CHAPTER 5

High-Frequency Data Analysis and Market Microstructure

High-frequency data are observations taken at fine time intervals. In finance, they often mean observations taken daily or at a finer time scale. These data have become available primarily due to advances in data acquisition and processing techniques, and they have attracted much attention because they are important in empirical study of market microstructure and realized volatility. The ultimate high-frequency data in finance are the transaction-by-transaction or trade-by-trade data in security markets. Here time is often measured in seconds. The Trades and Quotes (TAQ) database of the New York Stock Exchange (NYSE) contains all equity transactions reported on the *Consolidated Tape* from 1992 to the present, which includes transactions on the NYSE, AMEX, NASDAQ, and the regional exchanges. The Berkeley Options Data Base provides similar data for options transactions from August 1976 to December 1996. More high-frequency options data are also available; see the website of Chicago Board Options Exchange. Transactions data for many other securities and markets, both domestic and foreign, are continuously collected and processed. Wood (2000) provides some historical perspective of high-frequency financial study.

High-frequency financial data are important in studying a variety of issues related to the trading process and market microstructure. They can be used to compare the efficiency of different trading systems in price discovery (e.g., the open out-cry system of the NYSE and the computer trading system of NASDAQ). They can also be used to study the dynamics of bid-and-ask quotes of a particular stock (e.g., Hasbrouck, 1999; Zhang, Russell, and Tsay, 2008). In an order-driven stock market (e.g., the Taiwan Stock Exchange), high-frequency data can be used to study the order dynamics and, more interesting, to investigate the question of "who provides the market liquidity." Cho, Russell, Tiao, and Tsay (2003) use intraday 5-minute returns of more than 340 stocks traded on the Taiwan Stock Exchange to study the

Analysis of Financial Time Series, Third Edition, By Ruey S. Tsay

impact of daily stock price limits and find significant evidence of magnet effects toward the price ceiling.

However, high-frequency data have some unique characteristics that do not appear in lower frequencies. Analysis of these data thus introduces new challenges to financial economists and statisticians. In this chapter, we study these special characteristics, consider methods for analyzing high-frequency data, and discuss implications of the results obtained. In particular, we discuss nonsynchronous trading, bid–ask spread, duration models, price movements, and bivariate models for price changes and time durations between transactions associated with price changes. The models discussed are also applicable to other scientific areas such as telecommunications and environmental studies.

5.1 NONSYNCHRONOUS TRADING

We begin with nonsynchronous trading. Stock tradings such as those on the NYSE do not occur in a synchronous manner; different stocks have different trading frequencies, and even for a single stock the trading intensity varies from hour to hour and from day to day. Yet we often analyze a return series in a fixed time interval such as daily, weekly, or monthly. For daily series, the price of a stock is its *closing* price, which is the last transaction price of the stock in a trading day. The actual time of the last transaction of the stock varies from day to day. As such we incorrectly assume daily returns as an equally spaced time series with a 24-hour interval. It turns out that such an assumption can lead to erroneous conclusions about the predictability of stock returns even if the true return series are serially independent.

For daily stock returns, nonsynchronous trading can introduce (a) lag-1 cross correlation between stock returns, (b) lag-1 serial correlation in a portfolio return, and (c) in some situations negative serial correlations of the return series of a single stock. Consider stocks A and B. Assume that the two stocks are independent, and stock A is traded more frequently than stock B. For special news affecting the market that arrives near the closing hour on one day, stock A is more likely than B to show the effect of the news on the same day simply because A is traded more frequently. The effect of the news on B will eventually appear, but it may be delayed until the following trading day. If this situation indeed happens, return of stock A appears to lead that of stock B. Consequently, the return series may show a significant lag-1 cross correlation from A to B even though the two stocks are independent. For a portfolio that holds stocks A and B, the prior cross correlation would become a significant lag-1 serial correlation.

In a more complicated manner, nonsynchronous trading can also induce erroneous negative serial correlations for a single stock. There are several models available in the literature to study this phenomenon; see Campbell, Lo, and MacKinlay (1997) and the references therein. Here we adopt a simplified version of the model proposed in Lo and MacKinlay (1990). Let r_t be the continuously compounded return of a security at the time index t. For simplicity, assume that $\{r_t\}$

is a sequence of independent and identically distributed random variables with mean $E(r_t) = \mu$ and variance $\text{Var}(r_t) = \sigma^2$. For each time period, the probability that the security is not traded is π, which is time invariant and independent of r_t. Let r_t^o be the observed return. When there is no trade at time index t, we have $r_t^o = 0$ because there is no information available. Yet when there is a trade at time index t, we define r_t^o as the cumulative return from the previous trade (i.e., $r_t^o = r_t + r_{t-1} + \cdots + r_{t-k_t}$, where k_t is the largest nonnegative integer such that no trade occurred in the periods $t - k_t, t - k_t + 1, \ldots, t - 1$). Mathematically, the relationship between r_t and r_t^o is

$$r_t^o = \begin{cases} 0 & \text{with probability } \pi \\ r_t & \text{with probability } (1-\pi)^2 \\ r_t + r_{t-1} & \text{with probability } (1-\pi)^2\pi \\ r_t + r_{t-1} + r_{t-2} & \text{with probability } (1-\pi)^2\pi^2 \\ & \vdots \\ \sum_{i=0}^{k} r_{t-i} & \text{with probability } (1-\pi)^2\pi^k \\ & \vdots \end{cases} \tag{5.1}$$

These probabilities are easy to understand. For example, $r_t^o = r_t$ if and only if there are trades at both t and $t-1$, $r_t^o = r_t + r_{t-1}$ if and only if there are trades at t and $t-2$, but no trade at $t-1$, and $r_t^o = r_t + r_{t-1} + r_{t-2}$ if and only if there are trades at t and $t-3$, but no trades at $t-1$ and $t-2$, and so on. As expected, the total probability is 1 given by

$$\pi + (1-\pi)^2(1 + \pi + \pi^2 + \cdots) = \pi + (1-\pi)^2\frac{1}{1-\pi} = \pi + 1 - \pi = 1.$$

We are ready to consider the moment equations of the observed return series $\{r_t^o\}$. First, the expectation of r_t^o is

$$\begin{aligned} E(r_t^o) &= (1-\pi)^2 E(r_t) + (1-\pi)^2\pi E(r_t + r_{t-1}) + \cdots \\ &= (1-\pi)^2\mu + (1-\pi)^2\pi 2\mu + (1-\pi)^2\pi^2 3\mu + \cdots \\ &= (1-\pi)^2\mu(1 + 2\pi + 3\pi^2 + 4\pi^3 + \cdots) \\ &= (1-\pi)^2\mu\frac{1}{(1-\pi)^2} = \mu. \end{aligned} \tag{5.2}$$

In the prior derivation, we use the result $1 + 2\pi + 3\pi^2 + 4\pi^3 + \cdots = 1/(1-\pi)^2$. Next, for the variance of r_t^o, we use $\text{Var}(r_t^o) = E[(r_t^o)^2] - [E(r_t^o)]^2$ and

$$\begin{aligned} E(r_t^o)^2 &= (1-\pi)^2 E[(r_t)^2] + (1-\pi)^2\pi E[(r_t + r_{t-1})^2] + \cdots \\ &= (1-\pi)^2[(\sigma^2 + \mu^2) + \pi(2\sigma^2 + 4\mu^2) + \pi^2(3\sigma^2 + 9\mu^2) + \cdots] \end{aligned} \tag{5.3}$$

$$= (1-\pi)^2[\sigma^2(1+2\pi+3\pi^2+\cdots)+\mu^2(1+4\pi+9\pi^2+\cdots)] \quad (5.4)$$

$$= \sigma^2+\mu^2\left[\frac{2}{1-\pi}-1\right]. \quad (5.5)$$

In Eq. (5.3), we use

$$E\left(\sum_{i=0}^{k} r_{t-i}\right)^2 = \text{Var}\left(\sum_{i=0}^{k} r_{t-i}\right)+\left[E\left(\sum_{i=0}^{k} r_{t-i}\right)\right]^2 = (k+1)\sigma^2+[(k+1)\mu]^2$$

under the serial independence assumption of r_t. Using techniques similar to that of Eq. (5.2), we can show that the first term of Eq. (5.4) reduces to σ^2. For the second term of Eq. (5.4), we use the identity

$$1+4\pi+9\pi^2+16\pi^3+\cdots = \frac{2}{(1-\pi)^3}-\frac{1}{(1-\pi)^2},$$

which can be obtained as follows. Let

$$H = 1+4\pi+9\pi^2+16\pi^3+\cdots \quad \text{and} \quad G = 1+3\pi+5\pi^2+7\pi^3+\cdots.$$

Then $(1-\pi)H = G$ and

$$\begin{aligned}(1-\pi)G &= 1+2\pi+2\pi^2+2\pi^3+\cdots \\ &= 2(1+\pi+\pi^2+\cdots)-1 = \frac{2}{1-\pi}-1.\end{aligned}$$

Consequently, from Eqs. (5.2) and (5.5), we have

$$\text{Var}(r_t^o) = \sigma^2+\mu^2\left(\frac{2}{1-\pi}-1\right)-\mu^2 = \sigma^2+\frac{2\pi\mu^2}{1-\pi}. \quad (5.6)$$

Consider next the lag-1 autocovariance of $\{r_t^o\}$. Here we use $\text{Cov}(r_t^o, r_{t-1}^o) = E(r_t^o r_{t-1}^o)-E(r_t^0)E(r_{t-1}^o) = E(r_t^o r_{t-1}^o)-\mu^2$. The question then reduces to finding $E(r_t^o r_{t-1}^o)$. Notice that $r_t^o r_{t-1}^o$ is zero if there is no trade at t, no trade at $t-1$, or no trade at both t and $t-1$. Therefore, we have

$$r_t^o r_{t-1}^o = \begin{cases} 0 & \text{with probability } 2\pi-\pi^2 \\ r_t r_{t-1} & \text{with probability } (1-\pi)^3 \\ r_t(r_{t-1}+r_{t-2}) & \text{with probability } (1-\pi)^3\pi \\ r_t(r_{t-1}+r_{t-2}+r_{t-3}) & \text{with probability } (1-\pi)^3\pi^2 \\ & \vdots \\ r_t(\sum_{i=1}^{k} r_{t-i}) & \text{with probability } (1-\pi)^3\pi^{k-1} \\ & \vdots \end{cases} \quad (5.7)$$

Again the total probability is unity. To understand the prior result, notice that $r_t^o r_{t-1}^o = r_t r_{t-1}$ if and only if there are three consecutive trades at $t-2$, $t-1$, and t. Using Eq. (5.7) and the fact that $E(r_t r_{t-j}) = E(r_t)E(r_{t-j}) = \mu^2$ for $j > 0$, we have

$$\begin{aligned} &E(r_t^o r_{t-1}^o) \\ &\quad = (1-\pi)^3 \left\{ E(r_t r_{t-1}) + \pi E[r_t(r_{t-1} + r_{t-2})] + \pi^2 E\left[r_t \left(\sum_{i=1}^{3} r_{t-i} \right) \right] + \cdots \right\} \\ &\quad = (1-\pi)^3 \mu^2 (1 + 2\pi + 3\pi^2 + \cdots) = (1-\pi)\mu^2. \end{aligned}$$

The lag-1 autocovariance of $\{r_t^o\}$ is then

$$\text{Cov}(r_t^o, r_{t-1}^o) = -\pi \mu^2. \tag{5.8}$$

Provided that μ is not zero, the nonsynchronous trading induces a *negative* lag-1 autocorrelation in r_t^o given by

$$\rho_1(r_t^o) = \frac{-(1-\pi)\pi\mu^2}{(1-\pi)\sigma^2 + 2\pi\mu^2}.$$

In general, we can extend the prior result and show that

$$\text{Cov}(r_t^o, r_{t-j}^o) = -\mu^2 \pi^j, \qquad j \geq 1.$$

The magnitude of the lag-1 ACF depends on the choices of μ, π, and σ and can be substantial. Thus, when $\mu \neq 0$, the nonsynchronous trading induces negative autocorrelations in an observed security return series.

The previous discussion can be generalized to the return series of a portfolio that consists of N securities; see Campbell et al. (1997, Chapter 3). In the time series literature, effects of nonsynchronous trading on the return of a single security are equivalent to that of random temporal aggregation on a time series, with the trading probability π governing the mechanism of aggregation.

5.2 BID−ASK SPREAD

In some stock exchanges (e.g., NYSE), market makers play an important role in facilitating trades. They provide market liquidity by standing ready to buy or sell whenever the public wishes to buy or sell. By market liquidity, we mean the ability to buy or sell significant quantities of a security quickly, anonymously, and with little price impact. In return for providing liquidity, market makers are granted monopoly rights by the exchange to post different prices for purchases and sales of a security. They buy at the *bid* price P_b and sell at a higher ask price P_a. (For the

public, P_b is the sale price and P_a is the purchase price.) The difference $P_a - P_b$ is call the *bid–ask spread*, which is the primary source of compensation for market makers. Typically, the bid–ask spread is small—namely, one or two cents.

The existence of a bid–ask spread, although small in magnitude, has several important consequences in time series properties of asset returns. We briefly discuss the bid–ask bounce—namely, the bid–ask spread introduces *negative* lag-1 serial correlation in an asset return. Consider the simple model of Roll (1984). The observed market price P_t of an asset is assumed to satisfy

$$P_t = P_t^* + I_t \frac{S}{2}, \tag{5.9}$$

where $S = P_a - P_b$ is the bid–ask spread, P_t^* is the time-t fundamental value of the asset in a frictionless market, and $\{I_t\}$ is a sequence of independent binary random variables with equal probabilities (i.e., $I_t = 1$ with probability 0.5 and $= -1$ with probability 0.5). The I_t can be interpreted as an order-type indicator, with 1 signifying buyer-initiated transaction and -1 seller-initiated transaction. Alternatively, the model can be written as

$$P_t = P_t^* + \begin{cases} +S/2 & \text{with probability 0.5,} \\ -S/2 & \text{with probability 0.5.} \end{cases}$$

If there is no change in P_t^*, then the observed process of price changes is

$$\Delta P_t = (I_t - I_{t-1}) \frac{S}{2}. \tag{5.10}$$

Under the assumption of I_t in Eq. (5.9), $E(I_t) = 0$ and $\text{Var}(I_t) = 1$, and we have $E(\Delta P_t) = 0$ and

$$\text{Var}(\Delta P_t) = S^2/2, \tag{5.11}$$

$$\text{Cov}(\Delta P_t, \Delta P_{t-1}) = -S^2/4, \tag{5.12}$$

$$\text{Cov}(\Delta P_t, \Delta P_{t-j}) = 0, \qquad j > 1. \tag{5.13}$$

Therefore, the autocorrelation function of ΔP_t is

$$\rho_j(\Delta P_t) = \begin{cases} -0.5 & \text{if } j = 1, \\ 0 & \text{if } j > 1. \end{cases} \tag{5.14}$$

The bid–ask spread thus introduces a negative lag-1 serial correlation in the series of observed price changes. This is referred to as the *bid–ask bounce* in the finance literature. Intuitively, the bounce can be seen as follows. Assume that the fundamental price P_t^* is equal to $(P_a + P_b)/2$. Then P_t assumes the value P_a or P_b. If the previously observed price is P_a (the higher value), then the current observed price is either unchanged or lower at P_b. Thus, ΔP_t is either 0 or $-S$. However, if

the previous observed price is P_b (the lower value), then ΔP_t is either 0 or S. The negative lag-1 correlation in ΔP_t becomes apparent. The bid–ask spread does not introduce any serial correlation beyond lag 1, however.

A more realistic formulation is to assume that P_t^* follows a random walk so that $\Delta P_t^* = P_t^* - P_{t-1}^* = \epsilon_t$, which forms a sequence of independent and identically distributed random variables with mean zero and variance σ^2. In addition, $\{\epsilon_t\}$ is independent of $\{I_t\}$. In this case, $\text{Var}(\Delta P_t) = \sigma^2 + S^2/2$, but $\text{Cov}(\Delta P_t, \Delta P_{t-j})$ remains unchanged. Therefore,

$$\rho_1(\Delta P_t) = \frac{-S^2/4}{S^2/2 + \sigma^2} \leq 0.$$

The magnitude of the lag-1 autocorrelation of ΔP_t is reduced, but the negative effect remains when $S = P_a - P_b > 0$. In finance, it might be of interest to study the components of the bid–ask spread. Interested readers are referred to Campbell et al. (1997) and the references therein.

The effect of bid–ask spread continues to exist in portfolio returns and in multivariate financial time series. Consider the bivariate case. Denote the bivariate order-type indicator by $I_t = (I_{1t}, I_{2t})'$, where I_{1t} is for the first security and I_{2t} for the second security. If I_{1t} and I_{2t} are contemporaneously positively correlated, then the bid–ask spreads can introduce negative lag-1 cross correlations.

5.3 EMPIRICAL CHARACTERISTICS OF TRANSACTIONS DATA

Let t_i be the calendar time, measured in seconds from midnight, at which the ith transaction of an asset takes place. Associated with the transaction are several variables such as the transaction price, the transaction volume, the prevailing bid and ask quotes, and so on. The collection of t_i and the associated measurements are referred to as the *transactions data*. These data have several important characteristics that do not exist when the observations are aggregated over time. Some of the characteristics are given next.

1. *Unequally Spaced Time Intervals.* Transactions such as stock tradings on an exchange do not occur at equally spaced time intervals. As such, the observed transaction prices of an asset do not form an equally spaced time series. The time duration between trades becomes important and might contain useful information about market microstructure (e.g., trading intensity).
2. *Discrete-Valued Prices.* The price change of an asset from one transaction to the next only occurred in multiples of tick size before January 29, 2001. On the NYSE, the tick size was one-eighth of a dollar before June 24, 1997 and was one-sixteenth of a dollar before January 29, 2001. Therefore, the price was a discrete-valued variable in transactions data. Although all equity markets in the United States now use the decimal system, the price change in consecutive trades tends to occur in multiples of one cent and can be treated

approximately as a discrete-valued variable. In some markets, price change may also be subject to limit constraints set by regulators.

3. *Existence of a Daily Periodic or Diurnal Pattern.* Under the normal trading conditions, transaction activity can exhibit a periodic pattern. For instance, on the NYSE, transactions are "heavier" at the beginning and closing of the trading hours and "thinner" during lunch hour, resulting in a U-shaped transaction intensity. Consequently, time durations between transactions also exhibit a daily cyclical pattern.
4. *Multiple Transactions within a Single Second.* It is possible that multiple transactions, even with different prices, occur at the same time. This is partly due to the fact that time is measured in seconds, which may be too long a time scale in periods of heavy trading.

