High-Frequency Data Analysis and Market Microstructure
[Tsay (2005), chapter 5]

High-frequency data have some unique characteristics that do not appear in lower frequencies.

At this class we have:
- Nonsynchronous trading
- Bid-ask spread
- Duration models
- Price movements that are in multiples of tick size
- Bivariate models for price changes
- Time durations between transactions associated with price changes
Nonsynchronous Trading

Stock tradings 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.

We use for daily series the closing price, the last transaction price of the stock in a trading day. So, we incorrectly assume daily returns as an equally spaced time series with a 24-hour interval and it can lead to erroneous conclusions about the predictability of stock returns even if the true return series are serially independent.

For daily stock return, nonsynchronous trading can introduce:
  • Lag-1 cross-correlation between stock returns;
  • Lag-1 serial correlation in a portfolio return;
  • In some situations, negative serial correlations of the return series of a single stock.
Example: Two independent stocks A and B. A is traded more frequently than B.

Suppose a special new that arrives near the closing hour on one day.

- A is more likely than B to show the effect of the news on the same day.
- The effect on B will eventually appear, but it may be delayed until the following trading day.
- So, stock A appears to lead that of stock B.

- 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.
Lo and Mackinlay (1990) model (simplified):

Let $r_t$ be the continuously compounded return of a security at the time index $t$. 
\{r_t\} is a sequence of iid random variables with mean $E(r_t) = \mu$ and variance $Var(r_t) = \sigma^2$.

Let $\pi$ be the probability that the security is not traded.

Let $r_t^o$ be the observed return.

When there is no trade at time index $t$, so we have $r_t^o = 0$.

When there is a trade at time index $t$, we define $r_t^o$ as the cumulative return from the previous trade, $r_t^o = r_t + r_{t-1} + \cdots + r_{t-k_t}$,

$k_t$ is the largest non-negative integer such that no trade occurred in the periods $t - k_t, t - k_t + 1, \ldots, t - 1$. 

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

\[ r_t^o = r_t \iff \text{there are trades at both } t \text{ and } t-1. \text{ Probability is } (1-\pi)^2. \]

\[ r_t^o = r_t + r_{t-1} \iff \text{there are trades at both } t \text{ and } t-2, \text{ but no trade at } t-1. \text{ Probability is } (1-\pi)^2 \pi. \]

\[ r_t^o = r_t + r_{t-1} + r_{t-2} \iff \text{there are trades at both } t \text{ and } t-3, \text{ but no trade at } t-1 \text{ and } t-2, \text{ and so on. Probability is } (1-\pi)^2 \pi^2. \]
As expected,
\[
\pi + (1 - \pi)^2[1 + \pi + \pi^2 + \cdots] = \pi + (1 - \pi)^2 \frac{1}{1 - \pi} = \pi + 1 - \pi = 1
\]

What is the expectation of \( r_t^o \)?

\[
E(r_t^o) = (1 - \pi)^2 E(r_t) + (1 - \pi)^2 \pi E(r_t + r_{t-1}) + \cdots
\]
\[
= (1 - \pi)^2 \mu + (1 - \pi)^2 \pi 2 \mu + (1 - \pi)^2 \pi^2 3 \mu + \cdots
\]
\[
= (1 - \pi)^2 \mu[1 + 2\pi + 3\pi^2 + 4\pi^3 + \cdots]
\]
\[
= (1 - \pi)^2 \mu \frac{1}{(1 - \pi)^2} = \mu.
\]

We use the result \( 1 + 2\pi + 3\pi^2 + 4\pi^3 + \cdots = 1/(1 - \pi)^2 \)
What is the variance of \( r_t^o \)?

