論　文

Some Statistical Properties of Tick Data Set

(Part II)

Kangrong Tan

Abstract

In this paper, we explore some statistical properties of log returns by using the tick data set of the share market. We confirm the aggregation effect as the sampling time interval increases. And we also study the tail behavior of the distribution of log returns with a different sampling time interval. It shows that tail behavior of return distribution is much closer to student t distribution than normal distribution. And the calculation results show that there are also seasonal patterns in the tick data set.

Keyword

Tick data set, log returns, intraday trading, aggregation effect, evolution of returns distribution, tail behavior, Hill estimator, seasonality, trading volume

1 Introduction

Since the Geometric Brownian Motion model has been introduced in the spectrum of financial engineering, many quantitative models have been very much influenced by it. Under the Geometric Brownian Motion model, price P_t at time t of an asset is defined as follows.

$$P_t = P_0 exp(\mu t + \sigma W(t)) \tag{1}$$

where $W(t)$ is a Standard Brownian Motion and P_0 is the starting price at time zero, with $\mu, \sigma > 0$. Then, the log returns can be written as follows.

$$X_t = log\frac{P_t}{P_{t-1}} = \mu + \sigma(W(t) - W(t+1)) \tag{2}$$

Thus, the $\{X_t\}$ can be considered as an i.i.d Gaussian process with mean μ and variance σ. Many researches show that the longterm returns processes are close to the Geometric Brownian Motion model. If Geometric Brownian Motion is true, then the drift exponent under the Geometric Brownian Motion should be close to 0.5.

In this paper we check if the tick by tick trading data have those traditional statistical properties. Simultaneously, we study the evolution of the return distributions and the tail behavior with different sampling the intervals. The rest of this paper is organized as follows. In section 2, we simply describe the original data set which we use later. In section 3, we explore some statistical properties of the tick data sets. In section 4, we discuss the seasonal phenomena in the tick data sets. In section 5, we study some statistical properties of the trading volume. And section 6 gives the summary of this paper.

2 About the data set

In this paper, we use C stock price data set for the empirical analyses. The format of the data set being used here is shown in Table-1. One record is consisted of five information fields. There are Trading Date & Time, Ask, Bid, Transaction Price, and Transaction Volume. The range of the data set is from 1 of November, 1990 to 31 of January, 1991. It includes 63 trading days and 60328 transactions. According to Stoll and

Table-1 Data set Format

Date & Time	Ask	Bid	Trans Price	Volume
90110134228	102.5	102.375	102.375	15800
90110134236	102.5	102.375	102.375	300
90110134228	102.3	102.2	102.3	20000
90110134236	102.3	102.2	102.3	12000
		.....		

Whaley (1990), who conclude the difference of price process from late afternoon (4:00pm) to early morning (9:30) is much more different from that in intraday tradings and indicate that it is not a stable process. Thus, we only extract the trading data between 9:30 am to 4:00 pm and apply them to our analyses.

2.1 Some statistical properties of the data set

In this section we simply summarize some statistical properties of the original data set which is applied to the analyses later. Figure-1 and Figure-2 are the plots of time series of the original price data set and its log returns respectively. Figure-3 and Figure-4 show the plots of ACF levels of log returns and their absolute values respectively. Seen from those figures, the ACF levels of log returns are obviously lower compared to the ACF levels of their absolute values. It can be considered an evidence of the Long Range Dependence. There are some minus values seen from the Figure-3 at the very beginning stage of the lags. It means that there are some opposite buying-and-selling behavior during the very short trading time. And also seen from Figure-5, the density function of log returns standardized with mean 0 and variance 1 is far away from the standard normal distribution although it is almost symmetric. And it also indicates