To demonstrate these characteristics, we consider first the IBM transactions data from November 1, 1990, to January 31, 1991. These data are from the Trades, Orders Reports, and Quotes (TORQ) data set; see Hasbrouck (1992). There are 63 trading days and 60,328 transactions. To simplify the discussion, we ignore the price changes between trading days and focus on the transactions that occurred in the normal trading hours from 9:30 AM to 4:00 PM Eastern time. It is well known that overnight stock returns differ substantially from intraday returns; see Stoll and Whaley (1990) and the references therein. Table 5.1 gives the frequencies in percentages of price change measured in the tick size of $\$\frac{1}{8} = \$0.125$. From the table, we make the following observations:

1. About two-thirds of the intraday transactions were without price change.
2. The price changed in one tick approximately 29% of the intraday transactions.
3. Only 2.6% of the transactions were associated with two-tick price changes.
4. Only about 1.3% of the transactions resulted in price changes of three ticks or more.
5. The distribution of positive and negative price changes was approximately symmetric.

Consider next the number of transactions in a 5-minute time interval. Denote the series by x_t. That is, x_1 is the number of IBM transactions from 9:30 AM to 9:35 AM on November 1, 1990, Eastern time; x_2 is the number of transactions from 9:35 AM to 9:40 AM; and so on. The time gaps between trading days are ignored. Figure 5.1(a) shows the time plot of x_t, and Figure 5.1(b) shows the sample ACF

TABLE 5.1 Frequencies of Price Change in Multiples of Tick Size for IBM Stock from November 1, 1990, to January 31, 1991

Number (tick)	≤ -3	-2	-1	0	1	2	≥ 3
Percentage	0.66	1.33	14.53	67.06	14.53	1.27	0.63

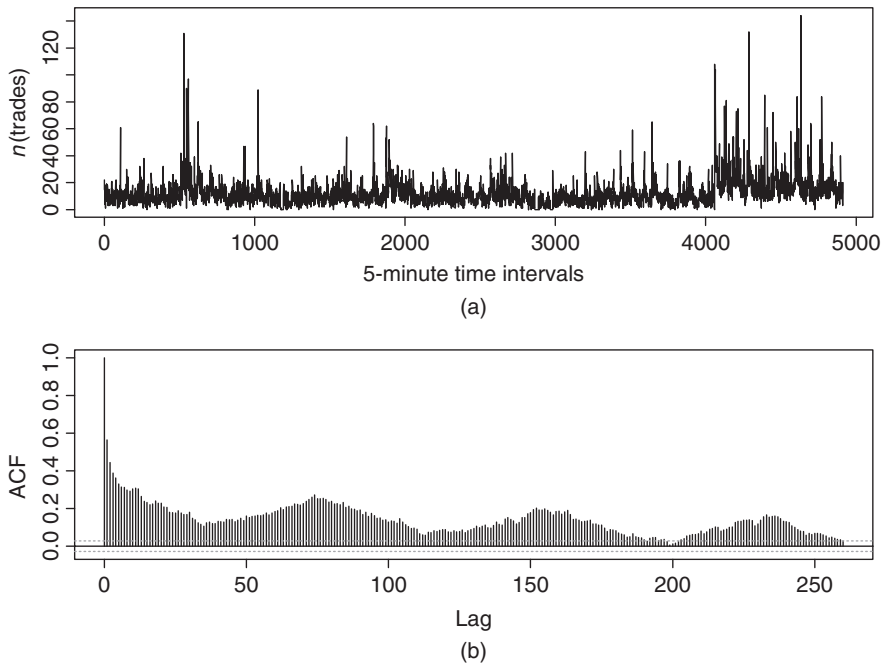

Figure 5.1 IBM intraday transactions data from 11/01/90 to 1/31/91: (a) number of transactions in 5-minute time intervals and (b) sample ACF of series in part (a).

of x_t for lags 1–260. Of particular interest is the cyclical pattern of the ACF with a periodicity of 78, which is the number of 5-minute intervals in a trading day. The number of transactions thus exhibits a daily pattern. To further illustrate the daily trading pattern, Figure 5.2 shows the average number of transactions within 5-minute time intervals over the 63 days. There are 78 such averages. The plot exhibits a "smiling" or U shape, indicating heavier trading at the opening and closing of the market and thinner trading during the lunch hours.

Since we focus on transactions that occurred during normal trading hours of a trading day, there are 59,838 time intervals in the data. These intervals are called the intraday *durations* between trades. For IBM stock, there were 6531 zero time intervals. That is, during the normal trading hours of the 63 trading days from November 1, 1990, to January 31, 1991, multiple transactions in a second occurred 6531 times, which is about 10.91%. Among these multiple transactions, 1002 of them had different prices, which is about 1.67% of the total number of intraday transactions. Therefore, multiple transactions (i.e., zero durations) may become an issue in statistical modeling of the time durations between trades.

Table 5.2 provides a two-way classification of price movements. Here price movements are classified into "up," "unchanged," and "down." We denote them by $+$, 0, and $-$, respectively. The table shows the price movements between two

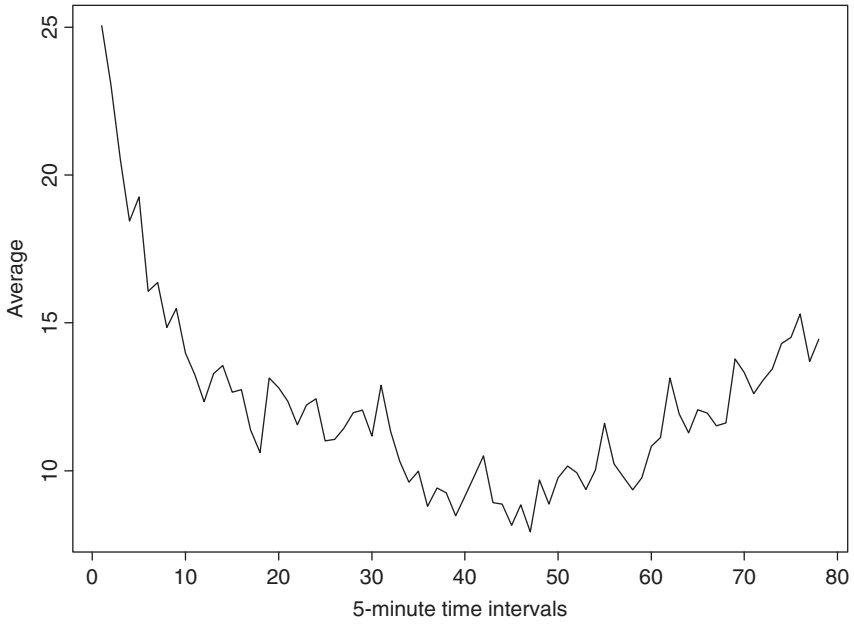

Figure 5.2 Time plot of average number of transactions in 5-minute time intervals. There are 78 observations, averaging over 63 trading days from 11/01/90 to 1/31/91 for IBM stock.

TABLE 5.2 Two-Way Classification of Price Movements in Consecutive Intraday Trades for IBM Stock[a]

	ith Trade			
$(i-1)$th trade	+	0	−	Margin
+	441	5498	3948	9887
0	4867	29779	5473	40119
−	4580	4841	410	9831
Margin	9888	40118	9831	59837

[a]The price movements are classified into "up," "unchanged," and "down." The data span is from November 1, 1990, to January 31, 1991.

consecutive trades [i.e., from the $(i-1)$th to the ith transaction] in the sample. From the table, trade-by-trade data show that:

1. Consecutive price increases or decreases are relatively rare, which are about $441/59837 = 0.74\%$ and $410/59837 = 0.69\%$, respectively.
2. There is a slight edge to move from up to unchanged rather than to down; see row 1 of the table.
3. There is a high tendency for the price to remain unchanged.

4. The probabilities of moving from down to up or unchanged are about the same; see row 3.

The first observation mentioned before is a clear demonstration of bid–ask bounce, showing *price reversals* in intraday transactions data. To confirm this phenomenon, we consider a directional series D_i for price movements, where D_i assumes the value $+1$, 0, and -1 for up, unchanged, and down price movement, respectively, for the ith transaction. The ACF of $\{D_i\}$ has a single spike at lag 1 with value -0.389, which is highly significant for a sample size of 59,837 and confirms the price reversal in consecutive trades.

As a second illustration, we consider the transactions data of IBM stock in December 1999 obtained from the TAQ database. The normal trading hours are from 9:30 AM to 4:00 PM Eastern time, except for December 31 when the market closed at 1:00 PM. Comparing with the 1990–1991 data, two important changes have occurred. First, the number of intraday tradings has increased sixfold. There were 134,120 intraday tradings in December 1999 alone. The increased trading intensity also increased the chance of multiple transactions within a second. The percentage of trades with zero time duration doubled to 22.98%. At the extreme, there were 42 transactions within a given second that happened twice on December 3, 1999. Second, the tick size of price movement was $\$\frac{1}{16} = \$0.0625$ instead of $\$\frac{1}{8}$. The change in tick size should reduce the bid–ask spread. Figure 5.3 shows the daily number of transactions in the new sample. Figure 5.4(a) shows the time plot of time durations between trades, measured in seconds, and Figure 5.4(b) is the

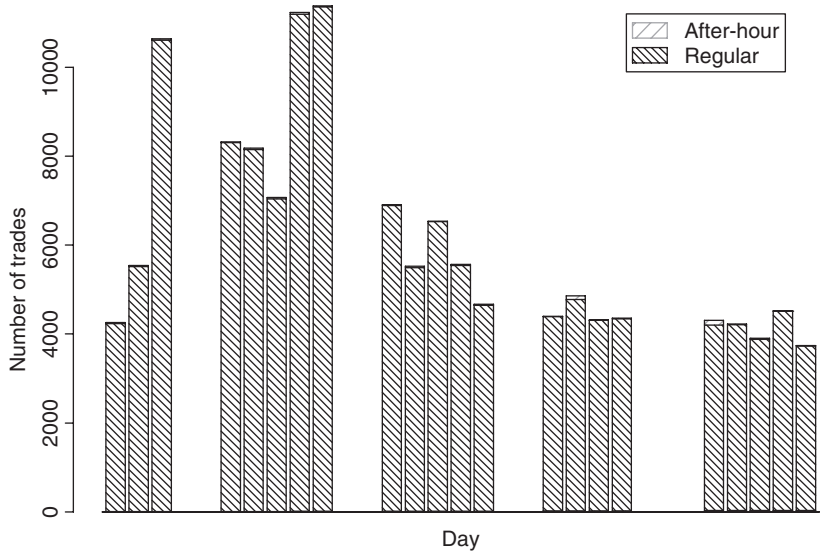

Figure 5.3 IBM transactions data for December 1999. Box plot shows the number of transactions in each trading day with after-hours portion denoting number of trades with time stamp after 4:00 PM.

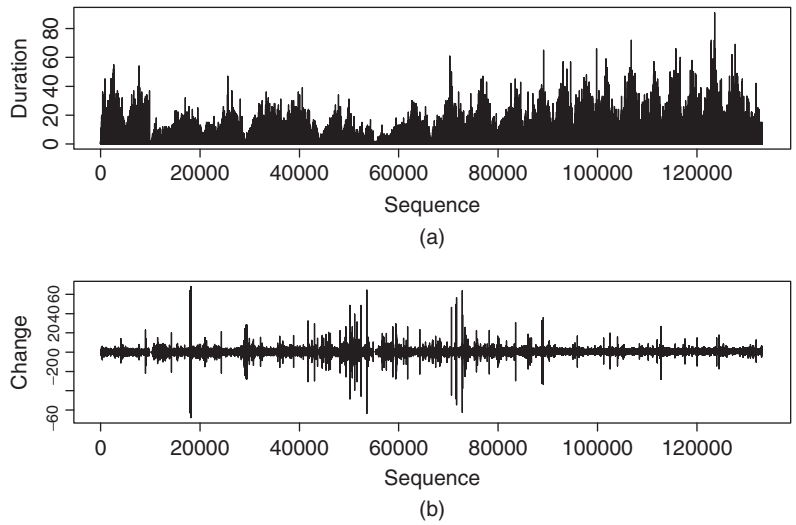

Figure 5.4 IBM transactions data for December 1999. (a) Time plot of time durations between trades and (b) time plot of price changes in consecutive trades measured in multiples of tick size of \$1/16. Only data during normal trading hours are included.

time plot of price changes in consecutive intraday trades, measured in multiples of the tick size of $\$\frac{1}{16}$. As expected, Figures 5.3 and 5.4(a) show clearly the inverse relationship between the daily number of transactions and the time interval between trades. Figure 5.4(b) shows two unusual price movements for IBM stock on December 3, 1999. They were a drop of 63 ticks followed by an immediate jump of 64 ticks and a drop of 68 ticks followed immediately by a jump of 68 ticks. Unusual price movements like these occurred infrequently in intraday transactions.

Focusing on trades recorded within regular trading hours, we have 61,149 trades out of 133,475 with no price change. This is about 45.8% and substantially lower than that between November 1990 and January 1991. It seems that reducing the tick size increased the chance of a price change. Table 5.3 gives the percentages of trades associated with a price change. The price movements remain approximately

TABLE 5.3 Percentages of Intraday Transactions Associated with a Price Change for IBM Stock Traded in December 1999[a]

Size	1	2	3	4	5	6	7	>7
	Downward Movements							
Percentage	18.03	5.80	1.79	0.66	0.25	0.15	0.09	0.32
	Upward Movements							
Percentage	18.24	5.57	1.79	0.71	0.24	0.17	0.10	0.31

[a]The percentage of transactions without price change is 45.8% and the total number of transactions recorded within regular trading hours is 133,475. The size is measured in multiples of tick size \$ 1/16.

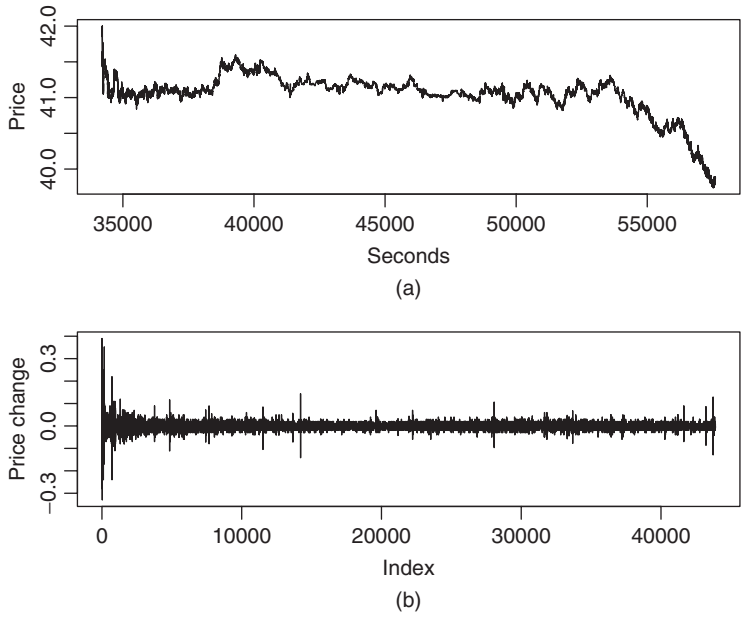

Figure 5.5 Transactions data of Boeing stock on December 1, 2008. (a) Price series over calendar time measured in seconds from midnight and (b) time plot of price changes in consecutive trades measured in cents. Only data during normal trading hours are included.

symmetric with respect to zero. Large price movements in intraday tradings are still relatively rare.

Finally, we consider the transactions data of Boeing stock on December 1, 2008. There are 43,894 transactions within the regular trading hours. Figure 5.5(a) shows the transaction prices versus the calendar time measured in seconds from the midnight, and Figure 5.5(b) shows the time plot of price changes. In this particular instance, the price shows a downward trend within the day, but the price changes continue to exhibit patterns similar to those before using the decimal system. Figure 5.6 shows the histogram of the price changes for the Boeing stock. The histogram shows some distinct characteristics. First, the price changes appear to be symmetric with respective to zero. Second, the price changes indeed concentrate on multiples of one cent. Out of the 43,894 transactions, 58.5% have no price change; see the big spike of the histogram. Details of the summary of price changes for the Boeing stock are given in Table 5.4. The remaining 4.59% of the price changes not shown in Table 5.4 are not in multiples of one cent.

Remark. The recordkeeping of high-frequency data is often not as good as that of observations taken at lower frequencies. Data cleaning becomes a necessity in high-frequency data analysis. For transactions data, missing observations may happen in many ways, and the accuracy of the exact transaction time might be questionable for some trades. For example, recorded trading times may be beyond

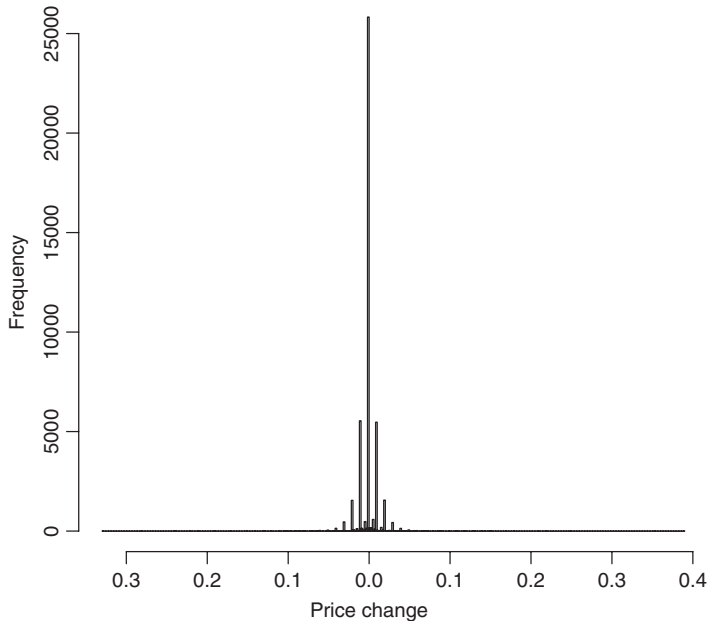

Figure 5.6 Histogram of price changes for Boeing stock on December 1, 2008.

TABLE 5.4 Frequencies of Price Change for Boeing Stock on December 1, 2008

Cents	< -3	-3	-2	-1	0	1	2	3	>3
Percentage	1.63	1.05	3.51	12.6	58.5	12.2	3.45	0.94	1.53

4:00 PM Eastern time even before the opening of after-hours tradings. How to handle these observations deserves a careful study. A proper method of data cleaning requires a deep understanding of the way in which the market operates. As such, it is important to specify clearly and precisely the methods used in data cleaning. These methods must be taken into consideration in making inference. □

Again, let t_i be the calendar time, measured in seconds from midnight, when the ith transaction took place. Let P_{t_i} be the transaction price. The price change from the $(i-1)$th to the ith trade is $y_i \equiv \Delta P_{t_i} = P_{t_i} - P_{t_{i-1}}$ and the time duration is $\Delta t_i = t_i - t_{i-1}$. Here it is understood that the subscript i in Δt_i and y_i denotes the time sequence of transactions, not the calendar time. In what follows, we consider models for y_i and Δt_i both individually and jointly.

