_{j\neq l} p^{\epsilon} |b_j^l/p - c/p| \\
&= \max_l \max_{j\neq l} p^{\epsilon} |x_j' \Sigma^{\frac{1}{2}} (S_{j,l} - zI)^{-1} \Sigma^{\frac{1}{2}} x_j/p - tr(\Sigma^{\frac{1}{2}} (S - zI)^{-1} \Sigma^{\frac{1}{2}}/p)| \\
&= \max_l \max_{j\neq l} p^{\epsilon} |x_j' \Sigma^{\frac{1}{2}} (S_{j,l} - zI)^{-1} \Sigma^{\frac{1}{2}} x_j/p - tr(\Sigma^{\frac{1}{2}} (S_{j,l} - zI)^{-1} \Sigma^{\frac{1}{2}}/p)| \\
&+ \max_l \max_{j\neq l} p^{\epsilon} |tr(\Sigma^{\frac{1}{2}} (S_{j,l} - zI)^{-1} \Sigma^{\frac{1}{2}}/p) - tr(\Sigma^{\frac{1}{2}} (S - zI)^{-1} \Sigma^{\frac{1}{2}}/p)|.
\end{aligned}$$

Use of Lemma 28, Lemma 27 with Borel Cantelli Lemma will give us for $\epsilon < \frac{1}{2} - \delta$ and $k > \frac{3}{1-2\delta-2\epsilon}$,

$$\max_l \max_{j\neq l} p^{\epsilon} |x_j' \Sigma^{\frac{1}{2}} (S_{j,l} - zI)^{-1} \Sigma^{\frac{1}{2}} x_j/p - tr(\Sigma^{\frac{1}{2}} (S_{j,l} - zI)^{-1} \Sigma^{\frac{1}{2}}/p)| \to 0 \text{ a.s.}$$

Also,

$$\begin{aligned}
&\max_l \max_{j\neq l} p^{\epsilon} |\frac{1}{p} tr(\Sigma^{\frac{1}{2}} (S_{j,l} - zI)^{-1} \Sigma^{\frac{1}{2}}) - \frac{1}{p} tr(\Sigma^{\frac{1}{2}} (S - zI)^{-1} \Sigma^{\frac{1}{2}})| \\
&= \max_l \max_{j\neq l} p^{\epsilon} |\frac{1}{p} tr[\Sigma^{\frac{1}{2}} \{(S_{j,l} - zI)^{-1} - (S - zI)^{-1}\} \Sigma^{\frac{1}{2}}]| \\
&= \max_l \max_{j\neq l} p^{\epsilon} |\frac{1}{p} tr[\{(S_{j,l} - zI)^{-1} - (S - zI)^{-1}\} \Sigma]| \\
&\leq \max_l \max_{j\leq l} p^{\epsilon} \frac{1}{p} \frac{||\Sigma||}{v} \\
&\leq \frac{1}{p} \frac{C p^{\delta+\epsilon}}{v} \to 0 \text{ a.s.}
\end{aligned}$$

This proves out claim.

The first and second inequalities are result of application of Lemma 26 and Assumption 7 respectively .

Proof of Lemma 34:

Proof. Using Lemma_ and and Markov Inequality it is easy to show that,

$$E(x_l'\Sigma^{\frac{1}{2}}(S_l - zI)^{-1}(\Sigma\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l} - zI)^{-1}\Sigma^{\frac{1}{2}}x_l - tr[\Sigma^{\frac{1}{2}}(S_l - zI)^{-1}(\Sigma\sum_{j\neq l}\frac{\tau_j}{1+\tau_j b_j^l} - zI)^{-1}\Sigma^{\frac{1}{2}}])$$
$$\leq C_k p^k \frac{p^{2\delta k}}{v^{2k}}.$$

After choosing appropriate value of k and using Borel Cantelli Lemma we can get the claim.