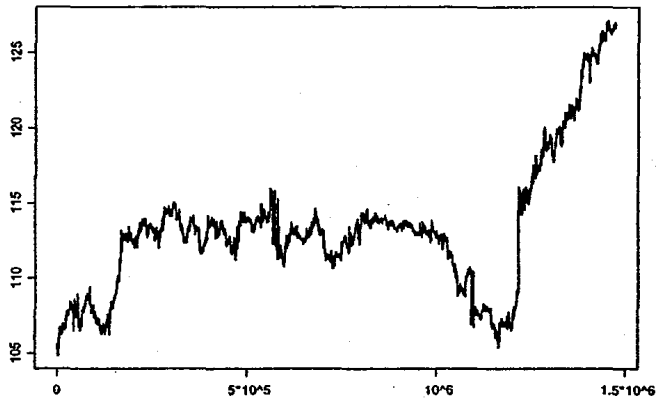

Figure 1: Plot of the price time series

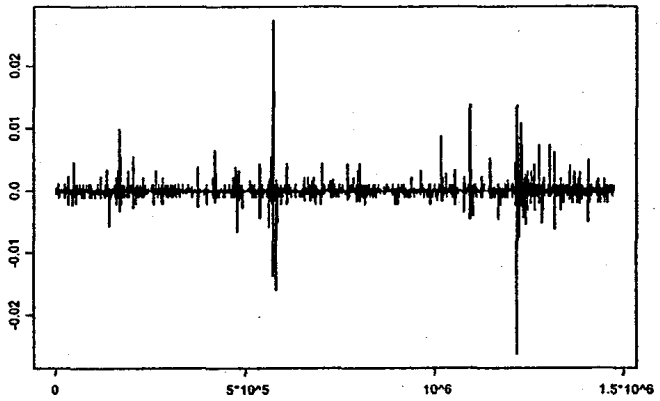

Figure 2: Plot of the log returns

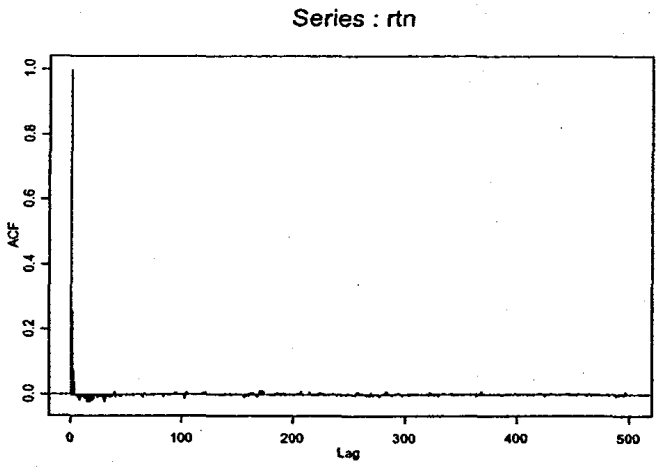

Figure 3: Plot of ACF of log returns

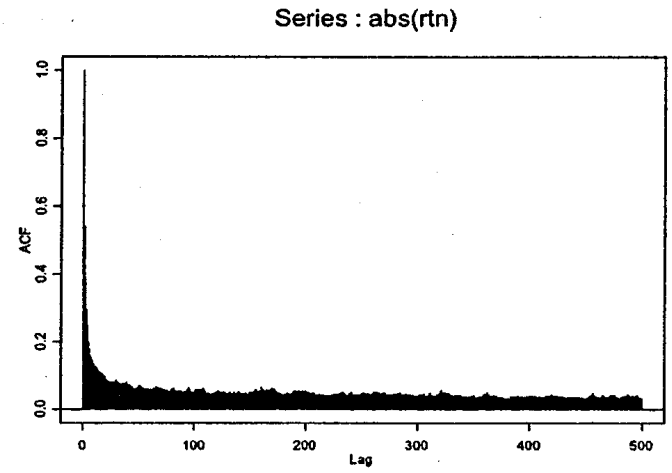

Figure 4: Plot of ACF of absolute values of log returns

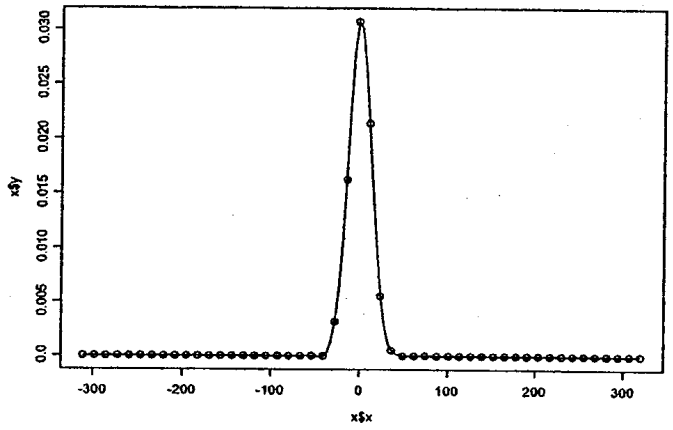

Figure 5: Plot of density of log returns

that the reality of the financial market does not meet the assumptions of Geometric Brownian Motion model in the case of tick by tick trading data set.