5.4 MODELS FOR PRICE CHANGES

The discreteness and concentration on "no change" make it difficult to model the intraday price changes. Campbell et al. (1997) discuss several econometric

models that have been proposed in the literature. Here we mention two models that have the advantage of employing explanatory variables to study the intraday price movements. The first model is the ordered probit model used by Hauseman, Lo, and MacKinlay (1992) to study the price movements in transactions data. The second model has been considered recently by McCulloch and Tsay (2000) and is a simplified version of the model proposed by Rydberg and Shephard (2003); see also Ghysels (2000).

5.4.1 Ordered Probit Model

Let y_i^* be the unobservable price change of the asset under study (i.e., $y_i^* = P_{t_i}^* - P_{t_{i-1}}^*$), where P_t^* is the *virtual* price of the asset at time t. The ordered probit model assumes that y_i^* is a continuous random variable and follows the model

$$y_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \tag{5.15}$$

where $\boldsymbol{x}_i$ is a p-dimensional row vector of explanatory variables available at time t_{i-1}, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector, $E(\epsilon_i|\boldsymbol{x}_i) = 0$, $\text{Var}(\epsilon_i|\boldsymbol{x}_i) = \sigma_i^2$, and $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. The conditional variance σ_i^2 is assumed to be a positive function of the explanatory variable $\boldsymbol{w}_i$, that is,

$$\sigma_i^2 = g(\boldsymbol{w}_i), \tag{5.16}$$

where $g(\cdot)$ is a positive function. For financial transactions data, $\boldsymbol{w}_i$ may contain the time interval $t_i - t_{i-1}$ and some conditional heteroscedastic variables. Typically, one also assumes that the conditional distribution of ϵ_i given $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ is Gaussian.

Suppose that the observed price change y_i may assume k possible values. In theory, k can be infinity, but countable. In practice, k is finite and may involve combining several categories into a single value. For example, we have $k = 7$ in Table 5.1, where the first value "-3 ticks" means that the price change is -3 ticks or lower. We denote the k possible values as $\{s_1, \ldots, s_k\}$. The ordered probit model postulates the relationship between y_i and y_i^* as

$$y_i = s_j \quad \text{if} \quad \alpha_{j-1} < y_i^* \leq \alpha_j, \qquad j = 1, \ldots, k, \tag{5.17}$$

where α_j are real numbers satisfying $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{k-1} < \alpha_k = \infty$. Under the assumption of conditional Gaussian distribution, we have

$$P(y_i = s_j|\boldsymbol{x}_i, \boldsymbol{w}_i) = P(\alpha_{j-1} < \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \leq \alpha_j|\boldsymbol{x}_i, \boldsymbol{w}_i)$$

$$= \begin{cases} P(\boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \leq \alpha_1|\boldsymbol{x}_i, \boldsymbol{w}_i) & \text{if } j = 1, \\ P(\alpha_{j-1} < \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \leq \alpha_j|\boldsymbol{x}_i, \boldsymbol{w}_i) & \text{if } j = 2, \ldots, k-1, \\ P(\alpha_{k-1} < \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i|\boldsymbol{x}_i, \boldsymbol{w}_i) & \text{if } j = k, \end{cases}$$

$$= \begin{cases} \Phi\left[\dfrac{\alpha_1 - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] & \text{if } j = 1, \\ \Phi\left[\dfrac{\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] - \Phi\left[\dfrac{\alpha_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] & \text{if } j = 2, \ldots, k-1, \\ 1 - \Phi\left[\dfrac{\alpha_{k-1} - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] & \text{if } j = k, \end{cases} \tag{5.18}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal random variable evaluated at x, and we write $\sigma_i(\boldsymbol{w}_i)$ to denote that σ_i^2 is a positive function of $\boldsymbol{w}_i$. From the definition, an ordered probit model is driven by an unobservable continuous random variable. The observed values, which have a natural ordering, can be regarded as categories representing the underlying process.

The ordered probit model contains parameters $\boldsymbol{\beta}$, α_i $(i = 1, \ldots, k-1)$, and those in the conditional variance function $\sigma_i(\boldsymbol{w}_i)$ in Eq. (5.16). These parameters can be estimated by the maximum-likelihood or Markov chain Monte Carlo methods.

Example 5.1. Hauseman et al. (1992) apply the ordered probit model to the 1988 transactions data of more than 100 stocks. Here we only report their result for IBM. There are 206,794 trades. The sample mean (standard deviation) of price change y_i, time duration Δt_i, and bid–ask spread are $-0.0010(0.753)$, $27.21(34.13)$, and $1.9470(1.4625)$, respectively. The bid–ask spread is measured in ticks. The model used has nine categories for price movement, and the functional specifications are

$$\begin{aligned} \boldsymbol{x}_i\boldsymbol{\beta} = {} & \beta_1 \Delta t_i^* + \sum_{v=1}^{3} \beta_{v+1} y_{i-v} + \sum_{v=1}^{3} \beta_{v+4} \mathrm{SP5}_{i-v} + \sum_{v=1}^{3} \beta_{v+7} \mathrm{IBS}_{i-v} \\ & + \sum_{v=1}^{3} \beta_{v+10} [T_\lambda(V_{i-v}) \times \mathrm{IBS}_{i-v}], \end{aligned} \tag{5.19}$$

$$\sigma_i^2(\boldsymbol{w}_i) = 1.0 + \gamma_1^2 \Delta t_i^* + \gamma_2^2 \mathrm{AB}_{i-1}, \tag{5.20}$$

where $T_\lambda(V) = (V^\lambda - 1)/\lambda$ is the Box–Cox (1964) transformation of V with $\lambda \in [0, 1]$ and the explanatory variables are defined by the following:

- $\Delta t_i^* = (t_i - t_{i-1})/100$ is a rescaled time duration between the $(i-1)$th and ith trades with time measured in seconds.
- AB_{i-1} is the bid–ask spread prevailing at time t_{i-1} in ticks.

- y_{i-v} ($v = 1, 2, 3$) is the lagged value of price change at t_{i-v} in ticks. With $k = 9$, the possible values of price changes are $\{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$ in ticks.
- V_{i-v} ($v = 1, 2, 3$) is the lagged value of dollar volume at the $(i - v)$th transaction, defined as the price of the $(i - v)$th transaction in dollars times the number of shares traded (denominated in hundreds of shares). That is, the dollar volume is in hundreds of dollars.
- SP5_{i-v} ($v = 1, 2, 3$) is the 5-minute continuously compounded returns of the Standard and Poor's 500 index futures price for the contract maturing in the closest month beyond the month in which transaction $(i - v)$ occurred, where the return is computed with the futures price recorded 1 minute before the nearest round minute *prior* to t_{i-v} and the price recorded 5 minutes before this.
- IBS_{i-v} ($v = 1, 2, 3$) is an indicator variable defined by

$$\text{IBS}_{i-v} = \begin{cases} 1 & \text{if } P_{i-v} > (P^a_{i-v} + P^b_{i-v})/2, \\ 0 & \text{if } P_{i-v} = (P^a_{i-v} + P^b_{i-v})/2, \\ -1 & \text{if } P_{i-v} < (P^a_{i-v} + P^b_{i-v})/2, \end{cases}$$

  where P^a_j and P^b_j are the ask and bid price at time t_j.

The parameter estimates and their t ratios are given in Table 5.5. All the t ratios are large except one, indicating that the estimates are highly significant. Such high t ratios are not surprising as the sample size is large. For the heavily traded IBM stock, the estimation results suggest the following conclusions:

1. The boundary partitions are not equally spaced but are almost symmetric with respect to zero.

TABLE 5.5 Parameter Estimates of Ordered Probit Model in Eqs. (5.19) and (5.20) for the 1988 Transaction Data of IBM, Where t Denotes the t Ratio

	Boundary Partitions of the Probit Model							
Parameter	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
Estimate	−4.67	−4.16	−3.11	−1.34	1.33	3.13	4.21	4.73
t	−145.7	−157.8	−171.6	−155.5	154.9	167.8	152.2	138.9
	Equation Parameters of the Probit Model							
Parameter	γ_1	γ_2	$\beta_1 : \Delta t^*_i$	$\beta_2 : y_{-1}$	β_3	β_4	β_5	β_6
Estimate	0.40	0.52	−0.12	−1.01	−0.53	−0.21	1.12	−0.26
t	15.6	71.1	−11.4	−135.6	−85.0	−47.2	54.2	−12.1
Parameter	β_7	β_8	β_9:	β_{10}	β_{11}	β_{12}	β_{13}	
Estimate	0.01	−1.14	−0.37	−0.17	0.12	0.05	0.02	
t	0.26	−63.6	−21.6	−10.3	47.4	18.6	7.7	

Source: Reprinted with permission of Elsevier from Journal of Financial Economics (1992, Vol. 31, p. 345)

2. The transaction duration Δt_i affects both the conditional mean and conditional variance of y_i in Eqs. (5.19) and (5.20).
3. The coefficients of lagged price changes are negative and highly significant, indicating *price reversals*.
4. As expected, the bid–ask spread at time t_{i-1} significantly affects the conditional variance.

5.4.2 Decomposition Model

An alternative approach to modeling price change is to decompose it into three components and use conditional specifications for the components; see Rydberg and Shephard (2003). The three components are an indicator for price change, the direction of price movement if there is a change, and the size of price change if a change occurs. Specifically, the price change at the ith transaction can be written as

$$y_i \equiv P_{t_i} - P_{t_{i-1}} = A_i D_i S_i, \tag{5.21}$$

where A_i is a binary variable defined as

$$A_i = \begin{cases} 1 & \text{if there is a price change at the } i\text{th trade,} \\ 0 & \text{if price remains the same at the } i\text{th trade,} \end{cases} \tag{5.22}$$

D_i is also a discrete variable signifying the *direction* of the price change if a change occurs, that is,

$$D_i|(A_i = 1) = \begin{cases} 1 & \text{if price increases at the } i\text{th trade,} \\ -1 & \text{if price drops at the } i\text{th trade,} \end{cases} \tag{5.23}$$

where $D_i|(A_i = 1)$ means that D_i is defined under the condition of $A_i = 1$, and S_i is the size of the price change in ticks if there is a change at the ith trade and $S_i = 0$ if there is no price change at the ith trade. When there is a price change, S_i is a positive integer-valued random variable.

Note that D_i is not needed when $A_i = 0$, and there is a natural ordering in the decomposition. D_i is well defined only when $A_i = 1$ and S_i is meaningful when $A_i = 1$ and D_i is given. Model specification under the decomposition makes use of the ordering.

Let F_i be the information set available at the ith transaction. Examples of elements in F_i are Δt_{i-j}, A_{i-j}, D_{i-j}, and S_{i-j} for $j \geq 0$. The evolution of price change under model (5.21) can then be partitioned as

$$\begin{aligned} P(y_i|F_{i-1}) &= P(A_i D_i S_i|F_{i-1}) \\ &= P(S_i|D_i, A_i, F_{i-1})P(D_i|A_i, F_{i-1})P(A_i|F_{i-1}). \end{aligned} \tag{5.24}$$

Since A_i is a binary variable, it suffices to consider the evolution of the probability $p_i = P(A_i = 1)$ over time. We assume that

$$\ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i\boldsymbol{\beta} \quad \text{or} \quad p_i = \frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}, \tag{5.25}$$

where $\boldsymbol{x}_i$ is a finite-dimensional vector consisting of elements of F_{i-1} and $\boldsymbol{\beta}$ is a parameter vector. Conditioned on $A_i = 1$, D_i is also a binary variable, and we use the following model for $\delta_i = P(D_i = 1|A_i = 1)$:

$$\ln\left(\frac{\delta_i}{1-\delta_i}\right) = \boldsymbol{z}_i\boldsymbol{\gamma} \quad \text{or} \quad \delta_i = \frac{e^{\boldsymbol{z}_i\boldsymbol{\gamma}}}{1+e^{\boldsymbol{z}_i\boldsymbol{\gamma}}}, \tag{5.26}$$

where $\boldsymbol{z}_i$ is a finite-dimensional vector consisting of elements of F_{i-1} and $\boldsymbol{\gamma}$ is a parameter vector. To allow for asymmetry between positive and negative price changes, we assume that

$$S_i|(D_i, A_i = 1) \sim 1 + \begin{cases} g(\lambda_{u,i}) & \text{if } D_i = 1, A_i = 1, \\ g(\lambda_{d,i}) & \text{if } D_i = -1, A_i = 1, \end{cases} \tag{5.27}$$

where $g(\lambda)$ is a geometric distribution with parameter λ and the parameters $\lambda_{j,i}$ evolve over time as

$$\ln\left(\frac{\lambda_{j,i}}{1-\lambda_{j,i}}\right) = \boldsymbol{w}_i\boldsymbol{\theta}_j \quad \text{or} \quad \lambda_{j,i} = \frac{e^{\boldsymbol{w}_i\boldsymbol{\theta}_j}}{1+e^{\boldsymbol{w}_i\boldsymbol{\theta}_j}}, \qquad j = u, d, \tag{5.28}$$

where $\boldsymbol{w}_i$ is again a finite-dimensional explanatory variable in F_{i-1} and $\boldsymbol{\theta}_j$ is a parameter vector.

In Eq. (5.27), the probability mass function of a random variable x, which follows the geometric distribution $g(\lambda)$, is

$$p(x = m) = \lambda(1-\lambda)^m, \qquad m = 0, 1, 2, \ldots.$$

We added 1 to the geometric distribution so that the price change, if it occurs, is at least 1 tick. In Eq. (5.28), we take the logistic transformation to ensure that $\lambda_{j,i} \in [0, 1]$.

The previous specification classifies the ith trade, or transaction, into one of three categories:

1. No price change: $A_i = 0$ and the associated probability is $(1 - p_i)$.
2. A price increase: $A_i = 1$, $D_i = 1$, and the associated probability is $p_i\delta_i$. The size of the price increase is governed by $1 + g(\lambda_{u,i})$.
3. A price drop: $A_i = 1$, $D_i = -1$, and the associated probability is $p_i(1 - \delta_i)$. The size of the price drop is governed by $1 + g(\lambda_{d,i})$.

Let $I_i(j)$ for $j = 1, 2, 3$ be the indicator variables of the prior three categories. That is, $I_i(j) = 1$ if the jth category occurs and $I_i(j) = 0$ otherwise. The log-likelihood function of Eq. (5.24) becomes

$$\begin{aligned}\ln[P(y_i|F_{i-1})] = {} & I_i(1)\ln[(1-p_i)] + I_i(2)[\ln(p_i) + \ln(\delta_i) \\ & + \ln(\lambda_{u,i}) + (S_i - 1)\ln(1-\lambda_{u,i})] \\ & + I_i(3)[\ln(p_i) + \ln(1-\delta_i) + \ln(\lambda_{d,i}) + (S_i - 1)\ln(1-\lambda_{d,i})],\end{aligned}$$

and the overall log-likelihood function is

$$\ln[P(y_1, \ldots, y_n|F_0)] = \sum_{i=1}^{n} \ln[P(y_i|F_{i-1})], \tag{5.29}$$

which is a function of parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\theta}_u$, and $\boldsymbol{\theta}_d$.

Example 5.2. We illustrate the decomposition model by analyzing the intraday transactions of IBM stock from November 1, 1990, to January 31, 1991. There were 63 trading days and 59,838 intraday transactions in the normal trading hours. The explanatory variables used are:

1. A_{i-1}: the action indicator of the previous trade [i.e., the $(i-1)$th trade within a trading day]
2. D_{i-1}: the direction indicator of the previous trade
3. S_{i-1}: the size of the previous trade
4. V_{i-1}: the volume of the previous trade, divided by 1000
5. Δt_{i-1}: time duration from the $(i-2)$th to $(i-1)$th trade
6. BA_i: the bid–ask spread prevailing at the time of transaction

Because we use lag-1 explanatory variables, the actual sample size is 59,775. It turns out that V_{i-1}, Δt_{i-1}, and BA_i are not statistically significant for the model entertained. Thus, only the first three explanatory variables are used. The model employed is

$$\begin{aligned}\ln\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 A_{i-1}, \\ \ln\left(\frac{\delta_i}{1-\delta_i}\right) &= \gamma_0 + \gamma_1 D_{i-1}, \\ \ln\left(\frac{\lambda_{u,i}}{1-\lambda_{u,i}}\right) &= \theta_{u,0} + \theta_{u,1} S_{i-1}, \\ \ln\left(\frac{\lambda_{d,i}}{1-\lambda_{d,i}}\right) &= \theta_{d,0} + \theta_{d,1} S_{i-1}.\end{aligned} \tag{5.30}$$

TABLE 5.6 Parameter Estimates of ADS Model in Eq. (5.30) for IBM Intraday Transactions from December 1, 1990, to January 31, 1991

Parameter	β_0	β_1	γ_0	γ_1
Estimate	−1.057	0.962	−0.067	−2.307
Standard Error	0.104	0.044	0.023	0.056
Parameter	$\theta_{u,0}$	$\theta_{u,1}$	$\theta_{d,0}$	$\theta_{d,1}$
Estimate	2.235	−0.670	2.085	−0.509
Standard Error	0.029	0.050	0.187	0.139

The parameter estimates, using the log-likelihood function in Eq. (5.29), are given in Table 5.6. The estimated simple model shows some dynamic dependence in the price change. In particular, the trade-by-trade price changes of IBM stock exhibit some appealing features:

1. The probability of a price change depends on the previous price change. Specifically, we have

$$P(A_i = 1|A_{i-1} = 0) = 0.258, \qquad P(A_i = 1|A_{i-1} = 1) = 0.476.$$

   The result indicates that a price change may occur in clusters and, as expected, most transactions are without price change. When no price change occurred at the $(i-1)$th trade, then only about one out of four trades in the subsequent transaction has a price change. When there is a price change at the $(i-1)$th transaction, the probability of a price change in the ith trade increases to about 0.5.
2. The direction of price change is governed by

$$P(D_i = 1|F_{i-1}, A_i) = \begin{cases} 0.483 & \text{if } D_{i-1} = 0 \text{ (i.e., } A_{i-1} = 0), \\ 0.085 & \text{if } D_{i-1} = 1, A_i = 1, \\ 0.904 & \text{if } D_{i-1} = -1, A_i = 1. \end{cases}$$

   This result says that (a) if no price change occurred at the $(i-1)$th trade, then the chances for a price increase or decrease at the ith trade are about even; and (b) the probabilities of consecutive price increases or decreases are very low. The probability of a price increase at the ith trade given that a price change occurs at the ith trade and there was a price increase at the $(i-1)$th trade is only 8.6%. However, the probability of a price increase is about 90% given that a price change occurs at the ith trade and there was a price decrease at the $(i-1)$th trade. Consequently, this result shows the effect of bid–ask bounce and supports price reversals in high-frequency trading.
3. There is weak evidence suggesting that big price changes have a higher probability to be followed by another big price change. Consider the size of

a price increase. We have

$$S_i|(D_i = 1) \sim 1 + g(\lambda_{u,i}), \qquad \lambda_{u,i} = 2.235 - 0.670S_{i-1}.$$

Using the probability mass function of a geometric distribution, we obtain that the probability of a price increase by one tick is 0.827 at the ith trade if the transaction results in a price increase and $S_{i-1} = 1$. The probability reduces to 0.709 if $S_{i-1} = 2$ and to 0.556 if $S_{i-1} = 3$. Consequently, the probability of a large S_i is proportional to S_{i-1} given that there is a price increase at the ith trade.