□

7.4 Chapter 5

Remark 1:

$$
\begin{aligned}
M &= \sqrt{\frac{n}{p}}(WW^* - (\sum_l \tau_l)\Lambda) \\
&= \sqrt{\frac{n}{p}}(\sum_{k=1}^{p} e_k w_k^* W^* - (\sum_l \tau_l)\Lambda) \\
&= \sqrt{\frac{n}{p}}\sum_{k=1}^{p} e_k w_k^*(W_l^* + w_k e_k^*) - \sqrt{\frac{n}{p}}(\sum_l \tau_l)\Lambda \\
&= \sqrt{\frac{n}{p}}\sum_{k=1}^{p} e_k w_k^* W_k^* + \sqrt{\frac{n}{p}}\sum_{k=1}^{p} e_k w_k^* w_k e_k^* - \sqrt{\frac{n}{p}}(\sum_l \tau_l)\Lambda \\
&= \sqrt{\frac{n}{p}}\sum_{k=1}^{p} e_k (W_k w_k)^* + \sqrt{\frac{n}{p}}\sum_{k=1}^{p} e_k w_k^* w_k e_k^* - \sqrt{\frac{n}{p}}(\sum_l \tau_l)\sum_{k=1}^{p}\lambda_k e_k e_k^* \\
&= \sum_{k=1}^{p} e_k h_k^* + \sum_{k=1}^{p} e_k \sqrt{\frac{n}{p}}(w_k^* w_k - (\sum_l \tau_l)\lambda_k) e_k^* \\
&= \sum_{k=1}^{p} e_k h_k^* + \sum_{k=1}^{p} e_k \sqrt{\frac{n}{p}}(w_k^* w_k - (\sum_l \tau_l)\lambda_k) e_k^* \\
&= \sum_{k=1}^{p} e_k h_k^* + \sum_{k=1}^{p} e_k t_{kk} e_k^* \\
&= \sum_{k=1}^{p} e_k (h_k + t_{kk} e_k)^*,
\end{aligned}
$$

also it is easy to see that,

$$
\begin{aligned}
M &= \sqrt{\frac{n}{p}}(WW^* - (\sum_l \tau_l)\Lambda) \\
&= \sqrt{\frac{n}{p}}(W_k W^* + e_k w_k^* W^* - (\sum_l \tau_l)\Lambda_k - (\sum_l \tau_l) e_k e_k^* \lambda_k) \\
&= \sqrt{\frac{n}{p}}(W_k W^* - (\sum_l \tau_l)\Lambda_k) + \sqrt{\frac{n}{p}}[e_k w_k^*(W_k^* + w_k e_k^*) - (\sum_l \tau_l) e_k e_k^* \lambda_k] \\
&= M_k + e_k h_k^* + \sqrt{\frac{n}{p}} e_k w_k^* w_k e_k^* - \sqrt{\frac{n}{p}}(\sum_l \tau_l) e_k e_k^* \lambda_k \\
&= M_k + e_k h_k^* + e_k \sqrt{\frac{n}{p}}(w_k^* w_k - (\sum_l \tau_l)\lambda_k) e_k^* \\
&= M_k + e_k h_k^* + t_{kk} e_k e_k^* \\
&= M_k + e_k (h_k + t_{kk} e_k)^*.
\end{aligned}
$$

Remark 2:

$$
\begin{aligned}
& tr[e_k(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k] \\
&= (h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k tr[e_k(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}] \\
&= \{(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k\}\{(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k\} \\
&= \{(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k\}^2
\end{aligned}
$$

and

$$
\begin{aligned}
& tr[e_k(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}] \\
&= tr[(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k] \\
&= \{(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k\}^2.
\end{aligned}
$$

So,

$$
\begin{aligned}
LHS &= tr(M(M-zI)^{-1}) \\
&= tr\{e_k(h_k+t_{kk}e_k)^*(M_k-zI+e_k(h_k+t_{kk}e_k)^*)^{-1}\} \\
&= tr[e_k(h_k+t_{kk}e_k)^*(M_k(z)^{-1}-\frac{(M_k-zI)^{-1}e_k(h_k+t_{kk}e_k)^*(M_k-zI)^{-1}}{1+(h_k+t_{kk}e_k)^*(M_k-zI)^{-1}e_k})] \\
&= tr[e_k(h_k+t_{kk}e_k)^*((M_k-zI)^{-1}-\frac{(M_k-zI)^{-1}e_k(h_k+t_{kk}e_k)^*(M_k-zI)^{-1}}{1+(h_k+t_{kk}e_k)^*(M_k-zI)^{-1}e_k})] \\
&= tr[e_k(h_k+t_{kk}e_k)^*(\frac{(M_k-zI)^{-1}(1+(h_k+t_{kk}e_k)^*(M_k-zI)^{-1}e_k)-(M_k-zI)^{-1}e_k(h_k+t_{kk}e_k)^*(M_k}{1+(h_k+t_{kk}e_k)^*M_k(z)^{-1}e_k} \\
&= tr[\frac{e_k(h_k+t_{kk}e_k)^*(M_k-zI)^{-1}}{1+(h_k+t_{kk}e_k)^*(M_k-zI)^{-1}e_k}] \\
&= \frac{(h_k+t_{kk}e_k)^*(M_k-zI)^{-1}e_k}{1+(h_k+t_{kk}e_k)^*(M_k-zI)^{-1}e_k}
\end{aligned}
$$