2.2 Data sets sampling at different time intervals

In order to explore the statistical properties of the tick by tick tradings, we build new data sets at different sampling time intervals from the original data set, We organize 10 data sets from the original data set

Table-2 Number of samples in each data set

Name	Δt	Sample numbers
r_1	5	294848
r_2	10	147424
r_3	20	73712
r_4	40	36856
r_5	80	18428
r_6	160	9214
r_7	320	4607
r_8	640	2304
r_9	1280	1152
r_{10}	2560	576

with a different sampling time interval, say Δt, taking 5 seconds, 20 seconds and so on, till 2560 seconds. The number of samples of each new data set is shown in Table-2.

Here return is defined as follows.

$$r_{\Delta t} = logp(t+\Delta t) - logp(t) \tag{3}$$

where Δt takes 5, 10, 20, 40, 80, 160, 320, 640, 1280, and 2560 seconds. For simplicity, we call those new data sets as r_1, r_2, ..., r_{10}. And we call those 10 data sets standardized with mean 0 and variance 1 as z_1, z_2, ..., z_{10}. namely,

$$z_i = \frac{r_i - E[r_i]}{\sqrt{Var(r_i)}} \tag{4}$$

where $i = 1, 2, ..., 10$.

3 Statistical properties of tick by tick data sets

3.1 Descriptive statistics

In this section we exhibit some statistical properties of those 10 tick by tick data sets. The main descriptive statistics are listed in Table-3.

Seen from Table-3, with the increment of the sampling time interval, the kurtosis and skewness of log returns become smaller and smaller, and the standard deviation and mean are monotonically increasing (except r_{10}). Those phenomena can be considered as part of evidences of aggregation effect in the returns distributions of the financial markets. That is to say, the S.D in a wider trading time interval is always larger than that in a narrower trading time interval. With the increment of time scale, the returns distribution tends to be close to Gaussian distribution.

Table-3 Descriptive statistics of the data set

r_i	Kurtosis	Skewness	Standard Deviation	Mean
r_1	5804.67583	20.05188	2.84139-e004	6.29280-e007
r_2	2452.79534	11.85819	4.19388-e004	1.25741-e006
r_3	461.96845	1.29925	5.46683-e004	2.51023-e006
r_4	294.13239	1.03303	7.16722-e004	5.02702-e006
r_5	195.77727	0.80091	9.38418-e004	1.00362-e005
r_6	156.32127	1.18085	1.18068-e003	2.01271-e005
r_7	85.97425	−0.09408	1.66402-e003	4.00980-e005
r_8	60.10671	1.71411	2.24846-e003	8.01959-e005
r_9	36.85600	1.50138	3.11725-e003	1.60465-e004
r_{10}	59.56587	4.32272	4.31968-e003	3.22916-e004

3.2 Evolution of distributions

Seen from both Table-4 and Figure-6, 7 as the sampling time interval tends to be wider, the shape of the density function of the log returns is getting closer and closer to the standard normal distribution. Seen from the estimated modes listed in Table-4, eventually the mode of z_{10} becomes 0.3495, quite close to 0.39, which is the mode held by standard normal distribution.

Furthermore, the modes, and the estimated υ at 1% tails are listed up in Table-4 and Table-5. It is clearly that at most of the cases, those distributions of $z_1 \sim z_{10}$ have fat tails, and very close to Student t distributions which have different degrees of freedom(υ), standardized by $\sigma = \sqrt{\frac{\upsilon}{\upsilon - 2}}$.