We know that \( \text{Var}(r_t^o) = E[(r_t^o)^2] - [E(r_t^o)]^2 \), so first we need \( E(r_t^o)^2 \).

\[
E(r_t^o)^2 = (1 - \pi)^2 E[(r_t)^2] + (1 - \pi)^2 \pi E[(r_t + r_{t-1})^2] + \cdots \\
= (1 - \pi)^2[(\sigma^2 + \mu^2) + \pi(2\sigma^2 + 4\mu^2) + \pi^2(3\sigma^2 + 9\mu^2) + \cdots] \\
= (1 - \pi)^2\{\sigma^2[1 + 2\pi + 3\pi^2 + \cdots] + \mu^2[1 + 4\pi + 9\pi^2 + \cdots]\} \\
= \sigma^2 + \mu^2\left[\frac{2}{1 - \pi} - 1\right].
\]

To see the last equality, let 
\[
H = 1 + 4\pi + 9\pi^2 + 16\pi^3 + \cdots \quad \text{and} \quad G = 1 + 3\pi + 5\pi^2 + 7\pi^3 + \cdots
\]

It’s easy to see that \((1 - \pi)H = G\) and \((1 - \pi)G = 1 + 2\pi + 2\pi^2 + 2\pi^3 + 2\pi^4 + \cdots\)

So \((1 - \pi)^2 H = 2(1 + \pi + \pi^2 + \pi^3 + \cdots) - 1 = \text{term in brackets.}\)
Then,

$$\text{Var}(r_t^0) = \sigma^2 + \mu^2 \left[ \frac{2}{1 - \pi} - 1 \right] - \mu^2 = \sigma^2 + \frac{2\pi \mu^2}{1 - \pi}$$
What is the lag-1 autocovariance of $r_t^o$?

We know that $\text{Cov}(r_t^o, r_{t-1}^o) = E(r_t^o r_{t-1}^o) - E(r_t^o)E(r_{t-1}^o) = E(r_t^o r_{t-1}^o) - \mu^2$, so first we need $E(r_t^o r_{t-1}^o)$.

We have, $r_t^o r_{t-1}^o$ is zero if there is no trade at $t$, no trade at $t-1$, or no trade at both $t$ and $t-1$. So, the probability is the sum $\pi(1- \pi) + \pi(1- \pi) + \pi^2 = 2\pi - \pi^2$.

Notice that $r_t^o r_{t-1}^o = r_t r_{t-1}$ $\iff$ there are three consecutive trades at $t-2$, $t-1$ and $t$ with probability $(1- \pi)^3$.

And so on,
It's easy to see that: 

$$2\pi - \pi^2 + (1 - \pi)^3[1 + \pi + \pi^2 + \pi^3 + \ldots] = 1$$
Because \( E(r_t r_{t-j}) = E(r_t)E(r_{t-j}) = \mu^2 \) for \( j > 0 \), we have

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

And,

\[
\text{Cov}(r_t^o, r_{t-1}^o) = -\pi \mu^2
\]

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

**Conclusion:** the nonsynchronous trading induces a *negative* lag-1 autocorrelation in daily return, \( r_t^o \).
In general,

$$\text{Cov}(r_t^0, r_{t-j}^0) = -\mu^2 \pi^j, \quad j \geq 1$$

When $\mu \neq 0$, the nonsynchronous trading induces negative autocorrelations in an observed security return series.

Generalization to the return series of a portfolio see Campbell, Lo an Mackinlay (1997, chapter 3).
Bid-Ask Spread

Market makers have 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 $P_a$.

\[ P_a - P_b = \text{bid-ask spread} \] (source of compensation for market makers)

Typically, it’s small (one or two ticks)

*Bid-ask bounce* = bid-ask spread introduces *negative* lag-1 serial correlation in an asset return.
Roll (1984) model (simplified):

\[ P_t = P_t^* + I_t \frac{S}{2}, \]  

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 \( I_t = -1 \) 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

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

The change price process if the fundamental value of the asset does not change is,

\[ \Delta P_i = (I_t - I_{t-1}) \frac{S}{2}. \]
Suppose, \( E(I_t) = 0 \) and \( \text{Var}(I_t) = 1 \)

Then,
\[
\begin{align*}
E(\Delta P_t) &= 0 \\
\text{Var}(\Delta P_t) &= S^2/2 \\
\text{Cov}(\Delta P_t, \Delta P_{t-1}) &= -S^2/4, \\
\text{Cov}(\Delta P_t, \Delta P_{t-j}) &= 0, \quad j > 1 \\
\rho_j(\Delta P_t) &= \begin{cases} 
-0.5 & \text{if } j = 1 \\
0 & \text{if } j > 1
\end{cases}
\end{align*}
\]

Conclusion: The bid-ask spread introduces a negative lag-1 serial correlation in the series of observed price changes. But, it does not introduce any serial correlation beyond lag 1.