Remark 3:

To see this, manipulate the left hand side the following way-

$$
\begin{aligned}
(h_k+t_{kk}e_k)^*(M_k-zI)^{-1}e_k &= (h_k+t_{kk}e_k)^*(\sqrt{\frac{n}{p}}(W_kW^*-(\sum_l \tau_l)\Lambda_k)-zI)^{-1}e_k \\
&= (h_k+t_{kk}e_k)^*(\sqrt{\frac{n}{p}}(W_kW_k^*+W_kw_ke_k^*-(\sum_l \tau_l)\Lambda_k)-zI)^{-1}e_k \\
&= (h_k+t_{kk}e_k)^*(\sqrt{\frac{n}{p}}(W_kW_k^*-(\sum_l \tau_l)\Lambda_k)-zI+\sqrt{\frac{n}{p}}W_kw_ke_k^*)^{-1}e_k \\
&= (h_k+t_{kk}e_k)^*(\sqrt{\frac{n}{p}}(W_kW_k^*-(\sum_l \tau_l)\Lambda_k)-zI+h_ke_k^*)^{-1}e_k \\
&= (h_k+t_{kk}e_k)^*((\bar{M}_k-zI)+h_ke_k^*)^{-1}e_k \\
&= (h_k+t_{kk}e_k)^*((\bar{M}_k-zI)^{-1}-\frac{(\bar{M}_k-zI)^{-1}h_ke_k^*(\bar{M}_k-zI)^{-1}}{1+e_k^*(\bar{M}_k-zI)^{-1}h_k})e_k \\
&= (h_k+t_{kk}e_k)^*\bar{M}_k(z)^{-1}e_k-\frac{(h_k+t_{kk}e_k)^*(\bar{M}_k-zI)^{-1}h_ke_k^*(\bar{M}_k-zI)^{-1}e_k}{1+e_k^*(\bar{M}_k-zI)^{-1}h_k}.
\end{aligned}
$$

Observe that $h_k^* e_k = e^* h_k = 0$ because $W_k^* e_k = 0$. So,

$$\begin{aligned}(\bar{M}_k - zI)e_k &= \sqrt{\frac{n}{p}}(W_k W_k^* - (\sum_l \tau_l)\Lambda_k)e_k - ze_k \\ &= \sqrt{\frac{n}{p}}(W_k W_k^* e_k - (\sum_l \tau_l)\Lambda_k e_k) - ze_k \\ &= -ze_k.\end{aligned}$$

As a consequence $e_k^*(\bar{M}_k - zI)^{-1}h_k = -e_k^* h_k/z = 0$. And,

$$\begin{aligned}(h_k + t_{kk}e_k)^*(M_k - zI)^{-1}e_k &= -(h_k + t_{kk}e_k)^* e_k/z + (h_k + t_{kk}e_k)^*(\bar{M}_k - zI)^{-1}h_k e_k^* e_k/z \\ &= -t_{kk}/z + (h_k + t_{kk}e_k)^*(\bar{M}_k - zI)^{-1}h_k/z \\ &= \frac{-t_{kk} + h_k^*(\bar{M}_k - zI)^{-1}h_k + t_{kk}e_k{}^*(\bar{M}_k - zI)^{-1}h_k}{z} \\ &= \frac{-t_{kk} + h_k^*(\bar{M}_k - zI)^{-1}h_k}{z}.\end{aligned}$$