Table-4 Evolution of the modes of log returns distributions standardized with $\sigma = 1$ and $\mu = 0$

z_i	Estimated mode from z_i
z_1	0.0502005
z_2	0.0678106
z_3	0.1168644
z_4	0.1425529
z_5	0.1738793
z_6	0.1912392
z_7	0.2649319
z_8	0.2732431
z_9	0.3231789
z_{10}	0.3495165

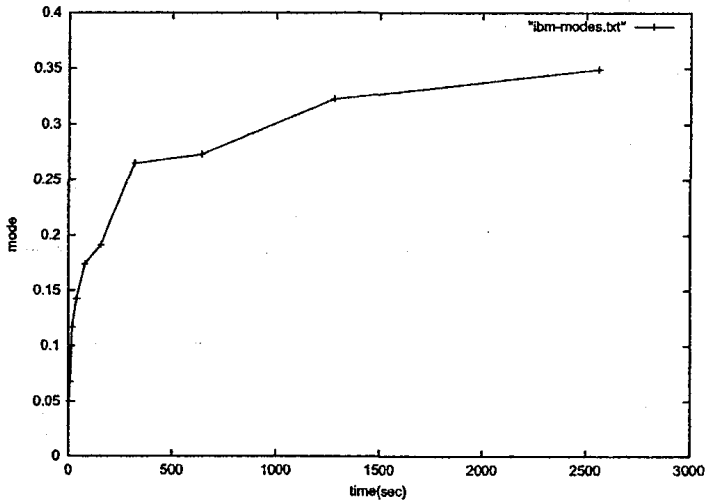

Figure 6: Evolution of the Modes

Table-5 Estimated υ in $t_\upsilon/\sqrt{\frac{\upsilon}{(\upsilon-2)}}$ at 1% tail

z_i	Estimated υ at 1% tail
z_1	2.3 ~ 2.4
z_2	2.5 ~ 2.6
z_3	2.3 ~ 2.4
z_4	2.4 ~ 2.5
z_5	2.3 ~ 2.4
z_6	2.5 ~ 2.6
z_7	2.4 ~ 2.5
z_8	5 ~ 6
z_9	6 ~ 7
z_{10}	2.4 ~ 2.5

3.3 Hill estimator

In practice, Hill estimator is widely adopted with great success, though, it also shows poor performance sometimes, known as "horror plots", see Resnick (1997). Hill plots of $z_1 \sim z_{10}$ are shown in Figure-8. Seen from the