Example:
Assuming that the fundamental price is equal to \((P_a + P_b)/2\), \(P_t\) assumes the value \(P_a\) or \(P_b\). Thus, \(\Delta P_t\) is either 0 or \(-S\), if the previous observed price is \(P_a\), and either 0 or \(S\), if the previous observed price is \(P_b\).
Now, assume a more realistic formulation that $P_t^*$ follows a random walk,

$$\Delta P_t^* = P_t^* - P_{t-1}^* = \epsilon_t,$$

$iid$ sequence with mean zero and variance $\sigma^2$ and independent of $\{I_t\}$.

It’s easy to see that, $\Delta P_t = \epsilon_t + (I_t - I_{t-1})S/2$ with the same mean and covariance, but

$$\text{Var}(\Delta P_t) = \sigma^2 + S^2/2$$

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

is reduced, but negative.

Furthermore,

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.
Empirical characteristics

1. Unequally spaced time intervals
2. Discrete-value prices
3. Existence of a daily periodic or diurnal pattern
4. Multiple transactions at the same time
Example: IBM transactions data from November 1, 1990 to January 31, 1991. (63 trading days and 60.328 transactions)

Ignore the price changes between trading days and focusing on the transactions that occurred in the normal trading hours from 9:30 am to 4:00 pm Eastern time.

<table>
<thead>
<tr>
<th>Number (tick)</th>
<th>( \leq -3 )</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>( \geq 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>0.66</td>
<td>1.33</td>
<td>14.53</td>
<td>67.06</td>
<td>14.53</td>
<td>1.27</td>
<td>0.63</td>
</tr>
</tbody>
</table>

The distribution of positive and negative price changes was approximately symmetric, with high frequency in zero and one ticks.

Consider the number of transactions in a 5-minute time interval, \( x_t \)
Figure 5.1. IBM intraday transactions data from 11/01/90 to 1/31/91: (a) the number of transactions in 5-minute time intervals and (b) the sample ACF of the series in part(a).
cycle pattern of the ACF with a periodicity of 78, which is the number of 5-minute intervals in a trading day.

In other words, the number of transactions exhibits a daily pattern.

Now, consider the average number of transactions within 5-minute time intervals over the 63 days. We have 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.
Figure 5.2. Time plot of the average number of transactions in 5-minute time intervals. There are 78 observations, averaging over the 63 trading days from 11/01/90 to 1/31/91 for IBM stock.
Now, consider the classification of price movements: *up* (+), *unchanged* (0) and *down* (-). And the price movements between two consecutive trades: \((i-1)\)th to \(i\)th transaction.

<table>
<thead>
<tr>
<th>((i - 1))th Trade</th>
<th>(i)th Trade</th>
<th>()</th>
<th>| ()</th>
<th>Margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>441</td>
<td>5498</td>
<td>3948</td>
<td>9887</td>
</tr>
<tr>
<td>0</td>
<td>4867</td>
<td>29779</td>
<td>5473</td>
<td>40119</td>
</tr>
<tr>
<td>–</td>
<td>4580</td>
<td>4841</td>
<td>410</td>
<td>9831</td>
</tr>
<tr>
<td><strong>Margin</strong></td>
<td><strong>9888</strong></td>
<td><strong>40118</strong></td>
<td><strong>9831</strong></td>
<td><strong>59837</strong></td>
</tr>
</tbody>
</table>

*The price movements are classified into “up,” “unchanged,” and “down.” The data span is from 11/01/90 to 1/31/91.*

> price reversals in intraday transactions data.
Models for price changes

1) Ordered Probit Model [Hauseman, Lo and MacKinlay (1992)]

Ordered Probit Model

Let $y_i^* \equiv P_{t_i}^* - P_{t_{i-1}}^*$, $P_t^*$ is the virtual price of the asset at time $t$.

Assumes that $y_i^*$ is a continuous random variable and follows a model:

$$ y_i^* = x_i \beta + \epsilon_i $$

where $x_i$ is a $p$-dimensional row vector of explanatory variables at time $t_{i-1}$.