Remark 4:

$$
\begin{aligned}
h_k^*(\bar{M}_k - zI)^{-1}h_k &= \frac{n}{p}w_k^*W_k^*(\bar{M}_k - zI)^{-1}W_k w_k \\
&= \frac{n}{p}tr(W_k w_k w_k^* W_k^*(\bar{M}_k - zI)^{-1}) \\
&= \frac{n\lambda_k}{pn}tr(W_k diag(\tau_1^{\frac{1}{2}},..,\tau_n^{\frac{1}{2}})Ux_k x_k^* U^* diag(\tau_1^{\frac{1}{2}},..,\tau_n^{\frac{1}{2}})W_k^*(\bar{M}_k - zI)^{-1}) \\
&= \frac{n\lambda_k}{p}tr(W_k diag(\tau_1^{\frac{1}{2}},..,\tau_n^{\frac{1}{2}})U(x_k x_k^* - I)U^* diag(\tau_1^{\frac{1}{2}},..,\tau_n^{\frac{1}{2}})W_k^*(\bar{M}_k - zI)^{-1}) \\
&+ \frac{n\lambda_k}{p}tr(W_k diag(\tau_1,..,\tau_n)W_k^*(\bar{M}_k - zI)^{-1}) \\
&= \frac{n\lambda_k}{p}tr(W_k diag(\tau_1^{\frac{1}{2}},..,\tau_n^{\frac{1}{2}})U(x_k x_k^* - I)U^* diag(\tau_1^{\frac{1}{2}},..,\tau_n^{\frac{1}{2}})W_k^*(\bar{M}_k - zI)^{-1}) \\
&+ \frac{n\lambda_k}{p}tr(\Lambda_k^{\frac{1}{2}}UX_n(diag(\tau_1,..,\tau_n))^2X_n^*U^*\Lambda_k^{\frac{1}{2}}(\bar{M}_k - zI)^{-1}) \\
&= \frac{n\lambda_k}{p}tr(W_k diag(\tau_1^{\frac{1}{2}},..,\tau_n^{\frac{1}{2}})U(x_k x_k^* - I)U^* diag(\tau_1^{\frac{1}{2}},..,\tau_n^{\frac{1}{2}})W_k^*(\bar{M}_k - zI)^{-1}) \\
&+ \frac{n\lambda_k}{p}tr(\Lambda_k^{\frac{1}{2}}UX_n(\sum_l \tau_n^2 e_l e_l^* - (\sum_l \tau_n^2)I)X_n^*U^*\Lambda_k^{\frac{1}{2}}(\bar{M}_k - zI)^{-1}) \\
&+ \frac{n\lambda_k}{p}tr(\Lambda_k^{\frac{1}{2}}UX_n(\sum_l \tau_n^2)X_n^*U^*\Lambda_k^{\frac{1}{2}}(\bar{M}_k - zI)^{-1}) \\
&= \frac{n\lambda_k}{p}tr(W_k diag(\tau_1^{\frac{1}{2}},..,\tau_n^{\frac{1}{2}})U(x_k x_k^* - I)U^* diag(\tau_1^{\frac{1}{2}},..,\tau_n^{\frac{1}{2}})W_k^*(\bar{M}_k - zI)^{-1}) \\
&+ \frac{n\lambda_k}{p}tr(\Lambda_k^{\frac{1}{2}}UX_n(\sum_l \tau_n^2 e_l e_l^* - (\sum_l \tau_n^2)I)X_n^*U^*\Lambda_k^{\frac{1}{2}}(\bar{M}_k - zI)^{-1}) \\
&+ \frac{n\lambda_k(\sum_l \tau_n^2)}{p}tr(\Lambda_k^{\frac{1}{2}}U(X_nX_n^* - I)U^*\Lambda_k^{\frac{1}{2}}(\bar{M}_k - zI)^{-1}) \\
&+ \frac{n\lambda_k(\sum_l \tau_n^2)}{p}tr(\Lambda_k(\bar{M}_k - zI)^{-1}).
\end{aligned}
$$

Bibliography

[1] Pierre Giot. Market risk models for intraday data. *The European Journal of Finance*, 11(4):309–324, 2005.

[2] CNV Krishnan, Ralitsa Petkova, and Peter Ritchken. Correlation risk. *Journal of Empirical Finance*, 16(3):353–367, 2009.

[3] Mark Salmon and Christoph Schleicher. Pricing multivariate currency options with copulas. *Copulas: From Theory to Application in Finance, Risk Books, London*, 2006.

[4] Paul Embrechts, Alexander McNeil, and Daniel Straumann. Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 1:176–223, 2002.