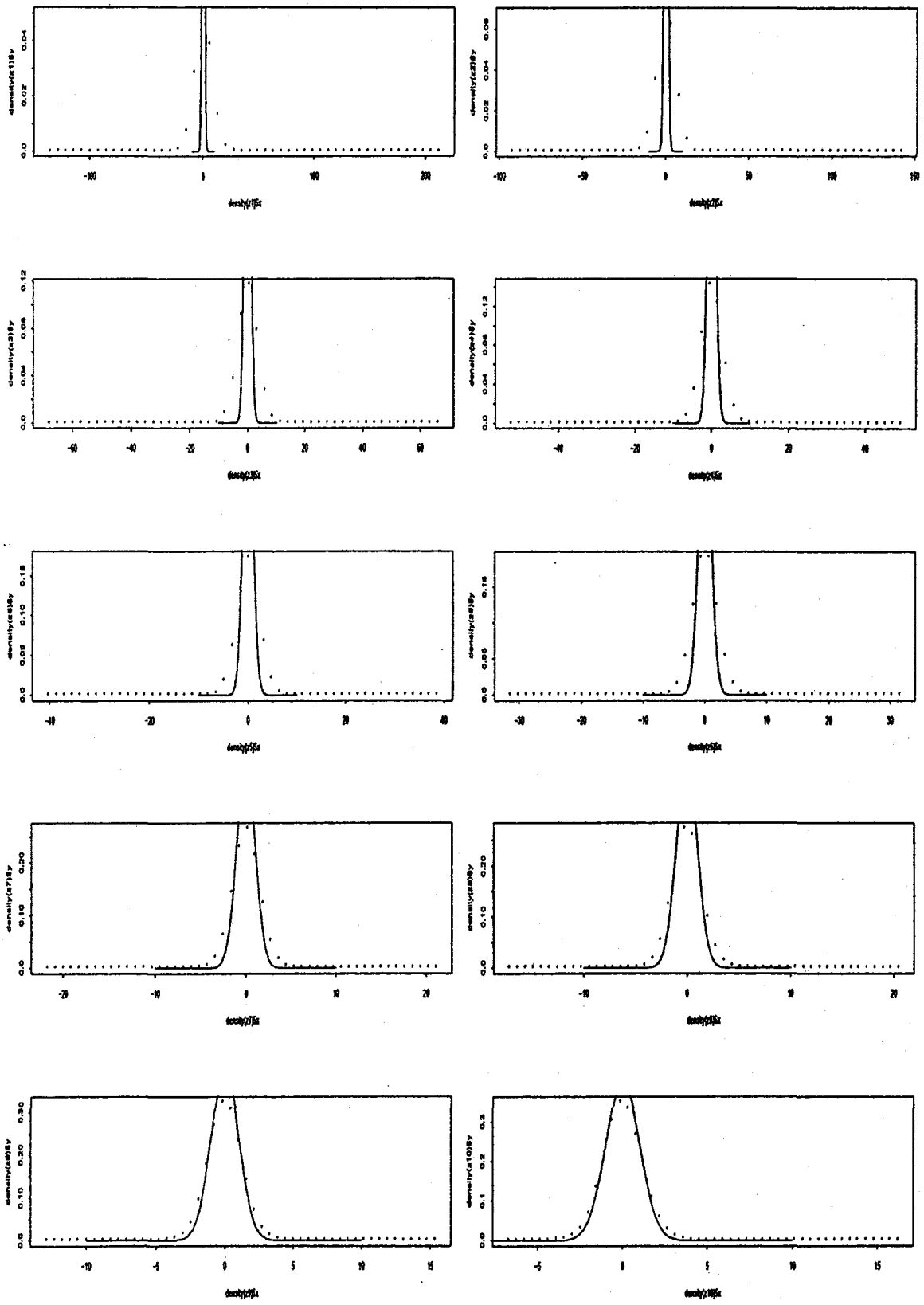

Figure 7: Plots of distributions of log returns with $\Delta t = 5, 10, 20, 40, 80, 160, 320, 640, 1280, 2560$(sec) and standardized with $\mu = 0$ and $\sigma = 1$(Dots are samples and lines are standard normal distributions)