$E(\epsilon_i|x_i) = 0$, $\text{Var}(\epsilon_i|x_i) = \sigma_i^2$, $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

The conditional variance is assumed to be a positive function of the explanatory variable $w_i$.

$$ \sigma_i^2 = g(w_i) $$

In $w_i$, we have the time interval $\Delta t_i = t_i - t_{i-1}$ (time duration) and some conditional heteroscedastic variables.

Furthermore, one also assumes that the conditional distribution of $\epsilon_i$ given $x_i$ and $w_i$ is Gaussian.
Assume that the observed price change $y_i$ may assume $k$ possible values as the number of ticks, in practice. Let $\{s_1, \ldots, s_k\}$ the set of $k$ possible values.

We observe,

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

where $\alpha_j$ are real numbers satisfying $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{k-1} < \alpha_k = \infty$

Under conditional Gaussian distribution, we have

$$P(y_i = s_j | x_i, w_i) = P(\alpha_{j-1} < x_i \beta + \epsilon_i \leq \alpha_j | x_i, w_i)$$

$$= \begin{cases} 
P(x_i \beta + \epsilon_i \leq \alpha_1 | x_i, w_i) & \text{if } j = 1, \\
\frac{\alpha_{j-1} < x_i \beta + \epsilon_i \leq \alpha_j | x_i, w_i)}{\alpha_{k-1} < x_i \beta + \epsilon_i | x_i, w_i)} & \text{if } j = 2, \ldots, k-1, \\
\frac{\alpha_{k-1} < x_i \beta + \epsilon_i | x_i, w_i)} & \text{if } j = k,
\end{cases}$$
Example: Hauseman, Lo and MacKinlay (1992)


Sample means (standard deviation):
\[ y_i = -0.001 \ (0.753) \quad \Delta t_i = 27.21 \ (34.13) \quad \text{bid-ask spread} = 1.947 \ (1.4625) \]
Model:

\[ x_i \beta = \beta_1 \Delta t_i^* + \sum_{v=1}^{3} \beta_{v+1} y_{i-v} + \sum_{v=1}^{3} \beta_{v+4} \text{SP5}_{i-v} + \sum_{v=1}^{3} \beta_{v+7} \text{IBS}_{i-v} \]

\[ + \sum_{v=1}^{3} \beta_{v+10} [T_\lambda(V_{i-v}) \times \text{IBS}_{i-v}] , \]

\[ \sigma_i^2(w_i) = 1.0 + \gamma_1^2 \Delta t_i^* + \gamma_2^2 \text{AB}_{i-1} \]

where,

\[ T_\lambda(V) = (V^\lambda - 1)/\lambda \] is the Box–Cox (1964) transformation of \( V \)

\[ \Delta t_i^* = (t_i - t_{i-1})/100 \]

\( \text{AB}_{i-1} \) 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
$V_{i-v}$ ($v = 1, 2, 3$) 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

$$IBS_{i-v} = \begin{cases} 
1 & \text{if } P_{i-v} > (P_{i-v}^a + P_{i-v}^b)/2, \\
0 & \text{if } P_{i-v} = (P_{i-v}^a + P_{i-v}^b)/2, \\
-1 & \text{if } P_{i-v} < (P_{i-v}^a + P_{i-v}^b)/2,
\end{cases}$$