A difference between the ADS of Eq. (5.21) and ordered probit models is that the former does not require any truncation or grouping in the size of a price change.

R Demonstration for Logistic Linear Regression

The following output has been edited:

```
> da=read.table("ibm91-ads.txt",header=T)
> da1=read.table("ibm91-adsx.txt",header=T)
> Ai=da[,1]  % Select the variables
> Di=da[,2]
> Aim1=da1[,4]
> Dim1=da1[,5]
>
> m1=glm(Ai~Aim1,family=binomial) %Fit a linear
   logistic model
> summary(m1)
Call:
glm(formula = Ai ~ Aim1, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1373  -0.7724  -0.7724   1.2180   1.6462

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.05667    0.01142  -92.55   <2e-16 ***
Aim1         0.96164    0.01827   52.62   <2e-16 ***
---
>
> di=Di[Ai==1]   % Select the cases in which Ai = 1.
> dim1=Dim1[Ai==1]
> di=(di+abs(di))/2  % Logistic regression works for 1 or 0,
       % but di is coded 1 or -1 so that change is needed.
> m2=glm(di~dim1,family=binomial)
> summary(m2)
Call:
glm(formula = di ~ dim1, family = binomial)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1640  -1.1493   0.4497   1.2058   2.2193

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.06663    0.01728  -3.855 0.000116 ***
dim1        -2.30693    0.03595 -64.171 <  2e-16 ***
```

5.5 DURATION MODELS

Duration models are concerned with time intervals between trades. Longer durations indicate lack of trading activities, which in turn signify a period of no new information. The dynamic behavior of durations thus contains useful information about intraday market activities. Using concepts similar to the ARCH models for volatility, Engle and Russell (1998) propose an autoregressive conditional duration (ACD) model to describe the evolution of time durations for (heavily traded) stocks. Zhang et al. (2001) extend the ACD model to account for nonlinearity and structural breaks in the data. In this section, we introduce some simple duration models. As mentioned before, intraday transactions exhibit some diurnal pattern. Therefore, we focus on the adjusted time duration

$$\Delta t_i^* = \Delta t_i / f(t_i), \tag{5.31}$$

where $f(t_i)$ is a deterministic function consisting of the cyclical component of Δt_i. Obviously, $f(t_i)$ depends on the underlying asset and the systematic behavior of the market. In practice, there are many ways to estimate $f(t_i)$, but no single method dominates the others in terms of statistical properties. A common approach is to use smoothing spline. Here we use simple quadratic functions and indicator variables to take care of the deterministic component of daily trading activities.

For the IBM data employed in the illustration of ADS models, we assume

$$f(t_i) = \exp[d(t_i)], \qquad d(t_i) = \beta_0 + \sum_{j=1}^{7} \beta_j f_j(t_i), \tag{5.32}$$

where

$$f_1(t_i) = -\left(\frac{t_i - 43200}{14400}\right)^2, \qquad f_3(t_i) = \begin{cases} -\left(\dfrac{t_i - 38700}{7500}\right)^2 & \text{if } t_i < 43200, \\ 0 & \text{otherwise,} \end{cases}$$

$$f_2(t_i) = -\left(\frac{t_i - 48300}{9300}\right)^2, \qquad f_4(t_i) = \begin{cases} -\left(\dfrac{t_i - 48600}{9000}\right)^2 & \text{if } t_i \geq 43200, \\ 0 & \text{otherwise,} \end{cases}$$

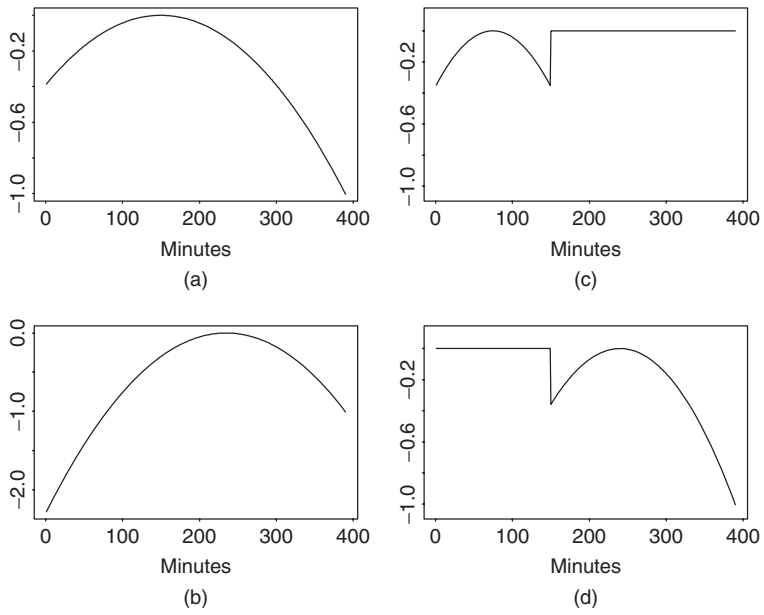

Figure 5.7 Quadratic functions used to remove deterministic component of IBM intraday trading durations: (a)–(d) are functions $f_1(\cdot)$ to $f_4(\cdot)$ of Eq. (5.32), respectively.

and $f_5(t_i)$ and $f_6(t_i)$ are indicator variables for the first and second 5 minutes of market opening [i.e., $f_5(\cdot) = 1$ if and only if t_i is between 9:30 AM and 9:35 AM Eastern time], and $f_7(t_i)$ is the indicator for the last 30 minutes of daily trading [i.e., $f_7(t_i) = 1$ if and only if the trade occurred between 3:30 PM and 4:00 PM Eastern time]. Figure 5.7 shows the plot of $f_i(\cdot)$ for $i = 1, \ldots, 4$, where the time scale on the x axis is in minutes. Note that $f_3(43200) = f_4(43200)$, where 43,200 corresponds to 12:00 noon.

The coefficients β_j of Eq. (5.32) are obtained by the least-squares method of the linear regression

$$\ln(\Delta t_i) = \beta_0 + \sum_{j=1}^{7} \beta_j f_j(t_i) + \epsilon_i.$$

The fitted model is

$$\begin{aligned}\widehat{\ln(\Delta t_i)} = {} & 2.555 + 0.159 f_1(t_i) + 0.270 f_2(t_i) + 0.384 f_3(t_i) \\ & + 0.061 f_4(t_i) - 0.611 f_5(t_i) - 0.157 f_6(t_i) + 0.073 f_7(t_i).\end{aligned}$$

Figure 5.8 shows the time plot of average durations in 5-minute time intervals over the 63 trading days before and after adjusting for the deterministic component. Figure 5.8(a) shows the average durations of Δt_i and, as expected, exhibits a

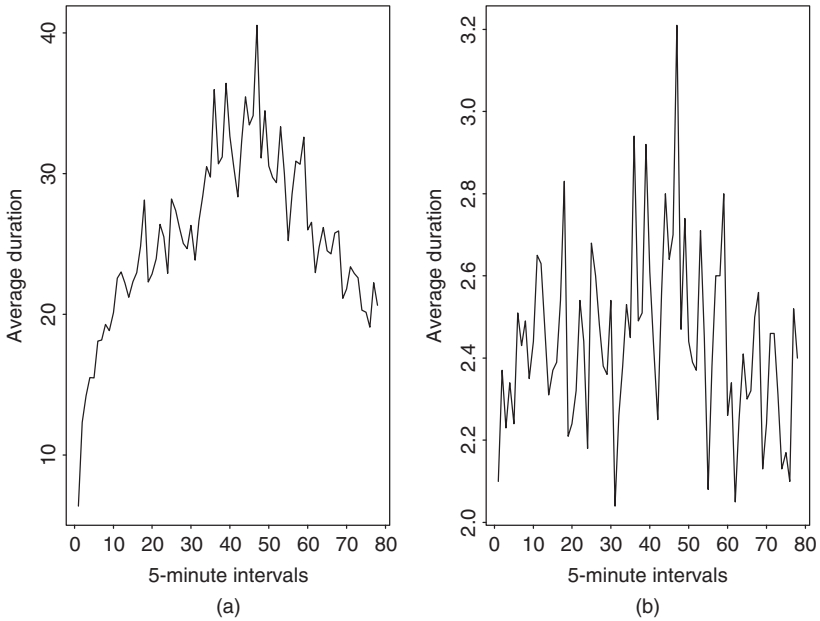

Figure 5.8 IBM transactions data from 11/01/90 to 1/31/91: (a) average durations in 5-minute time intervals and (b) average durations in 5-minute time intervals after adjusting for deterministic component.

diurnal pattern. Figure 5.8(b) shows the average durations of Δt_i^* (i.e., after the adjustment), and the diurnal pattern is largely removed.

5.5.1 The ACD Model

The autoregressive conditional duration (ACD) model uses the idea of GARCH models to study the dynamic structure of the adjusted duration Δt_i^* of Eq. (5.31). For ease in notation, we define $x_i = \Delta t_i^*$.

Let $\psi_i = E(x_i|F_{i-1})$ be the conditional expectation of the adjusted duration between the $(i-1)$th and ith trades, where F_{i-1} is the information set available at the $(i-1)$th trade. In other words, ψ_i is the expected adjusted duration given F_{i-1}. The basic ACD model is defined as

$$x_i = \psi_i \epsilon_i, \tag{5.33}$$

where $\{\epsilon_i\}$ is a sequence of independent and identically distributed nonnegative random variables such that $E(\epsilon_i) = 1$. In Engle and Russell (1998), ϵ_i follows a standard exponential or a standardized Weibull distribution, and ψ_i assumes the form

$$\psi_i = \omega + \sum_{j=1}^{r} \gamma_j x_{i-j} + \sum_{j=1}^{s} \omega_j \psi_{i-j}. \tag{5.34}$$

Such a model is referred to as an ACD(r, s) model. When the distribution of ϵ_i is exponential, the resulting model is called an EACD(r, s) model. Similarly, if ϵ_i follows a Weibull distribution, the model is a WACD(r, s) model. If necessary, readers are referred to Appendix A for a quick review of exponential and Weibull distributions.

Similar to GARCH models, the process $\eta_i = x_i - \psi_i$ is a martingale difference sequence [i.e., $E(\eta_i|F_{i-1}) = 0$], and the ACD(r, s) model can be written as

$$x_i = \omega + \sum_{j=1}^{\max(r,s)} (\gamma_j + \omega_j)x_{i-j} - \sum_{j=1}^{s} \omega_j \eta_{i-j} + \eta_j, \tag{5.35}$$

which is in the form of an ARMA process with non-Gaussian innovations. It is understood here that $\gamma_j = 0$ for $j > r$ and $\omega_j = 0$ for $j > s$. Such a representation can be used to obtain the basic conditions for weak stationarity of the ACD model. For instance, taking expectation on both sides of Eq. (5.35) and assuming weak stationarity, we have

$$E(x_i) = \frac{\omega}{1 - \sum_{j=1}^{\max(r,s)} (\gamma_j + \omega_j)}.$$

Therefore, we assume $\omega > 0$ and $1 > \sum_j (\gamma_j + \omega_j)$ because the expected duration is positive. As another application of Eq. (5.35), we study properties of the EACD(1,1) model.

EACD(1,1) Model

An EACD(1,1) model can be written as

$$x_i = \psi_i \epsilon_i, \qquad \psi_i = \omega + \gamma_1 x_{i-1} + \omega_1 \psi_{i-1}, \tag{5.36}$$

where ϵ_i follows the standard exponential distribution. Using the moments of a standard exponential distribution in Appendix A, we have $E(\epsilon_i) = 1$, $\text{Var}(\epsilon_i) = 1$, and $E(\epsilon_i^2) = \text{Var}(x_i) + [E(x_i)]^2 = 2$. Assuming that x_i is weakly stationary (i.e., the first two moments of x_i are time invariant), we derive the variance of x_i. First, taking the expectation of Eq. (5.36), we have

$$\begin{aligned} E(x_i) &= E[E(\psi_i \epsilon_i|F_{i-1})] = E(\psi_i), \\ E(\psi_i) &= \omega + \gamma_1 E(x_{i-1}) + \omega_1 E(\psi_{i-1}). \end{aligned} \tag{5.37}$$

Under weak stationarity, $E(\psi_i) = E(\psi_{i-1})$ so that Eq. (5.37) gives

$$\mu_x \equiv E(x_i) = E(\psi_i) = \frac{\omega}{1 - \gamma_1 - \omega_1}. \tag{5.38}$$

Next, because $E(\epsilon_i^2) = 2$, we have $E(x_i^2) = E[E(\psi_i^2 \epsilon_i^2|F_{i-1})] = 2E(\psi_i^2)$.

Taking the square of ψ_i in Eq. (5.36) and the expectation and using weak stationarity of ψ_i and x_i, we have, after some algebra, that

$$E(\psi_i^2) = \mu_x^2 \times \frac{1 - (\gamma_1 + \omega_1)^2}{1 - 2\gamma_1^2 - \omega_1^2 - 2\gamma_1\omega_1}. \tag{5.39}$$

Finally, using $\text{Var}(x_i) = E(x_i^2) - [E(x_i)]^2$ and $E(x_i^2) = 2E(\psi_i^2)$, we have

$$\text{Var}(x_i) = 2E(\psi_i^2) - \mu_x^2 = \mu_x^2 \times \frac{1 - \omega_1^2 - 2\gamma_1\omega_1}{1 - \omega_1^2 - 2\gamma_1\omega_1 - 2\gamma_1^2},$$

where μ_x is defined in Eq. (5.38). This result shows that, to have time-invariant unconditional variance, the EACD(1,1) model in Eq. (5.36) must satisfy $1 > 2\gamma_1^2 + \omega_1^2 + 2\gamma_1\omega_1$. The variance of a WACD(1,1) model can be obtained by using the same techniques and the first two moments of a standardized Weibull distribution.

ACD Models with a Generalized Gamma Distribution

In the statistical literature, intensity function is often expressed in terms of hazard function. As shown in Appendix B, the hazard function of an EACD model is constant over time and that of a WACD model is a monotonous function. These hazard functions are rather restrictive in application as the intensity function of stock transactions might not be constant or monotone over time. To increase the flexibility of the associated hazard function, Zhang et al. (2001) employ a (standardized) generalized gamma distribution for ϵ_i. See Appendix A for some basic properties of a generalized gamma distribution. The resulting hazard function may assume various patterns, including U shape or inverted U shape. We refer to an ACD model with innovations that follow a generalized gamma distribution as a GACD(r, s) model.

5.5.2 Simulation

To illustrate ACD processes, we generated 500 observations from the ACD(1,1) model:

$$x_i = \psi_i\epsilon_i, \qquad \psi_i = 0.3 + 0.2x_{i-1} + 0.7\psi_{i-1} \tag{5.40}$$

using two different innovational distributions for ϵ_i. In case 1, ϵ_i is assumed to follow a standardized Weibull distribution with parameter $\alpha = 1.5$. In case 2, ϵ_i follows a (standardized) generalized gamma distribution with parameters $\kappa = 1.5$ and $\alpha = 0.5$.

Figure 5.9(a) shows the time plot of the WACD(1,1) series, whereas Figure 5.10(a) is the GACD(1,1) series. Figure 5.11 plots the histograms of both simulated series. The difference between the two models is evident. Finally, the sample ACFs of the two simulated series are shown in Figures 5.12(a) and 5.13(b), respectively. The serial dependence of the data is clearly seen.

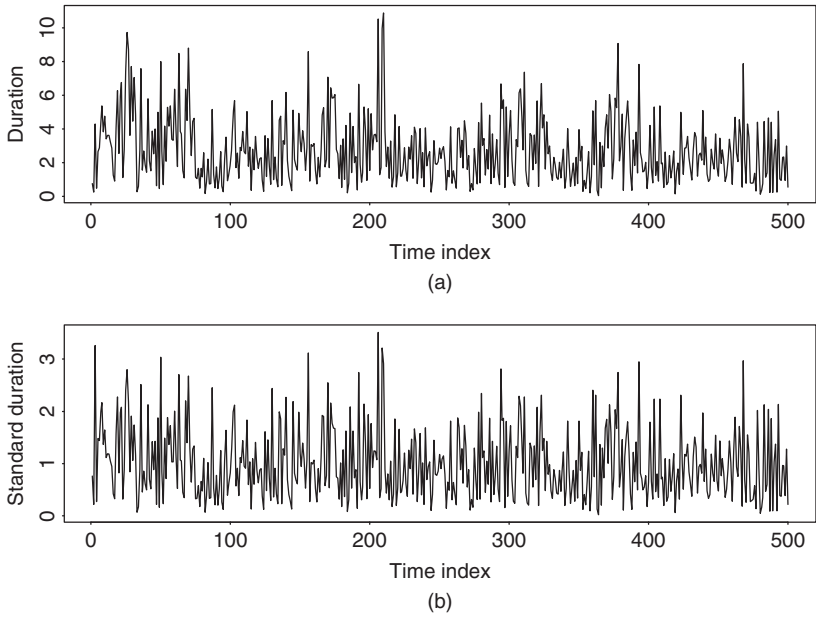

Figure 5.9 Simulated WACD(1,1) series in Eq. (5.40): (a) original series and (b) standardized series after estimation. There are 500 observations.