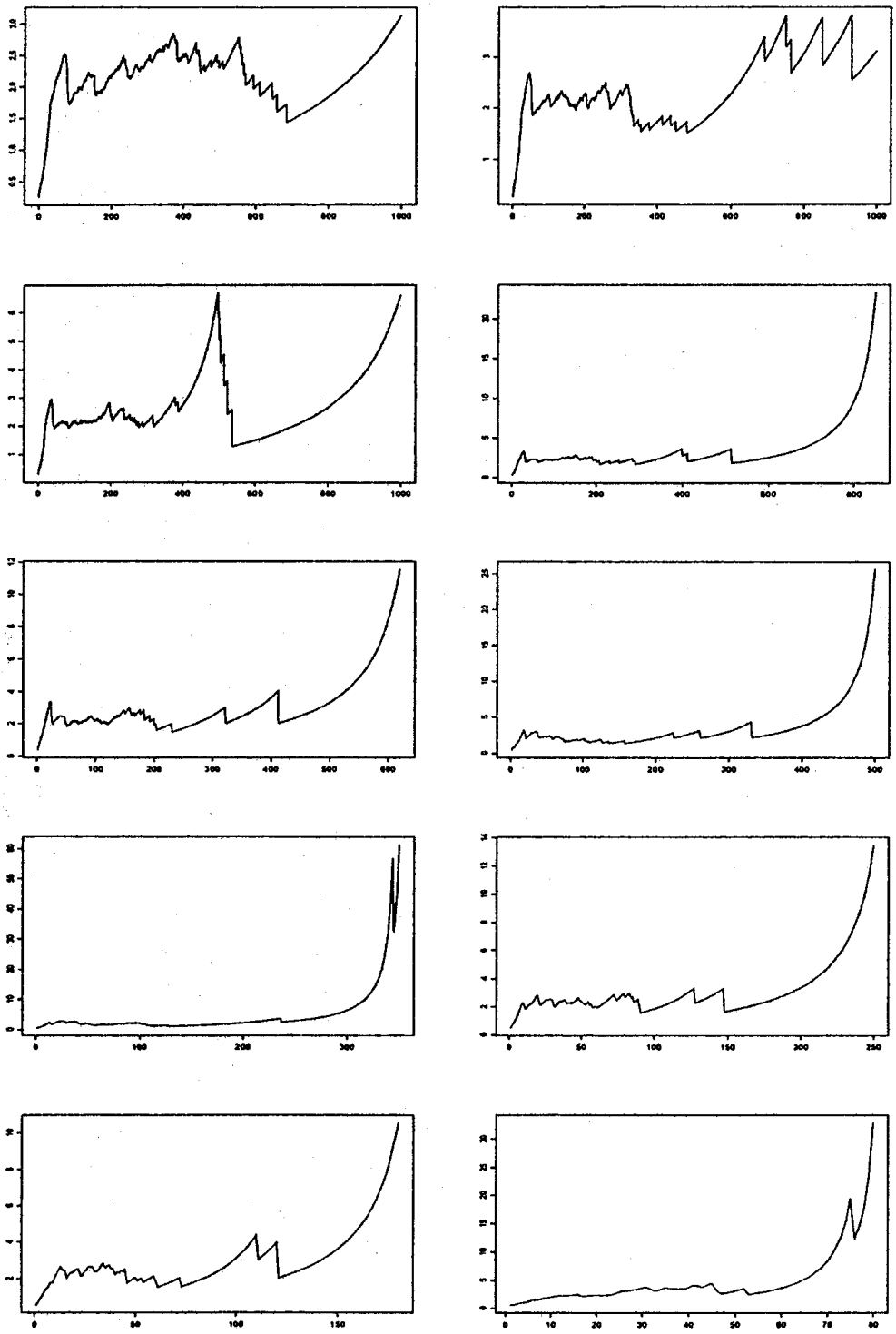

Figure 8: Plots of Hillplots for $z_1 - z_{10}$

Figure-8, it seems that the estimeated αs of those Hill plots are almost consistent with tail behavior from z_1 to z_5. But there are biases observed in the from z_6 to z_{10}. The reason which causes those biases could be the decreasing sampling numbers as time interval increases.

3.4 Volatility measures

In this section, we examine the behavior of the absolute returns based upon so-called Scaling Law or Volatility measures. It is an empirical study and is defined as follows.

$$\{E[|r|^p]\}^{1/p} = c(p)\,\Delta t^{\alpha(p)} \tag{5}$$

where $\alpha(p)$ is so-called the drift exponent. We manipulate logarithm operations to both sides of (5), then we have

$$\frac{1}{p} log E[|r|^p] = log c(p) + \alpha(p) log \Delta t \tag{6}$$

where, $c(p)$ and $\alpha(p)$ are determinstic functions. Thus we can do the regressions by equation (6) using $E[|r|^p]$ and different Δt to calculate the drift exponents. The calculation results are listed in Table-6.

Table-6 Drift exponents with different p

p	$\alpha(p)$	R^2
1	0.5456	0.9904
2	0.4228	0.9980

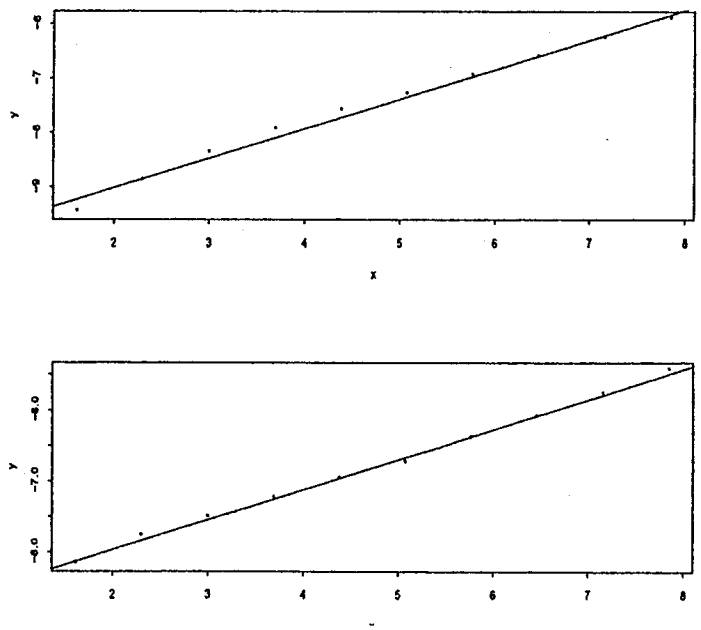

Figure 9: Plot of regressions of $\alpha(1)$, $\alpha(2)$

4 Seasonality in tick by tick data

Many preceded researches reported the phenomena of seasonality in tick data, for details, see Anderson (1997, 98) and Dacorogna (1993). In this section we simply discuss this issue and the methodology of deseasonalization.