where $P_{j}^a$ and $P_{j}^b$ are the ask and bid price at time $t_j$. 
Table 5.4. Parameter Estimates of the Ordered Probit Model in Eqs. (5.19) and (5.20) for the 1988 Transaction Data of IBM, Where $t$ Denotes the $t$-Ratio$^a$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\alpha_4$</th>
<th>$\alpha_5$</th>
<th>$\alpha_6$</th>
<th>$\alpha_7$</th>
<th>$\alpha_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>-4.67</td>
<td>-4.16</td>
<td>-3.11</td>
<td>-1.34</td>
<td>1.33</td>
<td>3.13</td>
<td>4.21</td>
<td>4.73</td>
</tr>
<tr>
<td>$t$</td>
<td>-145.7</td>
<td>-157.8</td>
<td>-171.6</td>
<td>-155.5</td>
<td>154.9</td>
<td>167.8</td>
<td>152.2</td>
<td>138.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\gamma_1$</th>
<th>$\gamma_2$</th>
<th>$\beta_1: \Delta t_i^*$</th>
<th>$\beta_2: \gamma_{-1}$</th>
<th>$\beta_3$</th>
<th>$\beta_4$</th>
<th>$\beta_5$</th>
<th>$\beta_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>0.40</td>
<td>0.52</td>
<td>-0.12</td>
<td>-1.01</td>
<td>-0.53</td>
<td>-0.21</td>
<td>1.12</td>
<td>-0.26</td>
</tr>
<tr>
<td>$t$</td>
<td>15.6</td>
<td>71.1</td>
<td>-11.4</td>
<td>-135.6</td>
<td>-85.0</td>
<td>-47.2</td>
<td>54.2</td>
<td>-12.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_7$</th>
<th>$\beta_8$</th>
<th>$\beta_9: \beta_{10}$</th>
<th>$\beta_{11}$</th>
<th>$\beta_{12}$</th>
<th>$\beta_{13}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>0.01</td>
<td>-1.14</td>
<td>-0.37</td>
<td>-0.17</td>
<td>0.12</td>
<td>0.05</td>
</tr>
<tr>
<td>$t$</td>
<td>0.26</td>
<td>-63.6</td>
<td>-21.6</td>
<td>-10.3</td>
<td>47.4</td>
<td>18.6</td>
</tr>
</tbody>
</table>

$^a$Reprinted with permission from Elsevier.
boundary partitions are not equally spaced, but almost symmetric with respect to zero.

the transaction duration, $\Delta t_i$, affects both the conditional mean and conditional variance of $y_i$.

the coefficients of lagged price changes are negative and highly significant, indicating *price reversals*.

the bid-ask spread at time $t_{i-1}$ significantly affects the conditional variance.
A Decomposition Model

Decomposition in 3 components: 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.

\[ y_i \equiv P_{t_i} - P_{t_{i-1}} = A_i D_i S_i \]

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

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

\( S_i \) is the size of the price change in ticks if there is a change at the \( i\)th trade and \( S_i = 0 \) if there is no price change at the \( i\)th trade. When there is a price change, \( S_i \) is a positive integer-valued random variable.
Let $F_i$ be the information set available at the $i$th transaction. We have an ordered decomposition of the price change probability.

$$P(y_i|F_{i-1}) = P(A_i D_i S_i|F_{i-1}) = P(S_i|D_i, A_i, F_{i-1}) P(D_i|A_i, F_{i-1}) P(A_i|F_{i-1})$$

Since $A_i$ is a binary variables,

$$p_i = P(A_i = 1)$$

$$\ln \left( \frac{p_i}{1 - p_i} \right) = x_i \beta \quad \text{or} \quad p_i = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}}$$

where $x_i$ is a finite-dimensional vector consisting of elements of $F_{i-1}$

Let,

$$\delta_i = P(D_i = 1|A_i = 1)$$

$$\ln \left( \frac{\delta_i}{1 - \delta_i} \right) = z_i \gamma \quad \text{or} \quad \delta_i = \frac{e^{z_i \gamma}}{1 + e^{z_i \gamma}}$$

where $z_i$ is a finite-dimensional vector consisting of elements of $F_{i-1}$
$S_i|(D_i, A_i = 1) \sim 1 + \begin{cases} g(\lambda_{u,i}) & \text{if } D_i = 1, A_i = 1, \\ g(\lambda_{d,i}) & \text{if } D_i = -1, A_i = 1, \end{cases}$

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

$$\ln \left( \frac{\lambda_{j,i}}{1 - \lambda_{j,i}} \right) = w_i \theta_j \quad \text{or} \quad \lambda_{j,i} = \frac{e^{w_i \theta_j}}{1 + e^{w_i \theta_j}}, \quad j = u, d,$$

where $w_i$ is again a finite-dimensional explanatory variable in $F_{i-1}$

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

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

Because the price change, if it occurs, is at least 1 tick, we added 1 to the geometric distribution. We take the logistic transformation to ensure that $\lambda_{j,i} \in [0, 1]$
Define 3 categories and $I_t(j), j = 1, 2, 3$ its indicator variables.

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 $1 + g(\lambda_{d,i})$.