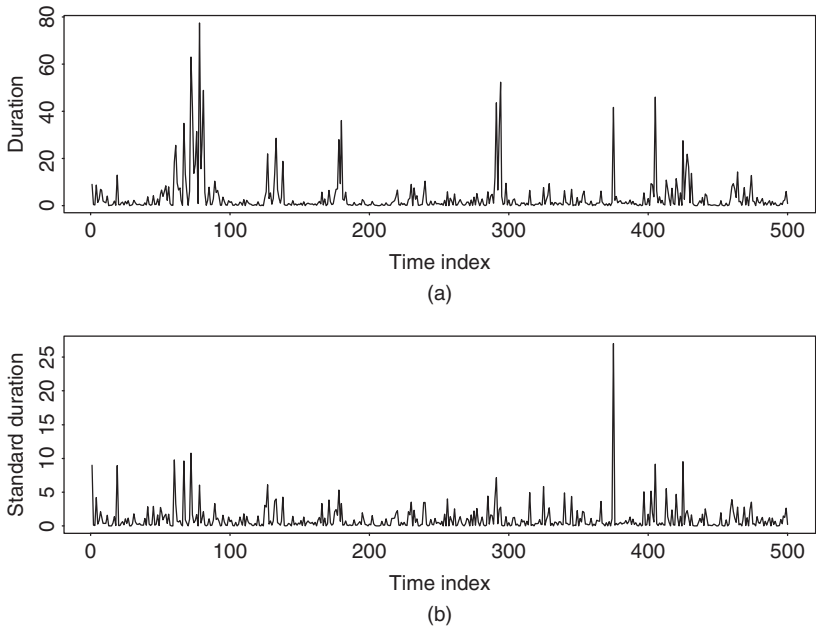

Figure 5.10 Simulated GACD(1,1) series in Eq. (5.40): (a) original series and (b) standardized series after estimation. There are 500 observations.

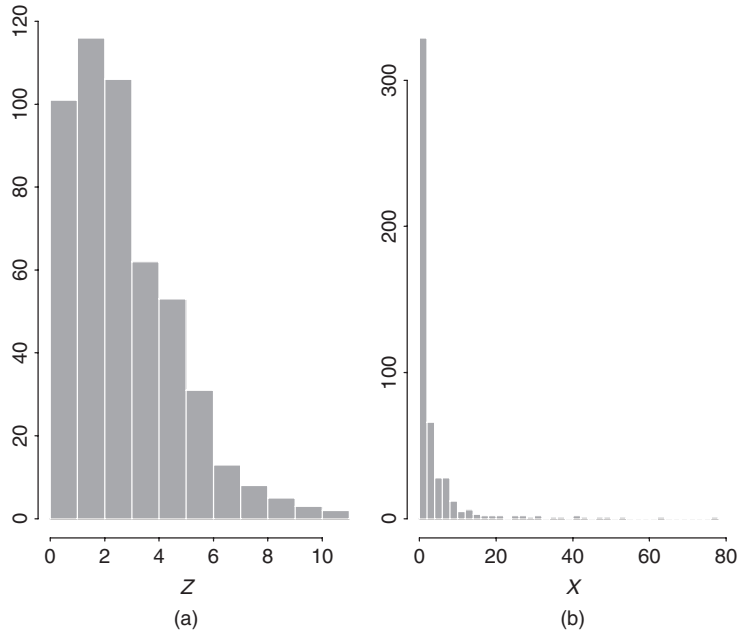

Figure 5.11 Histograms of simulated duration processes with 500 observations: (a) WACD(1,1) model and (b) GACD(1,1) model.

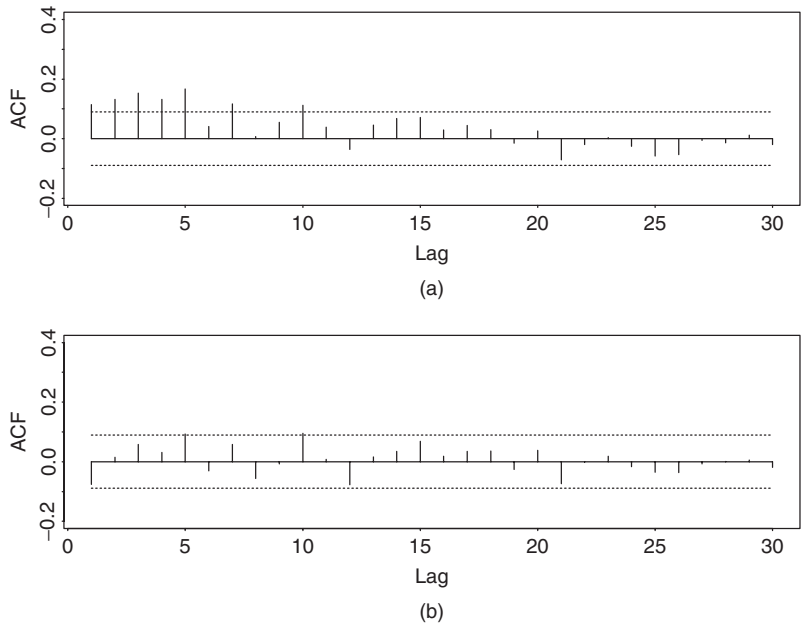

Figure 5.12 Sample autocorrelation function of simulated WACD(1,1) series with 500 observations: (a) original series and (b) standardized residual series.

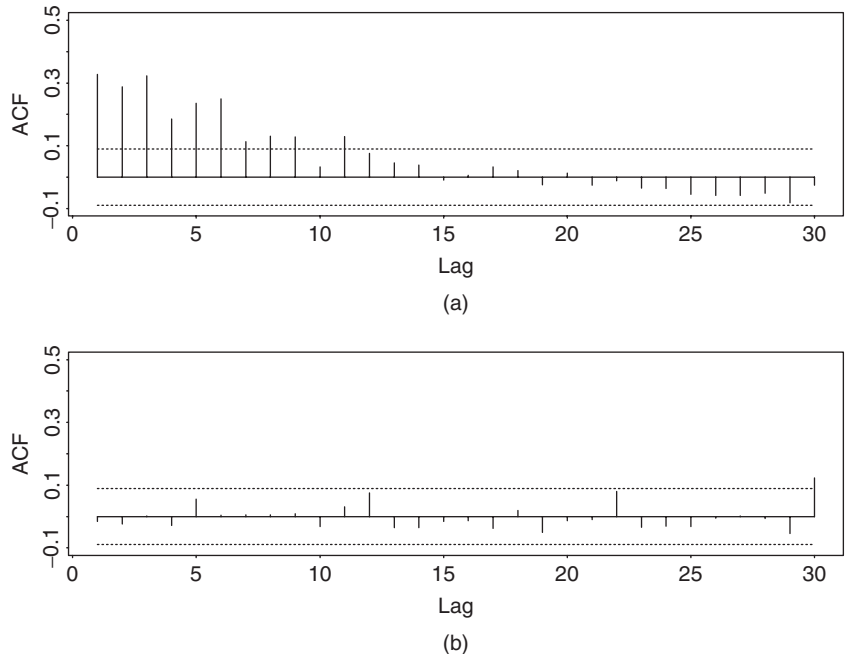

Figure 5.13 Sample autocorrelation function of simulated GACD(1,1) series with 500 observations: (a) original series and (b) standardized residual series.

5.5.3 Estimation

For an ACD(r, s) model, let $i_o = \max(r, s)$ and $\boldsymbol{x}_t = (x_1, \ldots, x_t)'$. The likelihood function of the durations $x_1, \ldots, x_T$ is

$$f(\boldsymbol{x}_T|\boldsymbol{\theta}) = \left[\prod_{i=i_o+1}^{T} f(x_i|F_{i-1}, \boldsymbol{\theta})\right] \times f(\boldsymbol{x}_{i_o}|\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ denotes the vector of model parameters, and T is the sample size. The marginal probability density function $f(\boldsymbol{x}_{i_o}|\boldsymbol{\theta})$ of the previous equation is rather complicated for a general ACD model. Because its impact on the likelihood function is diminishing as the sample size T increases, this marginal density is often ignored, resulting in use of the conditional-likelihood method. For a WACD model, we use the probability density function (pdf) of Eq. (5.56) and obtain the conditional log-likelihood function

$$\begin{aligned}\ell(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{x}_{i_o}) = &\sum_{i=i_0+1}^{T} \alpha \ln\left[\Gamma\left(1+\frac{1}{\alpha}\right)\right] \\ &+ \ln\left(\frac{\alpha}{x_i}\right) + \alpha \ln\left(\frac{x_i}{\psi_i}\right) - \left[\frac{\Gamma(1+1/\alpha)x_i}{\psi_i}\right]^{\alpha}, \end{aligned} \tag{5.41}$$

TABLE 5.7 Estimation Results for Simulated ACD(1,1) Series with 500 Observations: For WACD(1,1) Series and GACD(1,1) Series

	WACD(1,1) Model				
Parameter	ω	γ_1	ω_1	α	
True	0.3	0.2	0.7	1.5	
Estimate	0.364	0.100	0.767	1.477	
Standard Error	(0.139)	(0.025)	(0.060)	(0.052)	
	GACD(1,1) Model				
Parameter	ω	γ_1	ω_1	α	κ
True	0.3	0.2	0.7	0.5	1.5
Estimate	0.401	0.343	0.561	0.436	2.077
Standard Error	(0.117)	(0.074)	(0.065)	(0.078)	(0.653)

where $\psi_i = \omega + \sum_{j=1}^{r} \gamma_j x_{i-j} + \sum_{j=1}^{s} \omega_j \psi_{i-j}$, $\boldsymbol{\theta} = (\omega, \gamma_1, \ldots, \gamma_r, \omega_1, \ldots, \omega_s, \alpha)'$, and $\boldsymbol{x} = (x_{i_o+1}, \ldots, x_T)'$. When $\alpha = 1$, the (conditional) log-likelihood function reduces to that of an EACD(r, s) model.

For a GACD(r, s) model, the conditional log-likelihood function is

$$\ell(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{x}_{i_o}) = \sum_{i=i_o+1}^{T} \ln\left[\frac{\alpha}{\Gamma(\kappa)}\right] + (\kappa\alpha - 1)\ln(x_i) - \kappa\alpha \ln(\lambda\psi_i) - \left(\frac{x_i}{\lambda\psi_i}\right)^{\alpha}, \tag{5.42}$$

where $\lambda = \Gamma(\kappa)/\Gamma(\kappa + 1/\alpha)$ and the parameter vector $\boldsymbol{\theta}$ now also includes κ. As expected, when $\kappa = 1$, $\lambda = 1/\Gamma(1 + 1/\alpha)$ and the log-likelihood function in Eq. (5.42) reduces to that of a WACD(r, s) model in Eq. (5.41). This log-likelihood function can be rewritten in many ways to simplify the estimation.

Under some regularity conditions, the conditional maximum-likelihood estimates are asymptotically normal; see Engle and Russell (1998) and the references therein. In practice, simulation can be used to obtain finite-sample reference distributions for the problem of interest once a duration model is specified.

Example 5.3. (Simulated ACD(1,1) series, continued). Consider the simulated WACD(1,1) and GACD(1,1) series of Eq. (5.40). We apply the conditional-likelihood method and obtain the results in Table 5.7. The estimates appear to be reasonable. Let $\hat{\psi}_i$ be the 1-step-ahead prediction of ψ_i and $\hat{\epsilon}_i = x_i/\hat{\psi}_i$ be the standardized series, which can be regarded as standardized residuals of the series. If the model is adequately specified, $\{\hat{\epsilon}_i\}$ should behave as a sequence of independent and identically distributed random variables. Figures 5.9(b) and 5.10(b) show the time plot of $\hat{\epsilon}_i$ for both models. The sample ACF of $\hat{\epsilon}_i$ for both fitted models are shown in Figures 5.12(b) and 5.13(b), respectively. It is evident that no significant serial correlations are found in the $\hat{\epsilon}_i$ series.

Example 5.4. As an illustration of duration models, we consider the transaction durations of IBM stock on five consecutive trading days from November 1

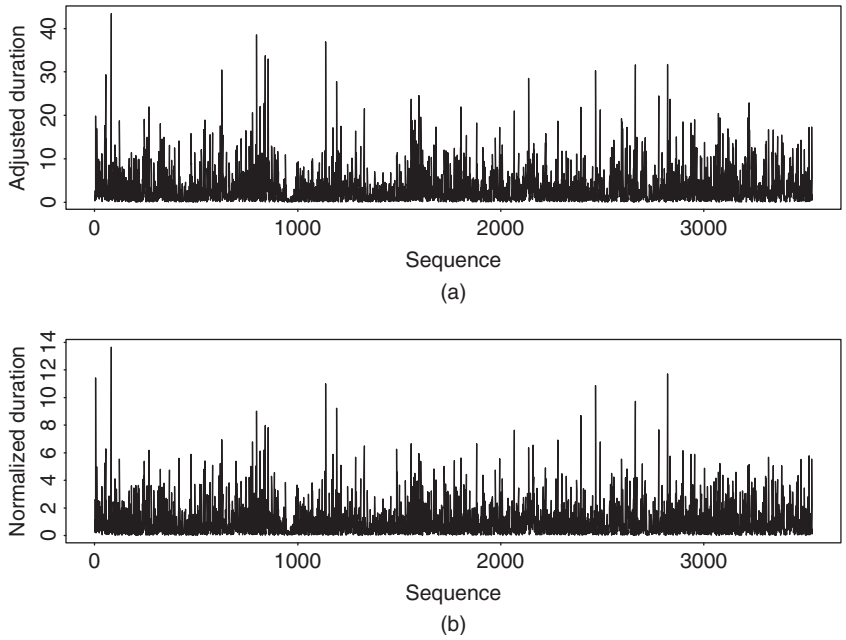

Figure 5.14 Time plots of durations for IBM stock traded in first five trading days of November 1990: (a) adjusted series and (b) normalized innovations of an WACD(1,1) model. There are 3534 nonzero durations.

to November 7, 1990. Focusing on positive transaction durations, we have 3534 observations. In addition, the data have been adjusted by removing the deterministic component in Eq. (5.32). That is, we employ 3534 positive adjusted durations as defined in Eq. (5.31).

Figure 5.14(a) shows the time plot of the adjusted (positive) durations for the first five trading days of November 1990, and Figure 5.15(a) gives the sample ACF of the series. There exist some serial correlations in the adjusted durations. We fit a WACD(1,1) model to the data and obtain the model

$$x_i = \psi_i \epsilon_i, \qquad \psi_i = 0.169 + 0.064 x_{i-1} + 0.885 \psi_{i-1}, \tag{5.43}$$

where $\{\epsilon_i\}$ is a sequence of independent and identically distributed random variates that follow the standardized Weibull distribution with parameter $\hat{\alpha} = 0.879(0.012)$, where 0.012 is the estimated standard error. Standard errors of the estimates in Eq. (5.43) are 0.039, 0.010, and 0.018, respectively. All t ratios of the estimates are greater than 4.2, indicating that the estimates are significant at the 1% level. Figure 5.14(b) shows the time plot of $\hat{\epsilon}_i = x_i/\hat{\psi}_i$, and Figure 5.15(b) provides the sample ACF of $\hat{\epsilon}_i$. The Ljung–Box statistics show $Q(10) = 4.96$ and $Q(20) = 10.75$ for the $\hat{\epsilon}_i$ series. Clearly, the standardized innovations have no significant serial correlations. In fact, the sample autocorrelations of the squared series $\{\hat{\epsilon}_i^2\}$

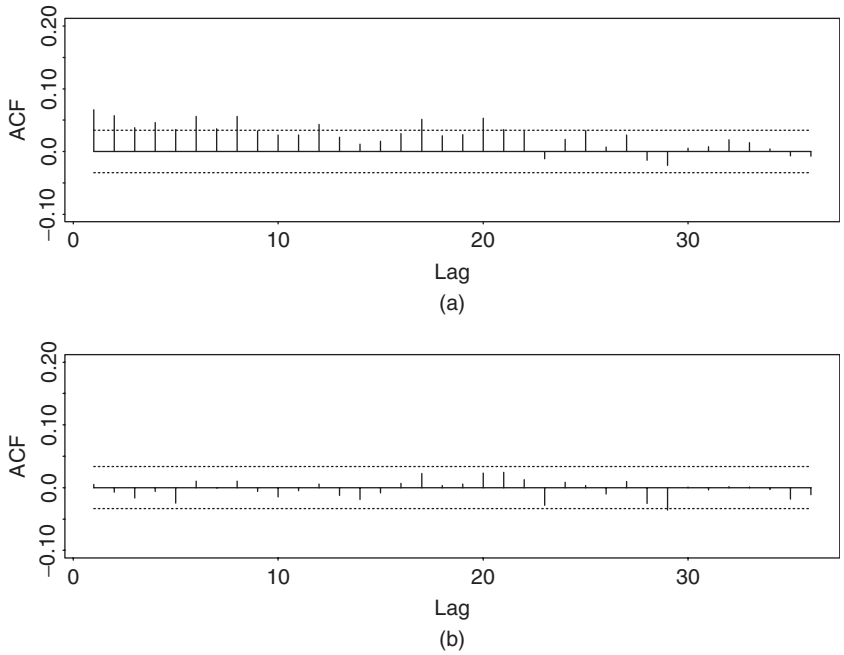

Figure 5.15 Sample autocorrelation function of adjusted durations for IBM stock traded in first five trading days of November 1990: (a) adjusted series and (b) normalized innovations for WACD(1,1) model.

are also small with $Q(10) = 6.20$ and $Q(20) = 11.16$, further confirming lack of serial dependence in the normalized innovations. In addition, the mean and standard deviation of a standardized Weibull distribution with $\alpha = 0.879$ are 1.00 and 1.14, respectively. These numbers are close to the sample mean and standard deviation of $\{\hat{\epsilon}_i\}$, which are 1.01 and 1.22, respectively. The fitted model seems adequate.

In model (5.43), the estimated coefficients show $\hat{\gamma}_1 + \hat{\omega}_1 \approx 0.949$, indicating certain persistence in the adjusted durations. The expected adjusted duration is $0.169/(1 - 0.064 - 0.885) = 3.31$ seconds, which is close to the sample mean 3.29 of the adjusted durations. The estimated α of the standardized Weibull distribution is 0.879, which is less than but close to 1. Thus, the conditional hazard function is monotonously decreasing at a slow rate.

If a generalized gamma distribution function is used for the innovations, then the fitted GACD(1,1) model is

$$x_i = \psi_i \epsilon_i, \qquad \psi_i = 0.141 + 0.063 x_{i-1} + 0.897 \psi_{i-1}, \tag{5.44}$$

where $\{\epsilon_i\}$ follows a standardized, generalized gamma distribution in Eq. (5.57) with parameters $\kappa = 4.248(1.046)$ and $\alpha = 0.395(0.053)$, where the number in parentheses denotes estimated standard error. Standard errors of the three parameters in Eq. (5.44) are 0.041, 0.010, and 0.019, respectively. All of the estimates are

statistically significant at the 1% level. Again, the normalized innovational process $\{\hat{\epsilon}_i\}$ and its squared series have no significant serial correlation, where $\hat{\epsilon}_i = x_i/\hat{\psi}_i$ based on model (5.44). Specifically, for the $\hat{\epsilon}_i$ process, we have $Q(10) = 4.95$ and $Q(20) = 10.28$. For the $\hat{\epsilon}_i^2$ series, we have $Q(10) = 6.36$ and $Q(20) = 10.89$.