4.1 Observed seasonality

At first, we just give some evidences of the seasonality in those 10 data sets. Here Figure-10 shows the plots of ACF of the absolute values of the log returns of all data sets being used in this paper. Clearly, there is a trend in ACF to indicate that seasonality appearing in the data sets.

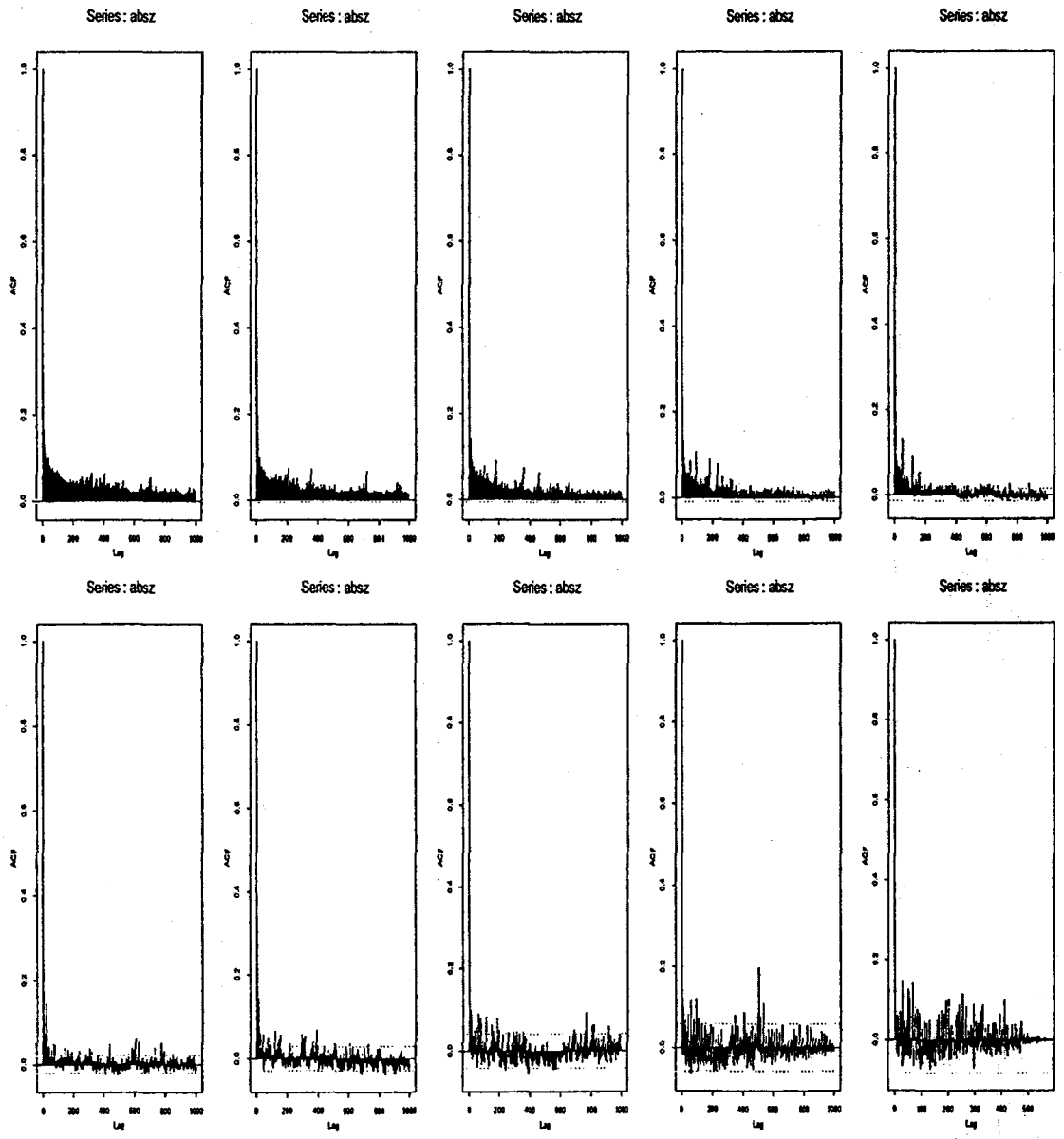

Figure 10: Plots of ACF of absolute returns of $z_1 - z_{10}$ (Upper: $z_1 - z_5$ (from left to right); Buttom: $z_6 - z_{10}$ (from left to right)

4.2 Deseasonalization

We assume the seasonality model as follows.

$$r_t = v_t s_t \varepsilon_t \tag{7}$$

where ε_t is i.i.d. random variable, and v_t and s_t are the long term volatility and the seasonality volatility respectively. In order to separate the long term volatility, we take the absolute operator, and then the logarithm operator to both sides of equation (7), thus we have,

$$log|r_t| = log|v_t| + log|s_t| + log|\varepsilon_t| \tag{8}$$

Then we apply the $log|r_t|$ for the MODWT, which is one of the Discrete Wavelet Transformations, with invariance in translation, non boundary adjustment and other good features by giving up the orthogonality. Details see Gencay and Selcuk and Whitcher(1999).