So,

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

Thus, it is possible to use the log likelihood function to estimate the parameters $\beta, \gamma, \theta_u, \theta_d$

$$
\ln[P(y_1, \ldots, y_n|F_0)] = \sum_{i=1}^{n} \ln[P(y_i|F_{i-1})]
$$
Example: intraday transactions of IBM stock from November 1, 1990 to January 31, 1991, with 63 trading days and 59.838 intraday transactions in the normal trading hours.

\[
\ln \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 A_{i-1}, \quad \ln \left( \frac{\delta_i}{1 - \delta_i} \right) = \gamma_0 + \gamma_1 D_{i-1},
\]

\[
\ln \left( \frac{\lambda_{u,i}}{1 - \lambda_{u,i}} \right) = \theta_{u,0} + \theta_{u,1} S_{i-1}, \quad \ln \left( \frac{\lambda_{d,i}}{1 - \lambda_{d,i}} \right) = \theta_{d,0} + \theta_{d,1} S_{i-1}.
\]

Table 5.5. Parameter Estimates of the ADS Model in Eq. (5.30) for IBM Intraday Transactions from 11/01/90 to 1/31/91

<table>
<thead>
<tr>
<th>Parameter</th>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
<th>( \gamma_0 )</th>
<th>( \gamma_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>-1.057</td>
<td>0.962</td>
<td>-0.067</td>
<td>-2.307</td>
</tr>
<tr>
<td>Standard error</td>
<td>0.104</td>
<td>0.044</td>
<td>0.023</td>
<td>0.056</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>( \theta_{u,0} )</th>
<th>( \theta_{u,1} )</th>
<th>( \theta_{d,0} )</th>
<th>( \theta_{d,1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>2.235</td>
<td>-0.670</td>
<td>2.085</td>
<td>-0.509</td>
</tr>
<tr>
<td>Standard error</td>
<td>0.029</td>
<td>0.050</td>
<td>0.187</td>
<td>0.139</td>
</tr>
</tbody>
</table>
Some features:

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

   \[ P(A_i = 1|A_{i-1} = 0) = 0.258, \quad P(A_i = 1|A_{i-1} = 1) = 0.476. \]

   \( \Rightarrow \) the price change may occur in clusters and, as expected, most transactions are without price change: \( [P(A_i = 0 \mid A_{i-1} = 0) = 0.742] \)

2. The direction of price change is governed by

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

   \( \Rightarrow \) if \( A_{i-1} = 0 \), then the chances for a price increase or decrease at the \( i \)th trade are about even

   \( \Rightarrow \) the probability of consecutive price increase or decrease are very low. Consequently, 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

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

So, the probability of a price increase by one tick is 0.827 at the \( i \)th 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 \( i \)th trade.

**Note:**
A difference between the ADS and ordered probit models is that the former does not require any truncation or grouping in the size of a price change.
Duration Models

- Engle and Russel (1998): ACD model - autoregressive conditional duration
- Concepts similar to the ARCH models for volatility
- For heavily traded stocks
- Duration is a positive number, so we need no-gaussian errors
- Used distributions: exponential (EACD), Weibull (WACD), generalized Gamma (GACD)

First, it’s necessary to adjust the time duration:

$$\Delta t_i^* = \Delta t_i / f(t_i)$$

where, $f(t_i)$ is a deterministic function consisting of the cyclical component of $\Delta t_i$. 
For example:

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

where

$$f_1(t_i) = - \left( \frac{t_i - 43200}{14400} \right)^2,$$

$$f_2(t_i) = - \left( \frac{t_i - 48300}{9300} \right)^2,$$

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

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

$f_5(t_i)$ and $f_6(t_i)$ are indicator variables for the first and second 5 minutes of market opening (i.e., $f_5(.) = 1$ 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 5.5 shows the plot of $f_i(.)$ 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.
Figure 5.5. Quadratic functions used to remove the deterministic component of IBM intraday trading durations: (a)–(d) are the functions $f_1(.)$ to $f_4(.)$ of Eq. (5.32), respectively.
For IBM data, the coefficients $\beta_j$ 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:

$$\ln(\widehat{\Delta t_i}) = 2.555 + 0.159 f_1(t_i) + 0.270 f_2(t_i) + 0.384 f_3(t_i)$$

$$+ 0.061 f_4(t_i) - 0.611 f_5(t_i) - 0.157 f_6(t_i) + 0.073 f_7(t_i).$$
Figure 5.6. IBM transactions data from 11/01/90 to 1/31/91: (a) the average durations in 5-minute time intervals and (b) the average durations in 5-minute time intervals after adjusting for the deterministic component.
The ACD Model