[5] Rüdiger Frey and Alexander J McNeil. Var and expected shortfall in portfolios of dependent credit risks: conceptual and practical insights. *Journal of banking & finance*, 26(7):1317–1334, 2002.

[6] Wolfgang Breymann, Alexandra Dias, and Paul Embrechts. Dependence structures for multivariate high-frequency data in finance. 2003.

[7] Yang An and Ngai Hang Chan. Short-term stock price prediction based on limit order book dynamics. *Journal of Forecasting*, 36(5):541–556, 2017.

[8] Thomas W Epps. Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 74(366a):291–298, 1979.

[9] Allan A Zebedee and Maria Kasch-Haroutounian. A closer look at comovements among stock returns. *Journal of Economics and Business*, 61 (4):279–294, 2009.

[10] Mark C Lundin, Michel M Dacorogna, and Ulrich A Müller. Correlation of high frequency financial time series. *Available at SSRN 79848*, 1998.

[11] Jayaram Muthuswamy, Sudipto Sarkar, Aaron Low, and Eric Terry. Time variation in the correlation structure of exchange rates: high-frequency analyses. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, 21(2):127–144, 2001.

[12] Roberto Renò. A closer look at the epps effect. *International Journal of theoretical and applied finance*, 6(01):87–102, 2003.

[13] Lan Zhang. Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47, 2011.

[14] Lan Zhang, Per A Mykland, and Yacine Aït-Sahalia. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411, 2005.

[15] Jianqing Fan, Yingying Li, and Ke Yu. Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association*, 107(497):412–428, 2012.

[16] Yacine Aït-Sahalia, Jianqing Fan, and Dacheng Xiu. High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association*, 105(492):1504–1517, 2010.

[17] Robert C Merton. On estimating the expected return on the market: An exploratory investigation. *Journal of financial economics*, 8(4):323–361, 1980.

[18] Robbin Tops. Copulas and correlation in credit risk. *Barchelor of Science Thesis*, 2010.

[19] Qing Xu. *Estimating and evaluating the Archimedean-copula-based models in financial risk management: a dissertation submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in Financial Economics, Massey University, Auckland, New Zealand*. PhD thesis, Massey University, 2008.

[20] Maxime Malgrat. Pricing of a 'worst of' option using a copula method, 2013.

[21] Lidan Grossmass and Ser-Huang Poon. Estimating dynamic copula dependence using intraday data. *Studies in Nonlinear Dynamics & Econometrics*, 19(4):501–529, 2015.

[22] Matthias R Fengler and Ostap Okhrin. Managing risk with a realized copula parameter. *Computational Statistics & Data Analysis*, 100:131–152, 2016.

[23] Lechuan Yin, Jiebin Chen, and Zhao-Rong Lai. Mixed-copula var for portfolio risk evaluation. In *2018 IEEE International Conference on Big Knowledge (ICBK)*, pages 400–408. IEEE, 2018.

[24] Michael McAleer and Marcelo C Medeiros. Realized volatility: A review. *Econometric Reviews*, 27(1-3):10–45, 2008.

[25] Tim Bollerslev. A conditionally heteroskedastic time series model for spec-

ulative prices and rates of return. *Review of economics and statistics*, 69 (3):542–547, 1987.

[26] M Angeles Carnero, Daniel Peña, and Esther Ruiz. Persistence and kurtosis in garch and stochastic volatility models. *Journal of Financial Econometrics*, 2(2):319–342, 2004.

[27] Hans Malmsten and Timo Teräsvirta. Stylized facts of financial time series and three popular models of volatility. *SSE/EFI Working Paper Series in Economics and Finance*, 563:1–44, 2004.

[28] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, STANFORD UNIVERSITY STANFORD United States, 1956.

[29] Yazhen Wang and Jian Zou. Volatility analysis in high-frequency financial data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):393–404, 2014.

[30] Xin Guo, Howard Shek, Tze Leung Lai, and Samuel Po-Shing Wong. *Quantitative Trading: Algorithms, Analytics, Data, Models, Optimization*. Chapman and Hall/CRC, 2017.

[31] Yacine Ait-Sahalia, Per A Mykland, and Lan Zhang. How often to sample a continuous-time process in the presence of market microstructure noise. *The review of financial studies*, 18(2):351–416, 2005.

[32] Stefano Demarta and Alexander J McNeil. The t copula and related copulas. *International statistical review*, 73(1):111–129, 2005.