The expected duration of model (5.44) is 3.52, which is slightly greater than that of the WACD(1,1) model in Eq. (5.43). Similarly, the persistence parameter $\hat{\gamma}_1 + \hat{\omega}_1$ of model (5.44) is also slightly higher at 0.96.

Remark. Estimation of EACD models can be carried out by using programs for ARCH models with some minor modification; see Engle and Russell (1998). In this book, we use either the RATS program or some Fortran programs developed by the author to estimate the duration models. Limited experience indicates that it is harder to estimate a GACD model than an EACD or a WACD model. RATS programs used to estimate WACD and GACD models are given in Appendix C. □

5.6 NONLINEAR DURATION MODELS

Nonlinear features are also commonly found in high-frequency data. As an illustration, we apply some nonlinearity tests discussed in Chapter 4 to the normalized innovations $\hat{\epsilon}_i$ of the WACD(1,1) model for the IBM transaction durations in Example 5.4; see Eq. (5.43). Based on an AR(4) model, the test results are given in part (a) of Table 5.8. As expected from the model diagnostics of Example 5.4, the Ori-F test indicates no quadratic nonlinearity in the normalized innovations. However, the TAR-F test statistics suggest strong nonlinearity.

Based on the test results in Table 5.8, we entertain a threshold duration model with two regimes for the IBM intraday durations. The threshold variable is x_{t-1} (i.e., lag-1 adjusted duration). The estimated threshold value is 3.79. The fitted threshold WACD(1,1) model is $x_i = \psi_i \epsilon_i$, where

$$\psi_i = \begin{cases} 0.020 + 0.257x_{i-1} + 0.847\psi_{i-1}, & \epsilon_i \sim w(0.901) \quad \text{if } x_{i-1} \le 3.79, \\ 1.808 + 0.027x_{i-1} + 0.501\psi_{i-1}, & \epsilon_i \sim w(0.845) \quad \text{if } x_{i-1} > 3.79, \end{cases} \tag{5.45}$$

where $w(\alpha)$ denotes a standardized Weibull distribution with parameter α. The number of observations in the two regimes are 2503 and 1030, respectively. In Eq. (5.45), the standard errors of the parameters for the first regime are 0.043, 0.041, 0.024, and 0.014, whereas those for the second regime are 0.526, 0.020, 0.147, and 0.020, respectively.

Consider the normalized innovations $\hat{\epsilon}_i = x_i/\hat{\psi}_i$ of the threshold WACD(1,1) model in Eq. (5.45). We obtain $Q(12) = 9.8$ and $Q(24) = 23.9$ for $\hat{\epsilon}_i$ and $Q(12) = 8.0$ and $Q(24) = 16.7$ for $\hat{\epsilon}_i^2$. Thus, there are no significant serial correlations in the $\hat{\epsilon}_i$ and $\hat{\epsilon}_i^2$ series. Furthermore, applying the same nonlinearity tests as before to this newly normalized innovational series $\hat{\epsilon}_i$, we detect no nonlinearity; see part

TABLE 5.8 Nonlinearity Tests for IBM Transaction Durations from November 1 to November 7, 1990[a]

	(a) Normalized Innovations of a WACD(1,1) Model				
Type	Ori-F	TAR-$F(1)$	TAR-$F(2)$	TAR-$F(3)$	TAR-$F(4)$
Test	0.343	3.288	3.142	3.128	0.297
p Value	0.969	0.006	0.008	0.008	0.915
	(b) Normalized Innovations of a Threshold WACD(1,1) Model				
Test	0.163	0.746	1.899	1.752	0.270
p Value	0.998	0.589	0.091	0.119	0.929

[a]Only intraday durations are used. The number in parentheses of TAR-F tests denotes time delay.

(b) of Table 5.8. Consequently, the two-regime threshold WACD(1,1) model in Eq. (5.45) is adequate.

If we classify the two regimes as heavy and thin trading periods, then the threshold model suggests that the trading dynamics measured by intraday transaction durations are different between heavy and thin trading periods for IBM stock even after the adjustment of diurnal pattern. This is not surprising as market activities are often driven by the arrival of news and other information.

The estimated threshold WACD(1,1) model in Eq. (5.45) contains some insignificant parameters. We refine the model and obtain the result:

$$\psi_i = \begin{cases} 0.225x_{i-1} + 0.867\psi_{i-1}, & \epsilon_i \sim w(0.902) \quad \text{if } x_{i-1} \le 3.79, \\ 1.618 + 0.614\psi_{i-1}, & \epsilon_i \sim w(0.846) \quad \text{if } x_{i-1} > 3.79. \end{cases}$$

All of the estimates of the refined model are highly significant. The Ljung–Box statistics of the standardized innovations $\hat{\epsilon}_i = x_i/\hat{\psi}_i$ show $Q(10) = 5.91(0.82)$ and $Q(20) = 16.04(0.71)$ and those of $\hat{\epsilon}_i^2$ give $Q(10) = 5.35(0.87)$ and $Q(20) = 15.20(0.76)$, where the number in parentheses is the p value. Therefore, the refined model is adequate. The RATS program used to estimate the prior model is given in Appendix C.

5.7 BIVARIATE MODELS FOR PRICE CHANGE AND DURATION

In this section, we introduce a model that considers jointly the process of price change and the associated duration. As mentioned before, many intraday transactions of a stock result in no price change. Those transactions are highly relevant to trading intensity, but they do not contain direct information on price movement. Therefore, to simplify the complexity involved in modeling price change, we focus on transactions that result in a price change and consider a price change and duration (PCD) model to describe the multivariate dynamics of price change and the associated time duration.

We continue to use the same notation as before, but the definition is changed to transactions with a price change. Let t_i be the calendar time of the ith price change of an asset. As before, t_i is measured in seconds from midnight of a trading day. Let P_{t_i} be the transaction price when the ith price change occurred and $\Delta t_i = t_i - t_{i-1}$ be the time duration between price changes. In addition, let N_i be the number of trades in the time interval (t_{i-1}, t_i) that result in no price change. This new variable is used to represent trading intensity during a period of no price change. Finally, let D_i be the direction of the ith price change with $D_i = 1$ when price goes up and $D_i = -1$ when the price comes down, and let S_i be the size of the ith price change measured in ticks. Under the new definitions, the price of a stock evolves over time by

$$P_{t_i} = P_{t_{i-1}} + D_i S_i, \tag{5.46}$$

and the transactions data consist of $\{\Delta t_i, N_i, D_i, S_i\}$ for the ith price change. The PCD model is concerned with the joint analysis of $(\Delta t_i, N_i, D_i, S_i)$.

Remark. Focusing on transactions associated with a price change can reduce the sample size dramatically. For example, consider the intraday data of IBM stock from November 1, 1990 to January 31, 1991. There were 60,265 intraday trades, but only 19,022 of them resulted in a price change. In addition, there is no diurnal pattern in time durations between price changes. □

To illustrate the relationship among the price movements of all transactions and those of transactions associated with a price change, we consider the intraday tradings of IBM stock on November 21, 1990. There were 726 transactions on that day during normal trading hours, but only 195 trades resulted in a price change. Figure 5.16 shows the time plot of the price series for both cases. As expected, the price series are the same.

The PCD model decomposes the joint distribution of $(\Delta t_i, N_i, D_i, S_i)$ given F_{i-1} as

$$\begin{aligned} &f(\Delta t_i, N_i, D_i, S_i | F_{i-1}) \\ &= f(S_i | D_i, N_i, \Delta t_i, F_{i-1}) f(D_i | N_i, \Delta t_i, F_{i-1}) f(N_i | \Delta t_i, F_{i-1}) f(\Delta t_i | F_{i-1}). \end{aligned} \tag{5.47}$$

This partition enables us to specify suitable econometric models for the conditional distributions and, hence, to simplify the modeling task. There are many ways to specify models for the conditional distributions. A proper specification might depend on the asset under study. Here we employ the specifications used by McCulloch and Tsay (2000), who use generalized linear models for the discrete-valued variables and a time series model for the continuous variable $\ln(\Delta t_i)$.

For the time duration between price changes, we use the model

$$\ln(\Delta t_i) = \beta_0 + \beta_1 \ln(\Delta t_{i-1}) + \beta_2 S_{i-1} + \sigma \epsilon_i, \tag{5.48}$$

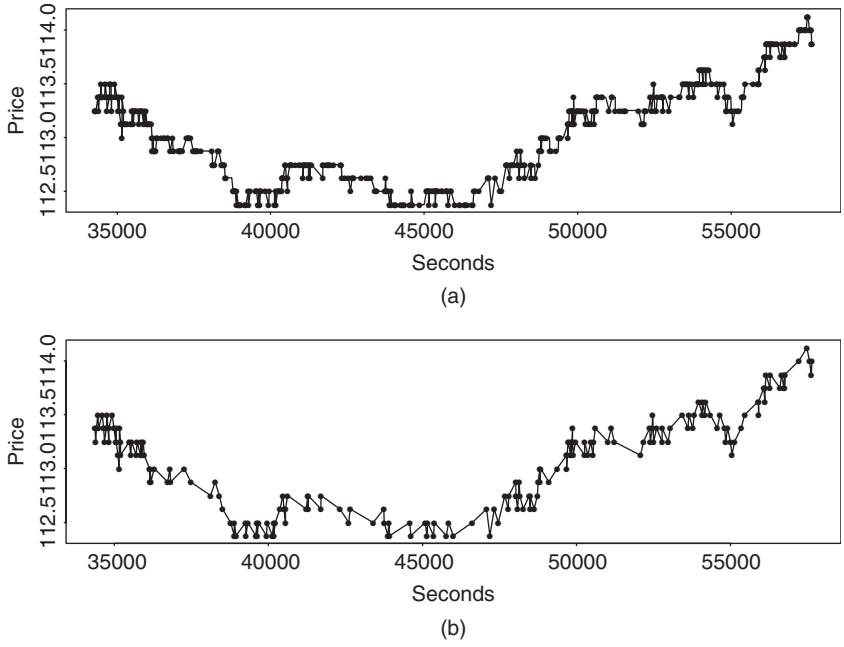

Figure 5.16 Time plots of intraday transaction prices of IBM stock on November 21, 1990: (a) all transactions and (b) transactions that resulted in price change.

where σ is a positive number and $\{\epsilon_i\}$ is a sequence of iid $N(0, 1)$ random variables. This is a multiple linear regression model with lagged variables. Other explanatory variables can be added if necessary. The log transformation is used to ensure the positiveness of time duration.

The conditional model for N_i is further partitioned into two parts because empirical data suggest a concentration of N_i at 0. The first part of the model for N_i is the logit model

$$p(N_i = 0|\Delta t_i, F_{i-1}) = \text{logit}[\alpha_0 + \alpha_1 \ln(\Delta t_i)], \tag{5.49}$$

where $\text{logit}(x) = \exp(x)/[1 + \exp(x)]$, whereas the second part of the model is

$$N_i|(N_i > 0, \Delta t_i, F_{i-1}) \sim 1 + g(\lambda_i), \quad \lambda_i = \frac{\exp[\gamma_0 + \gamma_1 \ln(\Delta t_i)]}{1 + \exp[\gamma_0 + \gamma_1 \ln(\Delta t_i)]}, \tag{5.50}$$

where $\sim$ means "is distributed as," and $g(\lambda)$ denotes a geometric distribution with parameter λ, which is in the interval (0, 1).

The model for direction D_i is

$$D_i|(N_i, \Delta t_i, F_{i-1}) = \text{sign}(\mu_i + \sigma_i \epsilon), \tag{5.51}$$

where ϵ is a $N(0, 1)$ random variable, and

$$\mu_i = \omega_0 + \omega_1 D_{i-1} + \omega_2 \ln(\Delta t_i),$$
$$\ln(\sigma_i) = \beta \left| \sum_{j=1}^{4} D_{i-j} \right| = \beta |D_{i-1} + D_{i-2} + D_{i-3} + D_{i-4}|.$$

In other words, D_i is governed by the sign of a normal random variable with mean μ_i and variance σ_i^2. A special characteristic of the prior model is the function for $\ln(\sigma_i)$. For intraday transactions, a key feature is the *price reversal* between consecutive price changes. This feature is modeled by the dependence of D_i on D_{i-1} in the mean equation with a negative ω_1 parameter. However, there exists an occasional local trend in the price movement. The previous variance equation allows for such a local trend by increasing the uncertainty in the direction of price movement when the past data showed evidence of a local trend. For a normal distribution with a fixed mean, increasing its variance makes a random draw have the same chance to be positive and negative. This in turn increases the chance for a sequence of all positive or all negative draws. Such a sequence produces a local trend in price movement.

To allow for different dynamics between positive and negative price movements, we use different models for the size of a price change. Specifically, we have

$$S_i | (D_i = -1, N_i, \Delta t_i, F_{i-1}) \sim p(\lambda_{d,i}) + 1, \quad \text{with} \tag{5.52}$$
$$\ln(\lambda_{d,i}) = \eta_{d,0} + \eta_{d,1} N_i + \eta_{d,2} \ln(\Delta t_i) + \eta_{d,3} S_{i-1}$$
$$S_i | (D_i = 1, N_i, \Delta t_i, F_{i-1}) \sim p(\lambda_{u,i}) + 1, \quad \text{with} \tag{5.53}$$
$$\ln(\lambda_{u,i}) = \eta_{u,0} + \eta_{u,1} N_i + \eta_{u,2} \ln(\Delta t_i) + \eta_{u,3} S_{i-1},$$

where $p(\lambda)$ denotes a Poisson distribution with parameter λ, and 1 is added to the size because the minimum size is 1 tick when there is a price change.

The specified models in Eqs. (5.48)–(5.53) can be estimated jointly by either the maximum-likelihood method or the Markov chain Monte Carlo methods. Based on Eq. (5.47), the models consist of six conditional models that can be estimated separately.

Example 5.5. Consider the intraday transactions of IBM stock on November 21, 1990. There are 194 price changes within normal trading hours. Figure 5.17 shows the histograms of $\ln(\Delta t_i)$, N_i, D_i, and S_i. The data for D_i are about equally distributed between "upward" and "downward" movements. Only a few transactions resulted in a price change of more than 1 tick; as a matter of fact, there were 7 changes with 2 ticks and 1 change with 3 ticks. Using Markov chain Monte Carlo (MCMC) methods (see Chapter 12), we obtained the following models for the data. The reported estimates and their standard deviations are the posterior means and

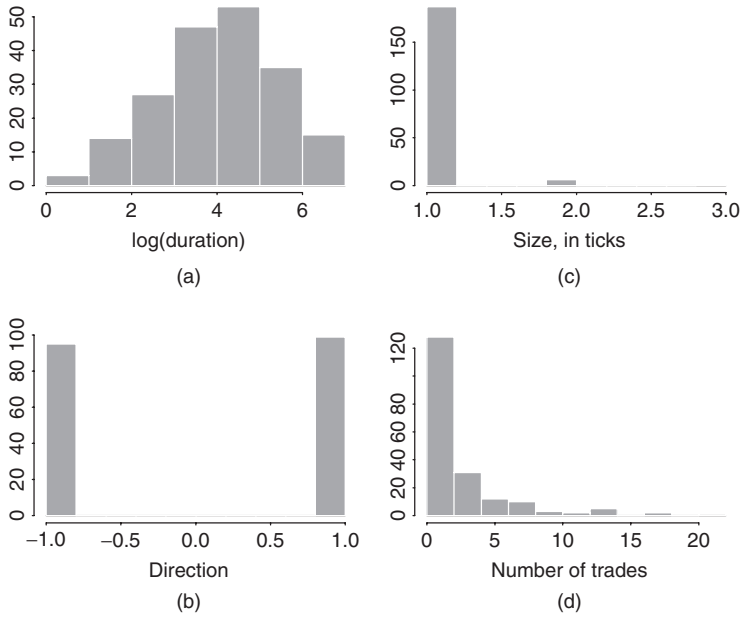

Figure 5.17 Histograms of intraday transactions data for IBM stock on November 21, 1990: (a) log durations between price changes, (b) direction of price movement, (c) size of price change measured in ticks, and (d) number of trades without price change.

standard deviations of MCMC draws with 9500 iterations. The model for the time duration between price changes is

$$\ln(\Delta t_i) = 4.023 + 0.032\ln(\Delta t_{i-1}) - 0.025S_{i-1} + 1.403\epsilon_i,$$

where standard deviations of the coefficients are 0.415, 0.073, 0.384, and 0.073, respectively. The fitted model indicates that there was no dynamic dependence in the time duration. For the N_i variable, we have

$$\Pr(N_i > 0|\Delta t_i, F_{i-1}) = \text{logit}[-0.637 + 1.740\ln(\Delta t_i)],$$

where standard deviations of the estimates are 0.238 and 0.248, respectively. Thus, as expected, the number of trades with no price change in the time interval (t_{i-1}, t_i) depends positively on the length of the interval. The magnitude of N_i when it is positive is

$$N_i|(N_i > 0, \Delta t_i, F_{i-1}) \sim 1 + g(\lambda_i), \qquad \lambda_i = \frac{\exp[0.178 - 0.910\ln(\Delta t_i)]}{1 + \exp[0.178 - 0.910\ln(\Delta t_i)]},$$

where standard deviations of the estimates are 0.246 and 0.138, respectively. The negative and significant coefficient of $\ln(\Delta t_i)$ means that N_i is positively related to the length of the duration Δt_i because a large $\ln(\Delta t_i)$ implies a small λ_i, which

in turn implies higher probabilities for larger N_i; see the geometric distribution in Eq. (5.27).

The fitted model for D_i is

$$\mu_i = 0.049 - 0.840 D_{i-1} - 0.004 \ln(\Delta t_i),$$
$$\ln(\sigma_i) = 0.244 |D_{i-1} + D_{i-2} + D_{i-3} + D_{i-4}|,$$

where standard deviations of the parameters in the mean equation are 0.129, 0.132, and 0.082, respectively, whereas the standard error for the parameter in the variance equation is 0.182. The price reversal is clearly shown by the highly significant negative coefficient of D_{i-1}. The marginally significant parameter in the variance equation is exactly as expected. Finally, the fitted models for the size of a price change are

$$\ln(\lambda_{d,i}) = 1.024 - 0.327 N_i + 0.412 \ln(\Delta t_i) - 4.474 S_{i-1},$$
$$\ln(\lambda_{u,i}) = -3.683 - 1.542 N_i + 0.419 \ln(\Delta t_i) + 0.921 S_{i-1},$$

where standard deviations of the parameters for the "down size" are 3.350, 0.319, 0.599, and 3.188, respectively, whereas those for the "up size" are 1.734, 0.976, 0.453, and 1.459. The interesting estimates of the prior two equations are the negative estimates of the coefficient of N_i. A large N_i means there were more transactions in the time interval (t_{i-1}, t_i) with no price change. This can be taken as evidence of no new information available in the time interval (t_{i-1}, t_i). Consequently, the size for the price change at t_i should be small. A small $\lambda_{u,i}$ or $\lambda_{d,i}$ for a Poisson distribution gives precisely that.