5 Volume of the intraday tradings

We summarize the trading volume within 63 days. The descripitive statistics are shown in Table-7. Seen from Table-7, 8 the volume between 9:30 and 12:00. of the intraday tradings, is larger than that between the 12:00 and 14:00. And it increases again when it comes near to 15:00.

Recently we are working on the modeling of relation between the price change and trading volume, Trading Volume could be an engine of price change and vice versa.

Table-7 Descripitive Statistics of Trading Volume

Total	
Mean	124.0714
Median	114.5
Mode	89
Standard Deviation	58.76815
Variance	3453.695
Kurtosis	3.746287
Skewness	1.471465
Range	451
Min	0
Max	451

Table-8 Trading Volume during different time intervals

	h11	h12	h13	h14	h15	h16
Mean	142	128	108	100	120	146
Median	123	106	97	92	108	134
Mode	113	78	88	101	100	139
Standard Deviation	72.1	58.0	44.7	46.7	52.8	61.8
Variance	5205	3364	1994	2180	2789	3815
Kurtosis	5.23	1.04	0.38	6.27	2.94	1.63
Skewness	1.7	1.27	0.94	1.77	1.35	0.95
Range	451	257	194	287	315	320
Min	0	52	42	31	1	0
Max	451	309	236	318	316	320

Where h11 means from 10:00 to 11:00, h12 means from 11:00 to 12:00 and so on.

6 Conclusion

In this paper, we analyze statistical properties of the tick market trading data, and we find that the log returns distributions are getting closer and closer to the standard normal distribution. One of the main evidences can be acquired from the evolution of the modes of these returns distributions sampling at different time intervals, though, they are not standard normal distributions. We also check the behavior of those distributions tails. We find that they are of fat tails and much more close to student t distributions. And for Scaling law, our results show that the drift exponents are getting smaller and smaller as p increases, eventually it becomes less than 0.5. Furthermore, we find the seasonal pattern becomes sustantially stronger in absolute values of the log returns as the sampling time interval is near to one hour. At last we analyze the relationship between the trading time and the trading volume. We find there are some trading patterns during the intraday market. Namely, the trading volumes are definitely different during the different trading hours.

The next work for us is to study the relation between the trading volume and the log returns or return distributions, by building up rational and reasonable mathematical models, to identify the evolution of distributions with some probability distribution family.

〔 References 〕

[1] T. G. Anderson and T. Bollerslev, *Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long-Run in High Frequency Returns*, J. of Finance, **52**(1997), 975-1005.

[2] T. G. Anderson and T. Bollerslev, *DM-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer-Run Dependencies*, J. of Finance, **53**(1998), 219-265.

[3] F. Black and M. Scholes, *The princing of options and corporate liabilities*, J. Polotical Econom. **81**(1973), 637-654.

[4] M. M. Dacorogna, U. A. Muller, R. J. Nagaler, R. B. Olsen and O. V. Pictet, A geographical mdel for the daily and weekly seasonal volatility in the foreign exchange markets, *Journal of international and money and finance*, **12**(1993), 413-438.

[5] R. Gencay and F. Selcuk and B. Whitcher, *Differentiating Intraday Seasonalities Through Wavelet Multi-Scaling*, Physica A, 2001.

[6] P. Praetz, *The distribution of share price changes*, J. Business **45** (1972), 49-45.

[7] S. I. Resnick, *Heavy tail modeling and teletraffic data*, Ann. Statist. 25(1997), 1805-1869.

[8] H. Stoll and R. Whaley, *Stock market structure of volatility*, Review of Financial studies, **3**(1990), 37-71.

[9] K. Tan, *Fractality in stock market*, J. Industrial and Economic Studies, **43**(2001), 33-42.