[33] Hong-Bin Fang, Kai-Tai Fang, and Samuel Kotz. The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82(1): 1–16, 2002.

[34] Umberto Cherubini, Elisa Luciano, and Walter Vecchiato. *Copula methods in finance*. John Wiley & Sons, 2004.

[35] Ayumi Ida, Naoyuki Ishimura, and MasaAki Nakamura. Note on the measures of dependence in terms of copulas. *Procedia Economics and Finance*, 14:273–279, 2014.

[36] Ole E Barndorff-Nielsen, Peter Reinhard Hansen, Asger Lunde, and Neil Shephard. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics*, 162(2):149–169, 2011.

[37] Takaki Hayashi and Nakahiro Yoshida. On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 11(2):359–379, 2005.

[38] Valeri Voev and Asger Lunde. Integrated covariance estimation using high-frequency data in the presence of noise. *Journal of Financial Econometrics*, 5(1):68–104, 2006.

[39] Nikolaus Hautsch, Lada M Kyj, and Roel CA Oomen. A blocking and regularization approach to high-dimensional realized covariance estimation. *Journal of Applied Econometrics*, 27(4):625–645, 2012.

[40] Olivier Ledoit and Michael Wolf. Honey, i shrunk the sample covariance matrix. *UPF economics and business working paper*, (691), 2003.

[41] Federico M Bandi and Jeffrey R Russell. Separating microstructure noise from volatility. *Journal of Financial Economics*, 79(3):655–692, 2006.

[42] Karim Bannouh, Martin Martens, Roel CA Oomen, and Dick JC van Dijk. Realized mixed-frequency factor models for vast dimensional covariance estimation. *ERIM Report Series Reference No. ERS-2012-017-F&A*, 2012.

[43] Asger Lunde, NEIL Shephard, and Kevin Sheppard. Econometric analysis of vast covariance matrices using composite realized kernels. *Manuscript, University of Aarhus*, 2011.

[44] Xinghua Zheng and Yingying Li. On the estimation of integrated covariance matrices of high dimensional diffusion processes. 2011.

[45] Ningning Xia and Xinghua Zheng. On the inference about the spectral distribution of high-dimensional covariance matrix based on high-frequency noisy observations. *The Annals of Statistics*, 46(2):500–525, 2018.

[46] Christian Y Robert and Mathieu Rosenbaum. On the limiting spectral distribution of the covariance matrices of time-lagged processes. *Journal of multivariate analysis*, 101(10):2434–2451, 2010.

[47] Cheng Wang, Baisuo Jin, and Baiqi Miao. On limiting spectral distribution of large sample covariance matrices by varma (p, q). *Journal of Time Series Analysis*, 32(5):539–546, 2011.

[48] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[49] ZD Bai. Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica*, pages 611–662, 1999.

[50] Jeffrey S Geronimo and Theodore P Hill. Necessary and sufficient condition that the limit of stieltjes transforms is a stieltjes transform. *Journal of Approximation Theory*, 121(1):54–60, 2003.

[51] Zhidong D Bai and Yong Q Yin. Convergence to the semicircle law. *The Annals of Probability*, pages 863–875, 1988.

[52] Zhigang Bao. Strong convergence of esd for the generalized sample covariance matrices when p/n? 0. *Statistics & Probability Letters*, 82(5):894–901, 2012.

[53] ZD Bai and LX Zhang. The limiting spectral distribution of the product of the wigner matrix and a nonnegative definite matrix. *Journal of Multivariate Analysis*, 101(9):1927–1949, 2010.

[54] GM Pan and JT Gao. Asymptotic theory for sample covariance matrix under cross-sectional dependence. *Preprint*, 2012.

[55] Lili Wang and Debashis Paul. Limiting spectral distribution of renormalized separable sample covariance matrices when p/n? 0. *Journal of Multivariate Analysis*, 126:25–52, 2014.

[56] Colin McDiarmid. Centering sequences with bounded differences. *Combinatorics, Probability and Computing*, 6(1):79–86, 1997.

[57] Rémy Mariétan. Statistical applications of random matrix theory: Comparison of two populations. Technical report, EPFL, 2019.

[58] Marc Potters, Jean-Philippe Bouchaud, and Laurent Laloux. Financial applications of random matrix theory: Old laces and new pieces. *arXiv preprint physics/0507111*, 2005.