In summary, granted that a sample of 194 observations in a given day may not contain sufficient information about the trading dynamics of IBM stock, but the fitted models appear to provide some sensible results. McCulloch and Tsay (2000) extend the PCD model to a hierarchical framework to handle all the data of the 63 trading days between November 1, 1990, and January 31, 1991. Many of the parameter estimates become significant in this extended sample, which has more than 19,000 observations. For example, the overall estimate of the coefficient of $\ln(\Delta t_{i-1})$ in the model for time duration ranges from 0.04 to 0.1, which is small, but significant.

Finally, using transactions data to test microstructure theory often requires a careful specification of the variables used. It also requires a deep understanding of the way by which the market operates and the data are collected. However, ideas of the econometric models discussed in this chapter are useful and widely applicable in analysis of high-frequency data.

5.8 APPLICATION

In this section we apply the ACD model to stock volatility modeling. Consider the daily range of the log price of Apple stock from January 4, 1999, to November 20,

2007. The data are obtained from Yahoo Finance and consist of 2235 observations. This series was analyzed in Tsay (2009). The range of daily log prices has been used in the literature as a robust alternative to volatility modeling; see Chapter 3 and Chou (2005) and the references therein. Apple stock had two-for-one splits on June 21, 2000, and February 28, 2005, during the sample period, but no adjustments are needed for the splits because we use daily range of log price. As mentioned before, stock prices in the U.S. markets switched from the tick size $\frac{1}{16}$ of a dollar to the decimal system on January 29, 2001. Such a change affected the bid–ask spread of stock prices. We shall employ intervention analysis to study the impact of such a policy change on the stock volatility.

The sample mean, standard deviation, minimum, and maximum of the range of log prices are 0.0407, 0.0218, 0.0068, and 0.1468, respectively. The sample skewness and excess kurtosis are 1.3 and 2.13, respectively. Figure 5.18(a) shows the time plot of the range series. The volatility seems to be increasing from 2000 to 2001, then decreasing to a stable level after 2002. It seems to increase somewhat at the end of the series. Figure 5.19(a) shows the sample ACF of the daily range series. The sample ACFs are highly significant and decay slowly.

We fit EACD(1,1), WACD(1,1), and GACD(1,1) models to the daily range series. The estimation results, along with the Ljung–Box statistics for the standardized residual series and its squared process, are given in Table 5.9. The parameter estimates for the duration equation are stable for all three models, except for the

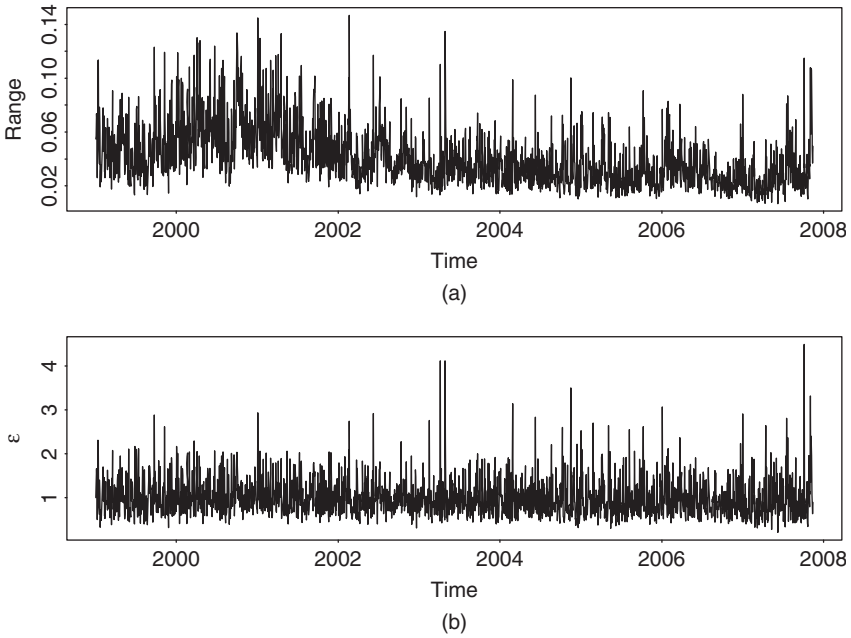

Figure 5.18 Time plots of daily range of log price of Apple stock from January 4, 1999, to November 20, 2007: (a) observed daily range and (b) standardized residuals of a GACD(1,1) model.

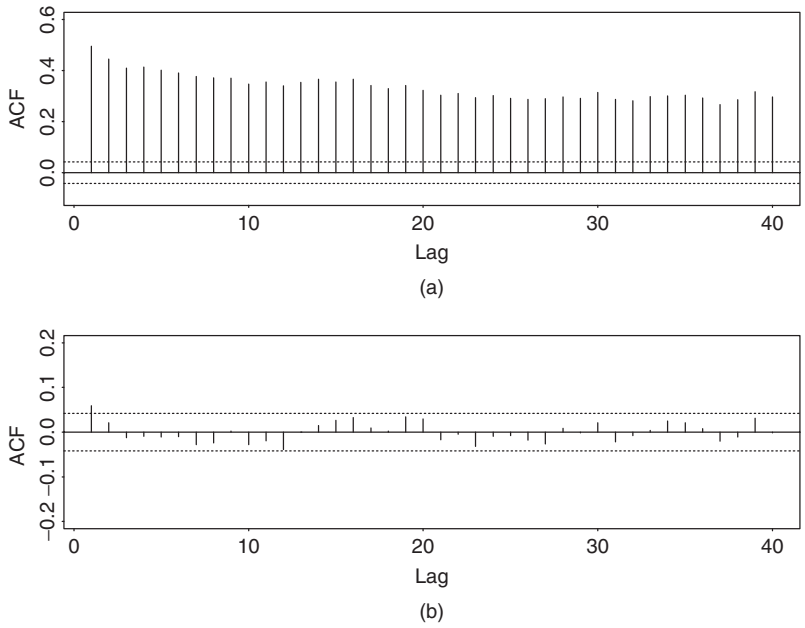

Figure 5.19 Sample autocorrelation function of daily range of log prices of Apple stock from January 4, 1999, to November 20, 2007: (a) ACF of daily range and (b) ACF of standardized residual series of GACD(1,1) model.

TABLE 5.9 Estimation Results of EACD(1,1), WACD(1,1), and GACD(1,1) Models for Daily Range of Log Prices of Apple Stock from January 4, 1999 to November 20, 2007[a]

	Parameters					Checking	
Model	α_0	α_1	β_1	α	κ	$Q(10)$	$Q^*(10)$
EACD	0.0007	0.133	0.849			16.65	12.12
	(0.0005)	(0.036)	(0.044)			(0.082)	(0.277)
WACD	0.0013	0.131	0.835	2.377		13.66	9.74
	(0.0003)	(0.015)	(0.021)	(0.031)		(0.189)	(0.464)
GACD	0.0010	0.133	0.843	1.622	2.104	14.62	11.21
	(0.0002)	(0.015)	(0.019)	(0.029)	(0.040)	(0.147)	(0.341)

[a]The standard errors of the estimates and the p values of the Ljung–Box statistics are in parentheses, where $Q(10)$ and $Q^*(10)$ are for standardized residual series and its squared process, respectively.

constant term of the EACD model, which appears to be statistically insignificant at the usual 5% level. Indeed, in this particular instance, the EACD(1,1) model fares slightly worse than the other two ACD models. Between the WACD(1,1) and GACD(1,1) models, we slightly prefer the GACD(1,1) model because it fits the data better and is more flexible.

Figure 5.19(b) shows the sample ACFs of the standardized residuals of the fitted GACD(1,1) model. From the plot, the standardized residuals do not have significant serial correlations, even though the lag-1 sample ACF is slightly above its two standard error limit. The lag-1 serial correlation is removed when we use nonlinear ACD models later. Figure 5.18(b) shows the time plot of the standardized residuals of the GACD(1,1) model. The residuals do not show any pattern of model inadequacy. The mean, standard deviation, minimum, and maximum of the standardized residuals are 0.203, 4.497, 0.999, and 0.436, respectively.

It is interesting to see that the estimates of the shape parameter α are greater than 1 for both WACD(1,1) and GACD(1,1) models, indicating that the hazard function of the daily range is monotonously increasing. This is consistent with the idea of volatility clustering, for large volatility tends to be followed by another large volatility.

Threshold ACD model

To refine the GACD(1,1) model for the daily range of log prices of Apple stock, we employ a two-regime threshold WACD(1,1) model. Some preliminary analysis of the threshold WACD models indicates that the major difference in the parameter estimates between the two regimes is the shape parameter of the Weibull distribution. Thus, we focus on a TWACD(2;1,1) model with different shape parameters for the two regimes.

Table 5.10 gives the maximized log-likelihood value of a TWACD(2;1,1) model with delay $d = 1$ and threshold $r \in \{x_{(q)} | q = 60, 65, \ldots, 95\}$, where $x_{(q)}$ denotes the sample qth percentile. From the table, the threshold 0.04753 is selected, which is the 70th percentile of the data. The fitted model is

$$x_i = \psi_i \epsilon_i, \qquad \psi_i = 0.0013 + 0.1539 x_{i-1} + 0.8131 \psi_{i-1},$$

where the standard errors of the coefficients are 0.0003, 0.0164, and 0.0215, respectively, and ϵ_i follows the standardized Weibull distribution as

$$\epsilon_i \sim \begin{cases} W(2.2756) & \text{if } x_{i-1} \leq 0.04753, \\ W(2.7119) & \text{otherwise,} \end{cases}$$

where the standard errors of the two shape parameters are 0.0394 and 0.0717, respectively.

TABLE 5.10 Selection of Threshold of TWACD(2;1,1) Model for Daily Range of Log Prices of Apple Stock from January 4, 1999, to November 20, 2007[a]

Quantile	60	65	70	75	80	85	90	95
$r \times 100$	4.03	4.37	4.75	5.15	5.58	6.16	7.07	8.47
$\ell(r) \times 10^3$	6.073	6.076	6.079	6.076	6.078	6.074	6.072	6.066

[a]The threshold variable is x_{i-1}.

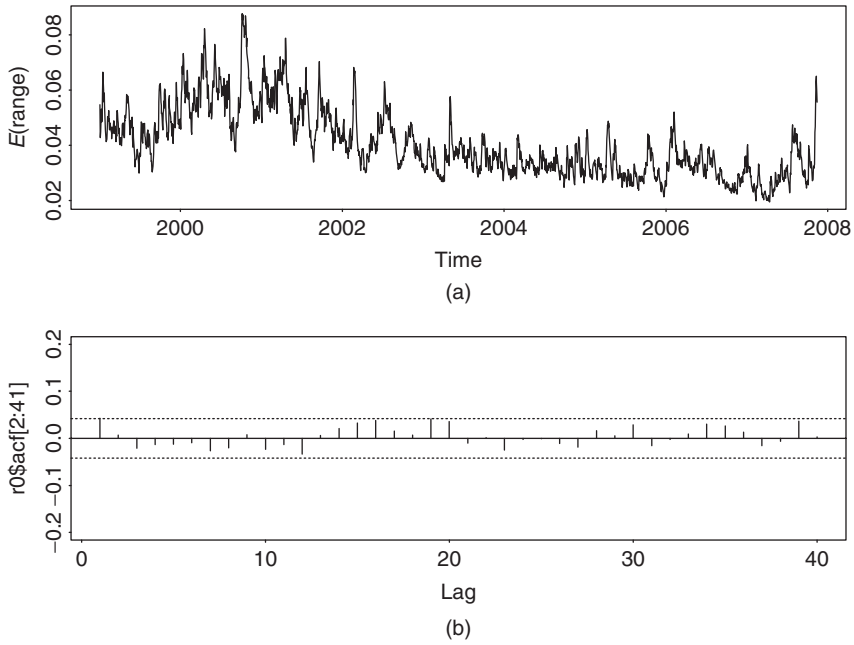

Figure 5.20 Model fitting for daily range of log price of Apple stock from January 4, 1999, to November 20, 2007: (a) conditional expected durations of fitted TWACD(2;1,1) model and (b) sample ACF of standardized residuals.

Figure 5.20(a) shows the time plot of the conditional expected duration for the fitted TWACD(2;1,1) model, that is, $\hat{\psi}_i$, whereas Figure 5.20(b) gives the residual ACFs for the fitted model. All residual ACFs are within the two standard error limits. Indeed, we have $Q(1) = 4.01(0.05)$ and $Q(10) = 9.84(0.45)$ for the standardized residuals and $Q^*(1) = 0.83(0.36)$ and $Q^*(10) = 9.35(0.50)$ for the squared series of the standardized residuals, where the number in parentheses denotes p value. Note that the threshold variable x_{i-1} is also selected based on the value of the log-likelihood function. For instance, the log-likelihood function of the TWACD(2;1,1) model assumes the value 6.069×10^3 and 6.070×10^3, respectively, for $d = 2$ and 3 when the threshold is 0.04753. These values are lower than that when $d = 1$.

Intervention Analysis

High-frequency financial data are often influenced by external events, for example, an increase or drop in interest rates by the U.S. Federal Open Market Committee or a jump in the oil price. Applications of ACD models in finance are often faced with the problem of outside interventions. To handle the effects of external events, the intervention analysis of Box and Tiao (1975) can be used. Here we apply the analysis to the daily range series of Apple stock to study the impact of change in tick size on the stock volatility.

Let t_o be the time of intervention. For the Apple stock, $t_o = 522$, which corresponds to January 26, 2001, the last trading day before the change in tick size. Since more observations in the sample are after the intervention, we define the indicator variable

$$I_i^{(t_o)} = \begin{cases} 1 & \text{if } i \le t_o, \\ 0 & \text{otherwise,} \end{cases}$$

to signify the absence of intervention. Since a larger tick size tends to increase the observed daily price range, it is reasonable to assume that the conditional expected range would be higher before the intervention. A simple intervention model for the daily range of Apple stock is then given by

$$x_i = \psi_i \begin{cases} \epsilon_{1i} & \text{if } x_{i-1} \le 0.04753, \\ \epsilon_{2i} & \text{otherwise,} \end{cases}$$

where ψ_i follows the model

$$\psi_i = \alpha_0 + \gamma I_i^{(t_o)} + \alpha_1 x_{i-1} + \beta_1 \psi_{i-1}, \tag{5.54}$$

where γ denotes the decrease in expected duration due to the decimalization of stock prices. In other words, the expected durations before and after the intervention are

$$\frac{\alpha_0 + \gamma}{1 - \alpha_1 - \beta_1} \quad \text{and} \quad \frac{\alpha_0}{1 - \alpha_1 - \beta_1},$$

respectively. We expect $\gamma > 0$.

The fitted duration equation for the intervention model is

$$\psi_i = 0.0021 + 0.0011 I_i^{(522)} + 0.1595 x_{i-1} + 0.7828 \psi_{i-1},$$

where the standard errors of the estimates are 0.0004, 0.0003, 0.0177, and 0.0264, respectively. The estimate $\hat{\gamma}$ is significant at the 1% level. For the innovations, we have

$$\epsilon_i \sim \begin{cases} W(2.2835) & \text{if } x_{i-1} \le 0.04753, \\ W(2.7322) & \text{otherwise.} \end{cases}$$

The standard errors of the two estimates of the shape parameter are 0.0413 and 0.0780, respectively. Figure 5.21(a) shows the expected durations of the intervention model, and Figure 5.21(b) shows the ACF of the standardized residuals. All residual ACFs are within the two standard error limits. Indeed, for the standardized residuals, we have $Q(1) = 2.37(0.12)$ and $Q(10) = 6.24(0.79)$. For the squared series of the standardized residuals, we have $Q^*(1) = 0.34(0.56)$ and $Q^*(10) = 6.79(0.75)$. As expected, $\hat{\gamma} > 0$ so that the decimalization indeed reduces the expected value of the daily range. This simple analysis shows that, as expected, adopting the decimal system reduces the volatility of Apple stock.

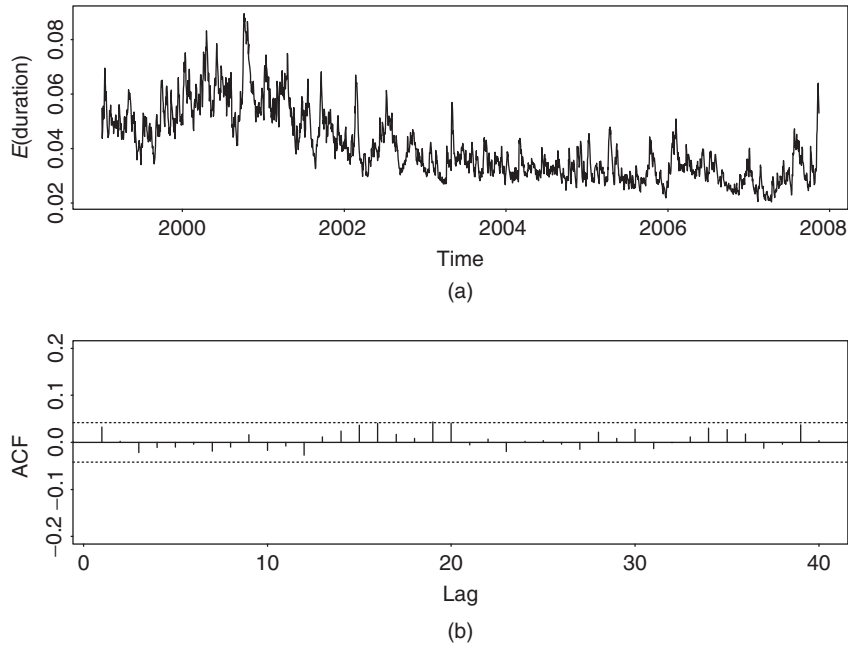

Figure 5.21 Model fitting for daily range of log price of Apple stock from January 4, 1999, to November 20, 2007: (a) conditional expected durations of fitted TWACD(2;1,1) model with intervention and (b) sample ACF of corresponding standardized residuals.

APPENDIX A: REVIEW OF SOME PROBABILITY DISTRIBUTIONS

Exponential Distribution

A random variable X has an exponential distribution with parameter $\beta > 0$ if its probability density function (pdf) is given by

$$f(x|\beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Denoting such a distribution by $X \sim \exp(\beta)$, we have $E(X) = \beta$ and $\text{Var}(X) = \beta^2$. The cumulative distribution function (CDF) of X is

$$F(x|\beta) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - e^{-x/\beta} & \text{if } x \geq 0. \end{cases}$$

When $\beta = 1$, X is said to have a standard exponential distribution.