Let \( x_i = \Delta t_i^* \)

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

\[
x_i = \psi_i \epsilon_i,
\]

(5.33)

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

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

(5.34)

We have a ACD(r,s) model
So,
\(\varepsilon_i\) exponential \(\Rightarrow\) EACD(r,s) model
\(\varepsilon_i\) Weibull \(\Rightarrow\) WACD(r,s) model
\(\varepsilon_i\) Gamma \(\Rightarrow\) GACD(r,s) model

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

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

which is in the form of an ARMA process with non-Gaussian innovations.

Assuming weak stationarity,

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

Therefore, we assume \(\omega > 0\) and \(1 > \sum_j (\gamma_j + \omega_j)\) because the expected duration is positive.
EACD(1,1) Model

\[ x_i = \psi_i \epsilon_i, \quad \psi_i = \omega + \gamma_1 x_{i-1} + \omega_1 \psi_{i-1}, \]

Where \( \epsilon_i \) follows the standard exponential distribution.
\( E(\epsilon_i) = 1, \ Var(\epsilon_i) = 1, \) and \( E(\epsilon_i^2) = \text{Var}(x_i) + [E(x_i)]^2 = 2. \)

Under weak stationarity of \( x_i, \)
\[ E(x_i) = E[E(\psi_i \epsilon_i | F_{i-1})] = E(\psi_i), \quad E(\psi_i) = \omega + \gamma_1 E(x_{i-1}) + \omega_1 E(\psi_{i-1}). \]

Under weak stationarity, \( E(\psi_i) = E(\psi_{i-1}) \)

\[ \mu_x \equiv E(x_i) = E(\psi_i) = \frac{\omega}{1 - \gamma_1 - \omega_1} \]

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

And
\[ E(\psi_i^2) = \mu_x^2 \times \frac{1 - (\gamma_1 + \omega_1)^2}{1 - 2\gamma_1^2 - \omega_1^2 - 2\gamma_1\omega_1}. \]

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

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

**Simulation**

500 observations from the ACD(1,1) model

\[ x_i = \psi_i \epsilon_i, \quad \psi_i = 0.3 + 0.2x_{i-1} + 0.7\psi_{i-1} \]

Using:

- Standardized Weibull distribution with parameter \( \alpha = 1.5 \)
- Standardized generalized gamma distribution with parameters \( k = 1.5 \) and \( \alpha = 1.5 \).
Figure 5.7. A simulated WACD(1,1) series in Eq. (5.40): (a) the original series and (b) the standardized series after estimation. There are 500 observations.
Figure 5.8. A simulated GACD(1,1) series in Eq. (5.40): (a) the original series and (b) the standardized series after estimation. There are 500 observations.
Figure 5.9. Histograms of simulated duration processes with 500 observations: (a) WACD(1,1) model and (b) GACD(1,1) model.
Figure 5.10. The sample autocorrelation function of a simulated WACD(1,1) series with 500 observations: (a) the original series and (b) the standardized residual series.
Figure 5.11. The sample autocorrelation function of a simulated GACD(1,1) series with 500 observations: (a) the original series and (b) the standardized residual series.
**Estimation**

ACD(r,s) model

The likelihood function of the durations $x_1, \cdots, x_T$

$$f(x_T | \theta) = \left[ \prod_{i=i_0+1}^{T} f(x_i | F_{i-1}, \theta) \right] \times f(x_{i_0} | \theta),$$

where $\theta$ denotes the vector of model parameters, and $T$ is the sample size.