[59] Feher Kristen, Whelan James, and Müller Samuel. Assessing modularity using a random matrix theory approach. *Statistical Applications in Genetics and Molecular Biology*, 10(1):1–34, 2011.

[60] Martin Hyrš and Josef Schwarz. Elliptical and archimedean copulas in estimation of distribution algorithm with model migration. In *Computational Intelligence (IJCCI), 2015 7th International Joint Conference on*, volume 1, pages 212–219. IEEE, 2015.

[61] Christian Genest and R Jock MacKay. Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canadian Journal of Statistics*, 14(2):145–159, 1986.

[62] Harry Joe and James Jianmeng Xu. The estimation method of inference functions for margins for multivariate models. 1996.

[63] Paul Embrechts and Marius Hofert. Statistical inference for copulas in high dimensions: A simulation study. *ASTIN Bulletin: The Journal of the IAA*, 43(2):81–95, 2013.

[64] Arthur Charpentier, Jean-David Fermanian, and Olivier Scaillet. The estimation of copulas: Theory and practice. 2007.

[65] Eugene P Wigner. On the distribution of the roots of certain symmetric matrices. *Ann. Math*, 67(2):325–327, 1958.

[66] Arup Bose and Monika Bhattacharjee. *Large Covariance and Autocovariance Matrices*. Chapman and Hall/CRC, 2018.

[67] Ludwig Arnold. On the asymptotic distribution of the eigenvalues of random matrices. Technical report, WISCONSIN UNIV MADISON MATHEMATICS RESEARCH CENTER, 1967.

[68] Ulf Grenander. *Probabilities on algebraic structures*. Courier Corporation, 2008.

[69] Ulf Grenander and Jack W Silverstein. Spectral analysis of networks with random topologies. *SIAM Journal on Applied Mathematics*, 32(2):499–519, 1977.

[70] Viacheslav Leonidovich Girko. *Theory of random determinants*, volume 45. Springer Science & Business Media, 2012.

[71] James R Schott. *Matrix analysis for statistics*. John Wiley & Sons, 2016.

[72] Montserrat Fuentes. Testing for separability of spatial–temporal covariance functions. *Journal of statistical planning and inference*, 136(2):447–466, 2006.

[73] Matthew W Mitchell, Marc G Genton, and Marcia L Gumpertz. Testing for separability of space–time covariances. *Environmetrics: The official journal of the International Environmetrics Society*, 16(8):819–831, 2005.

[74] Matthew W Mitchell, Marc G Genton, and Marcia L Gumpertz. A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis*, 97(5):1025–1043, 2006.

[75] Nelson Lu and Dale L Zimmerman. The likelihood ratio test for a separable covariance matrix. *Statistics & probability letters*, 73(4):449–457, 2005.

[76] Anuradha Roy and Ravindra Khattree. On implementation of a test for kronecker product covariance structure for multivariate repeated measures data. *Statistical Methodology*, 2(4):297–306, 2005.

[77] Bo Li, Marc G Genton, and Michael Sherman. Testing the covariance structure of multivariate random fields. *Biometrika*, 95(4):813–829, 2008.

[78] Sean L Simpson. An adjusted likelihood ratio test for separability in unbalanced multivariate repeated measures data. *Statistical Methodology*, 7 (5):511–519, 2010.

[79] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[80] Leonid A Pastur. Spectra of random self adjoint operators. *Russian mathematical surveys*, 28(1):1, 1973.

[81] Leonid A Pastur. On the spectrum of random matrices. *Theoretical and Mathematical Physics*, 10(1):67–74, 1972.

[82] Yong-Quan Yin, Zhi-Dong Bai, and Pathak R Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability theory and related fields*, 78(4):509–521, 1988.

[83] Zhidong D Bai, Jack W Silverstein, and Yong Q Yin. A note on the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis*, 26(2):166–168, 1988.

[84] ZD Bai and YQ Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. In *Advances In Statistics*, pages 108–127. World Scientific, 2008.

[85] Jack W Silverstein. Some limit theorems on the eigenvectors of large dimensional sample covariance matrices. *Journal of multivariate Analysis*, 15(3):295–324, 1984.

[86] Arup Bose, Sourav Chatterjee, and Sreela Gangopadhyay. Limiting spectral distribution of large dimensional random matrices. 2003.

[87] Zhidong Bai and Hewa Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, pages 311–329, 1996.