Gamma Function

For $\kappa > 0$, the gamma function $\Gamma(\kappa)$ is defined by

$$\Gamma(\kappa) = \int_0^\infty x^{\kappa-1} e^{-x}\, dx.$$

The most important properties of the gamma function are:

1. For any $\kappa > 1$, $\Gamma(\kappa) = (\kappa - 1)\Gamma(\kappa - 1)$.
2. For any positive integer m, $\Gamma(m) = (m - 1)!$.
3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

The integration

$$\Gamma(y|\kappa) = \int_0^y x^{\kappa-1} e^{-x}\, dx, \qquad y > 0$$

is an *incomplete* gamma function. Its values have been tabulated in the literature. Computer programs are now available to evaluate the incomplete gamma function.

Gamma Distribution

A random variable X has a gamma distribution with parameter κ and β ($\kappa > 0$, $\beta > 0$) if its pdf is given by

$$f(x|\kappa, \beta) = \begin{cases} \dfrac{1}{\beta^\kappa \Gamma(\kappa)} x^{\kappa-1} e^{-x/\beta} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

By changing variable $y = x/\beta$, one can easily obtain the moments of X:

$$\begin{aligned} E(X^m) &= \int_0^\infty x^m f(x|\kappa, \beta)\, dx = \frac{1}{\beta^\kappa \Gamma(\kappa)} \int_0^\infty x^{\kappa+m-1} e^{-x/\beta}\, dx \\ &= \frac{\beta^m}{\Gamma(\kappa)} \int_0^\infty y^{\kappa+m-1} e^{-y}\, dy = \frac{\beta^m \Gamma(\kappa + m)}{\Gamma(\kappa)}. \end{aligned}$$

In particular, the mean and variance of X are $E(X) = \kappa\beta$ and $\text{Var}(X) = \kappa\beta^2$. When $\beta = 1$, the distribution is called a standard gamma distribution with parameter κ. We use the notation $G \sim \text{gamma}(\kappa)$ to denote that G follows a standard gamma distribution with parameter κ. The moments of G are

$$E(G^m) = \frac{\Gamma(\kappa + m)}{\Gamma(\kappa)}, \qquad m > 0. \tag{5.55}$$

Weibull Distribution

A random variable X has a Weibull distribution with parameters α and β ($\alpha > 0$, $\beta > 0$) if its pdf is given by

$$f(x|\alpha, \beta) = \begin{cases} \dfrac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

where β and α are the scale and shape parameters of the distribution. The mean and variance of X are

$$E(X) = \beta\Gamma\left(1 + \frac{1}{\alpha}\right), \qquad \text{Var}(X) = \beta^2 \left\{\Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^2\right\},$$

and the CDF of X is

$$F(x|\alpha, \beta) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - e^{-(x/\beta)^\alpha} & \text{if } x \geq 0. \end{cases}$$

When $\alpha = 1$, the Weibull distribution reduces to an exponential distribution.

Define $Y = X/[\beta\Gamma(1 + 1/\alpha)]$. We have $E(Y) = 1$ and the pdf of Y is

$$f(y|\alpha) = \begin{cases} \alpha \left[\Gamma\left(1 + \dfrac{1}{\alpha}\right)\right]^\alpha y^{\alpha-1} \exp\left\{-\left[\Gamma\left(1 + \dfrac{1}{\alpha}\right) y\right]^\alpha\right\} & \text{if } y \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{5.56}$$

where the scale parameter β disappears due to standardization. The CDF of the standardized Weibull distribution is

$$F(y|\alpha) = \begin{cases} 0 & \text{if } y < 0, \\ 1 - \exp\left\{-\left[\Gamma\left(1 + \dfrac{1}{\alpha}\right) y\right]^\alpha\right\} & \text{if } y > 0, \end{cases}$$

and we have $E(Y) = 1$ and $\text{Var}(Y) = \Gamma(1 + 2/\alpha)/[\Gamma(1 + 1/\alpha)]^2 - 1$. For a duration model with Weibull innovations, the pdf in Eq. (5.56) is used in the maximum-likelihood estimation.

Generalized Gamma Distribution

A random variable X has a generalized gamma distribution with parameter α, β, κ ($\alpha > 0$, $\beta > 0$, and $\kappa > 0$) if its pdf is given by

$$f(x|\alpha, \beta, \kappa) = \begin{cases} \dfrac{\alpha x^{\kappa\alpha-1}}{\beta^{\kappa\alpha}\Gamma(\kappa)} \exp\left[-\left(\dfrac{x}{\beta}\right)^\alpha\right] & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where β is a scale parameter, and α and κ are shape parameters. This distribution can be written as

$$G = \left(\frac{X}{\beta}\right)^{\alpha},$$

where G is a standard gamma random variable with parameter κ. The pdf of X can be obtained from that of G by the technique of changing variables. Similarly, the moments of X can be obtained from that of G in Eq. (5.55) by

$$E(X^m) = E[(\beta G^{1/\alpha})^m] = \beta^m E(G^{m/\alpha}) = \beta^m \frac{\Gamma(\kappa + m/\alpha)}{\Gamma(\kappa)} = \frac{\beta^m \Gamma(\kappa + m/\alpha)}{\Gamma(\kappa)}.$$

When $\kappa = 1$, the generalized gamma distribution reduces to that of a Weibull distribution. Thus, the exponential and Weibull distributions are special cases of the generalized gamma distribution.

The expectation of a generalized gamma distribution is $E(X) = \beta\Gamma(\kappa + 1/\alpha)/\Gamma(\kappa)$. In duration models, we need a distribution with unit expectation. Therefore, defining a random variable $Y = \lambda X/\beta$, where $\lambda = \Gamma(\kappa)/\Gamma(\kappa + 1/\alpha)$, we have $E(Y) = 1$ and the pdf of Y is

$$f(y|\alpha, \kappa) = \begin{cases} \dfrac{\alpha y^{\kappa\alpha - 1}}{\lambda^{\kappa\alpha}\Gamma(\kappa)} \exp\left[-\left(\dfrac{y}{\lambda}\right)^{\alpha}\right] & \text{if } y > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{5.57}$$

where again the scale parameter β disappears and $\lambda = \Gamma(\kappa)/\Gamma(\kappa + 1/\alpha)$.

APPENDIX B: HAZARD FUNCTION

A useful concept in modeling duration is the *hazard function* implied by a distribution function. For a random variable X, the *survival function* is defined as

$$S(x) \equiv P(X > x) = 1 - P(X \leq x) = 1 - \text{CDF}(x), \qquad x > 0,$$

which gives the probability that a subject, which follows the distribution of X, survives at the time x. The hazard function (or intensity function) of X is then defined by

$$h(x) = \frac{f(x)}{S(x)}, \tag{5.58}$$

where $f(\cdot)$ and $S(\cdot)$ are the pdf and survival function of X, respectively.

Example 5.6. For the Weibull distribution with parameters α and β, the survival function and hazard function are

$$S(x|\alpha, \beta) = \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right], \qquad h(x|\alpha, \beta) = \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1}, \qquad x > 0.$$

In particular, when $\alpha = 1$, we have $h(x|\beta) = 1/\beta$. Therefore, for an exponential distribution, the hazard function is constant. For a Weibull distribution, the hazard is a monotone function. If $\alpha > 1$, then the hazard function is monotonously increasing. If $\alpha < 1$, the hazard function is monotonously decreasing. For the generalized gamma distribution, the survival function and, hence, the hazard function involve the incomplete gamma function. Yet the hazard function may exhibit various patterns, including U shape or inverted U shape. Thus, the generalized gamma distribution provides a flexible approach to modeling the duration of stock transactions.

For the standardized Weibull distribution, the survival and hazard functions are

$$S(y|\alpha) = \exp\left\{-\left[\Gamma\left(1+\frac{1}{\alpha}\right) y\right]^{\alpha}\right\},$$
$$h(y|\alpha) = \alpha\left[\Gamma\left(1+\frac{1}{\alpha}\right)\right]^{\alpha} y^{\alpha-1}, \qquad y > 0.$$

APPENDIX C: SOME RATS PROGRAMS FOR DURATION MODELS

The data used are adjusted time durations of intraday transactions of IBM stock from November 1 to November 9, 1990. The file name is `ibm1to5.txt` and it has 3534 observations.

Program for Estimating a WACD(1,1) Model

```
all 0  3534:1
open data ibm1to5.txt
data(org=obs) / x r1
set psi = 1.0
nonlin a0 a1 b1 al
frml gvar = a0+a1*x(t-1)+b1*psi(t-1)
frml gma = %LNGAMMA(1.0+1.0/al)
frml gln =al*gma(t)+log(al)-log(x(t)) $
  +al*log(x(t)/(psi(t)=gvar(t)))-(exp(gma(t))*x(t)/psi(t))**al
smpl 2 3534
compute a0 = 0.2, a1 = 0.1, b1 = 0.1, al = 0.8
maximize(method=bhhh,recursive,iterations=150) gln
set fv = gvar(t)
set resid = x(t)/fv(t)
set residsq = resid(t)*resid(t)
```

```
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

Program for Estimating a GACD(1,1) Model

```
all 0  3534:1
open data ibm1to5.txt
data(org=obs) / x r1
set psi = 1.0
nonlin a0 a1 b1 al ka
frml cv = a0+a1*x(t-1)+b1*psi(t-1)
frml gma = %LNGAMMA(ka)
frml lam = exp(gma(t))/exp(%LNGAMMA(ka+(1.0/al)))
frml xlam = x(t)/(lam(t)*(psi(t)=cv(t)))
frml gln =-gma(t)+log(al/x(t))+ka*al*log(xlam(t))
   -(xlam(t))**al
smpl 2 3534
compute a0 = 0.238, a1 = 0.075, b1 = 0.857, al = 0.5, ka = 4.0
nlpar(criterion=value,cvcrit=0.00001)
maximize(method=bhhh,recursive,iterations=150) gln
set fv = cv(t)
set resid = x(t)/fv(t)
set residsq = resid(t)*resid(t)
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

Program for Estimating a TAR-WACD(1,1) Model

The threshold 3.79 is prespecified.

```
all 0 3534:1
open data ibm1to5.txt
data(org=obs) / x rt
set psi = 1.0
nonlin a1 a2 al b0 b2 bl
frml u = ((x(t-1)-3.79)/abs(x(t-1)-3.79)+1.0)/2.0
frml cp1 = a1*x(t-1)+a2*psi(t-1)
frml gma1 = %LNGAMMA(1.0+1.0/al)
frml cp2 = b0+b2*psi(t-1)
frml gma2 = %LNGAMMA(1.0+1.0/bl)
frml cp = cp1(t)*(1-u(t))+cp2(t)*u(t)
frml gln1 =al*gma1(t)+log(al)-log(x(t)) $
 +al*log(x(t)/(psi(t)=cp(t)))-(exp(gma1(t))*x(t)/psi(t))**al
frml gln2 =bl*gma2(t)+log(bl)-log(x(t)) $
 +bl*log(x(t)/(psi(t)=cp(t)))-(exp(gma2(t))*x(t)/psi(t))**bl
frml gln = gln1(t)*(1-u(t))+gln2(t)*u(t)
smpl 2 3534
compute a1 = 0.2, a2 = 0.85, al = 0.9
```

```
compute b0 = 1.8, b2 = 0.5, b1 = 0.8
maximize(method=bhhh,recursive,iterations=150) gln
set fv = cp(t)
set resid = x(t)/fv(t)
set residsq = resid(t)*resid(t)
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

EXERCISES

5.1. Let r_t be the log return of an asset at time t. Assume that $\{r_t\}$ is a Gaussian white noise series with mean 0.05 and variance 1.5. Suppose that the probability of a trade at each time point is 40% and is independent of r_t. Denote the observed return by r_t^o. Is r_t^o serially correlated? If yes, calculate the first three lags of autocorrelations of r_t^o.

5.2. Let P_t be the observed market price of an asset, which is related to the fundamental value of the asset P_t^* via Eq. (5.9). Assume that $\Delta P_t^* = P_t^* - P_{t-1}^*$ forms a Gaussian white noise series with mean zero and variance 1.0. Suppose that the bid–ask spread is two ticks. What is the lag-1 autocorrelation of the price change series $\Delta P_t = P_t - P_{t-1}$ when the tick size is $\$\frac{1}{8}$? What is the lag-1 autocorrelation of the price change when the tick size is $\$\frac{1}{16}$?

5.3. The file `ibm-d2-dur.txt` contains the adjusted durations between trades of IBM stock on November 2, 1990. The file has three columns consisting of day, time of trade measured in seconds from midnight, and adjusted durations.

(a) Build an EACD model for the adjusted duration and check the fitted model.

(b) Build a WACD model for the adjusted duration and check the fitted model.

(c) Build a GACD model for the adjusted duration and check the fitted model.

(d) Compare the prior three duration models.

5.4. The file `mmm9912-dtp.txt` contains the transactions data of the stock of 3M Company in December 1999. There are three columns: day of the month, time of transaction in seconds from midnight, and transaction price. Transactions that occurred after 4:00 PM Eastern time are excluded.

(a) Is there a diurnal pattern in 3M stock trading? You may construct a time series n_t, which denotes the number of trades in a 5-minute time interval to answer this question.

(b) Use the price series to confirm the existence of a bid–ask bounce in intraday trading of 3M stock.

(c) Tabulate the frequencies of price change in multiples of tick size $\$\frac{1}{16}$. You may combine changes with 5 ticks or more into a category and those with -5 ticks or beyond into another category.

5.5. Consider again the transactions data of 3M stock in December 1999.

(a) Use the data to construct an intraday 5-minute log return series. Use the simple average of all transaction prices within a 5-minute interval as the stock price for the interval. Is the series serially correlated? You may use Ljung–Box statistics to test the hypothesis with the first 10 lags of the sample autocorrelation function.

(b) There are seventy-seven 5-minute returns in a normal trading day. Some researchers suggest that the sum of squares of the intraday 5-minute returns can be used as a measure of daily volatility. Apply this approach and calculate the daily volatility of the log return of 3M stock in December 1999. Discuss the validity of such a procedure to estimate daily volatility.

5.6. The file `mmm9912-adur.txt` contains an adjusted intraday trading duration of 3M stock in December 1999. There are thirty-nine 10-minute time intervals in a trading day. Let d_i be the average of all log durations for the ith 10-minute interval across all trading days in December 1999. Define an adjusted duration as $t_j / \exp(d_i)$, where j is in the ith 10-minute interval. Note that more sophisticated methods can be used to adjust the diurnal pattern of trading duration. Here we simply use a local average.

(a) Is there a diurnal pattern in the adjusted duration series? Why?

(b) Build a duration model for the adjusted series using exponential innovations. Check the fitted model.

(c) Build a duration model for the adjusted series using Weibull innovations. Check the fitted model.

(d) Build a duration model for the adjusted series using generalized gamma innovations. Check the fitted model.

(e) Compare and comment on the three duration models built before.

5.7. To gain experience in analyzing high-frequency financial data, consider the trade data of Boeing stock from December 1 to December 5, 2008. The data are in five files: `taq-td-ba12012008.txt` to `taq-td-ba12052008.txt`. Each file has five columns, namely hour, minute, second, price, and volume. Only transactions within the normal trading hours (9:30 AM to 4:00 PM Eastern time) are kept. Construct a time series of the number of trades in an intraday 5-minute time interval. Is there any diurnal pattern in the constructed series? You can simply compute the sample ACF of the series to answer this question.

5.8. Again, consider the high-frequency data of Boeing stock from December 1 to December 5, 2008. Construct an intraday 5-minute return series. Note that

the price of the stock in a 5-minute interval (e.g., 9:30 to 9:35 AM) is the last transaction price within the time interval. For simplicity, ignore overnight returns. Are there serial correlations in the 5-minute return series? Use 10 lags of the ACF and 5% significance level to perform of test.

5.9. Consider the same problem as in Exercise 5.8, but use 10-minute time intervals.

5.10. Again, consider the high-frequency data of Boeing stock. Compute the percentage of consecutive transactions without price change in the sample.

REFERENCES

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discusions). *Journal of the Royal Statistical Society*, Series **B 26**: 211–246.

Box, G. E. P. and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association* **70**: 70–79.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press, Princeton, NJ.

Cho, D., Russell, J. R., Tiao, G. C., and Tsay, R. S. (2003). The magnet effect of price limits: Evidence from high frequency data on Taiwan stock exchange. *Journal of Empirical Finance* **10**: 133–168.

Chou, R. Y. (2005). Forecasting financial volatilities with extreme values: The conditional autoregressive range (CARR) model. *Journal of Money, Credit and Banking* **37**: 561–582.

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica* **66**: 1127–1162.

Ghysels, E. (2000). Some econometric recipes for high-frequency data cooking. *Journal of Business and Economic Statistics* **18**: 154–163.

Hasbrouck, J. (1992). *Using the TORQ Database*. Stern School of Business, New York University, New York.

Hasbrouck, J. (1999). The dynamics of discrete bid and ask quotes. *Journal of Finance* **54**: 2109–2142.

Hauseman, J., Lo, A., and MacKinlay, C. (1992). An ordered probit analysis of transaction stock prices. *Journal of Financial Economics* **31**: 319–379.

Lo, A. and MacKinlay, A. C. (1990). An econometric analysis of nonsynchronous trading. *Journal of Econometrics* **45**: 181–212.

McCulloch, R. E. and Tsay, R. S. (2000). Nonlinearity in high frequency data and hierarchical models. *Studies in Nonlinear Dynamics and Econometrics* **5**: 1–17.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance* **39**: 1127–1140.

Rydberg, T. H. and Shephard, N. (2003). Dynamics of trade-by-trade price movements: Decomposition and models. *Journal of Financial Econometrics* **1**: 2–25.

Stoll, H. and Whaley, R. (1990). Stock market structure and volatility. *Review of Financial Studies* **3**: 37–71.

Tsay, R. S. (2009). Autoregressive conditional duration models. In *Applied Econometrics, Palgrave Handbook of Econometrics*, Vol. 2, T. C. Mills and K. Patterson (eds.), Basingstoke, Hampshire, UK.

Wood, R. A. (2000). Market microstructure research databases: History and projections. *Journal of Business & Economic Statistics* **18**: 140–145.

Zhang, M. Y., Russell, J. R., and Tsay, R. S. (2001). A nonlinear autoregressive conditional duration model with applications to financial transaction data. *Journal of Econometrics* **104**: 179–207.

Zhang, M. Y., Russell, J. R., and Tsay, R. S. (2008). Determinants of bid and ask quotes and implications for the cost of trading. *Journal of Empirical Finance* **15**: 656–678.