Use a conditional likelihood method, ignoring $f(x_{i_0} | \theta)$

For WACD:

$$\ell(x | \theta, x_{i_0}) = \sum_{i=i_0+1}^{T} \alpha \ln \left[ \Gamma \left( 1 + \frac{1}{\alpha} \right) \right] + \ln \left( \frac{\alpha}{x_i} \right) + \alpha \ln \left( \frac{x_i}{\psi_i} \right) - \left( \frac{\Gamma(1 + 1/\alpha) x_i}{\psi_i} \right)^{\alpha},$$

When $\alpha = 1$, the (conditional) log likelihood function reduces to that of an EACD(r,s) model.
For GACD(r,s) model:

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

See Example 5.3 and 5.4 (pp. 233-236)

**Note**: Estimation of EACD models can be carried out by using programs for ARCH models with some minor modification.
Bivariate models for price change and duration

Here, we consider jointly the process of price change and the associated duration.

**Problem:** many intraday transactions of a stock result in no price change and they are highly relevant to trading intensity, but don’t contain direct information on price movement.

- 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

  Note: This choice can reduce the sample size dramatically.

Example: IBM stock on November 21, 1990 (726 transactions intraday trades, but only 195 with a price change).
Figure 5.14. Time plots of the intraday transaction prices of IBM stock on November 21, 1990: (a) all transactions and (b) transactions that resulted in a price change.
Let

\[ t_i : \text{the calendar time of the } i \text{th change of an asset, measured in seconds.} \]
\[ \Delta t_i = t_i - t_{i-1} : \text{time duration between price changes} \]
\[ P_{t_i} : \text{transaction price when the } i \text{th price change occurred} \]
\[ N_i : \text{number of trades in the time interval } (t_i - t_{i-1}) \text{ that result in no price change} \]
\[ D_i : \text{the direction of the } i \text{th price change } [D_i = 1, \text{price goes up, and } D_i = -1, \text{price comes down}] \]
\[ S_i : \text{size of the } i \text{th price change measured in ticks} \]

\[ P_{t_i} = P_{t_{i-1}} + D_i S_i, \]

Each transaction consist of \{\Delta t_i, N_i, D_i, S_i\} for the \( i \)th price change.

**PCD Model:** decomposes the joint distribution of \{\Delta t_i, N_i, D_i, S_i\} given \( F_{i-1} \)

\[
f(\Delta t_i, N_i, D_i, S_i|F_{i-1}) = f(S_i|D_i, N_i, \Delta t_i, F_{i-1}) f(D_i|N_i, \Delta t_i, F_{i-1}) f(N_i|\Delta t_i, F_{i-1}) f(\Delta t_i|F_{i-1})
\]
McCulloch and Tsay (2000):

• For time duration between price changes:

\[ \ln(\Delta t_i) = \beta_0 + \beta_1 \ln(\Delta t_{i-1}) + \beta_2 S_{i-1} + \sigma \varepsilon_i \]

a multiple linear regression model with lagged variables, where \( \sigma \) is a positive number and \( \{\varepsilon_i\} \) is a sequence of iid \( N(0,1) \) random variables.

• For \( N_i \), it’s necessary to partitioned into two parts, because empirical data suggest a concentration of \( N_i \) at 0.

A logit model: 

\[ p(N_i = 0 \mid \Delta t, F_{i-1}) = \frac{\exp(\alpha_0 + \alpha_1 \ln(\Delta t_i))}{1 + \exp(\alpha_0 + \alpha_1 \ln(\Delta t_i))} \]

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

g(\lambda) denotes a geometric distribution with parameter \( \lambda \).
• For $D_i$, direction:

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

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

$$\mu_i = \omega_0 + \omega_1 D_{i-1} + \omega_2 \ln(\Delta t_i),$$

$$\ln(\sigma_i) = \beta \begin{vmatrix} D_{i-1} \\
\vdots \\
D_{i-4} \end{vmatrix} = \beta |D_{i-1} + D_{i-2} + D_{i-3} + D_{i-4}|.$$

$D_i$ is governed by the sign of a normal random variable with mean $\mu_i$ and variance $\sigma_i^2$.

*Price reversal* feature $\Rightarrow \varpi_1$ negative

But the variance equation allows for a local trend by increasing the uncertainty in the direction of price movement when the past data showed evidence of a local trend.
For size of a price change:

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

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

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

The model can be estimated jointly by the maximum likelihood method

See example 5.5 (pp. 240)