Financial Data Analysis

Market Microstructure and High-frequency data

Roman Liesenfeld
(University of Kiel)

October 9, 2008
## Contents

4 Market Microstructure and High-Frequency data 3

4.1 Bid-Ask Spread .......................... 4

4.2 Modelling transaction price changes .................. 9

4.3 Duration models .............................. 21
4 Market Microstructure and High-Frequency data

In this chapter, we consider statistical models to capture and model specific characteristics of high-frequency (HF) data in finance.

- HF-data are observations taken at fine time intervals.

Nowadays, in finance data at the finest intervals are available:

\[
{\text{trade-by-trade or transaction-by-transaction data}},
\]

with time intervals between successive observations often less than a few seconds.

- HF-data allow to study

the process and outcomes (like transaction volume, transaction price) of trading assets under explicit trading rules, i.e., under a specific market microstructure.

Example:

HF-data allow to compare the efficiency of different trading system, like

- the open out-cry system of, e.g., the NYSE.
- or the computer-trading system as, e.g.
  the German XETRA (eXchange Electronic TRAding) system.

- However, HF-data have some very specific features which should not be ignored and which do not appear in lower frequencies (e.g., in daily or weakly data).

- In the following, we study some of these features of HF-data.

In particular, we discuss

- bid–ask spread

- discreteness of price–changes (being multiples of the **tick size**)

- duration models (for the time-varying intervals between transactions).
4.1 Bid–Ask Spread

- On financial market we typically have a

  - bid-price \( (P_b) \) : Sale price of a security
    (i.e., the price I receive if I sell)
  - ask-price \( (P_a) \) : Purchase price
    (i.e., the price I have to pay if I buy)
  - bid–ask spread : \( S = P_a - P_b > 0 \)

- Bid– and ask-prices are typically set by “market makers”:

  - they stand ready to buy or to sell whenever someone wishes to sell or to buy.
  - Thus, he/she provides market liquidity.
  - As a compensation for providing liquidity, he/she is granted by the exchange monopoly
    rights to post different prices for purchases and sales.

- The existence of bid–ask spread has important consequences in the time series properties
  of asset return:

  The spread introduces negative lag–1 autocorrelation in the returns:
  “bid–ask–bounce”.

- To illustrate this autocorrelation, consider the following simple model form Roll (1984):

**Roll-model**

- The observed market price is assumed to satisfy

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

- $P_t$ = market price
- $P_t^*$ = fundamental value of the asset
- $I_t = \begin{cases} 1, & \text{if transaction is buyer initiated} \\ -1, & \text{if transaction is seller initiated} \end{cases}$ (order-type indicator)
- $S = P_a - P_b$; bid–ask spread

Thus, the price for a buyer–initiated transaction is larger than for a seller–initiated transaction.

- Now, assume that

  $I_t \sim \text{iid} \left\{ \begin{array}{ll} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2, \end{array} \right.$

so that

- $E[I_t] = 0; \ var[I_t] = E[I_t^2] = 1.$

- Furthermore, assume for a moment that there is no change in the fundamental value ($P_t^*$).

Then, the observed price change is

$$\Delta P_t = \Delta P_t^* + (I_t - I_{t-1}) \frac{S}{2} = (I_t - I_{t-1}) \frac{S}{2},$$

and we obtain:

- $E[\Delta P_t] = 0$
- $\var[\Delta P_t] = \frac{S^2}{4} (\var[I_t] + \var[I_{t-1}]) = \frac{S^2}{2}$
- $\cov(\Delta P_t, \Delta P_{t-1}) = E[\Delta P_t \cdot \Delta P_{t-1}]
  = E \left[ \frac{S^2}{4} (I_t - I_{t-1}) (I_{t-1} - I_{t-2}) \right]
  = \frac{S^2}{4} E[ -I_{t-1}^2]
  = \frac{-S^2}{4}$
- $\cov(\Delta P_t, \Delta P_{t-j}) = 0 \ \forall \ j > 1.$
• Therefore, the ACF of $\Delta P_t$ is

$$
\rho_j = \begin{cases} 
-0.5 & \text{if } j = 1 \\
0 & \text{if } j > 1.
\end{cases}
$$

Thus, the spread introduces a negative lag-1 serial correlation in observed price changes.

• The intuition is clear: Assume that

$$
P^*_t = (P_a + P_b) / 2,
$$

so that the transaction price is

$$
P_t = \frac{(P_a + P_b)}{2} + I_t \frac{(P_a - P_b)}{2} = \begin{cases} 
P_a & \text{if } I_t = 1 \text{ (buy)} \\
P_b & \text{if } I_t = -1 \text{ (sell)}
\end{cases}.
$$

Hence, as random buys and sells arrive at the market prices bounce back and forth between $P_a$ and $P_b$, creating spurious volatility and serial correlation even if $P^*_t$ is unchanged (see Figure 1)

![Bid–ask bounce](image)

Figure 1: Bid–ask bounce

- A more realistic specification for the fundamental value than in this simple Roll model is a RW:

$$
P^*_t = P^*_t + \varepsilon_t, \quad \varepsilon_t \sim \text{iid } (0, \sigma^2), \quad \varepsilon_t \perp \!\!\!\perp I_t.
$$
In this case

\[ E[\Delta P_t] = 0 \quad \text{and} \quad \text{var}[\Delta P_t] = \sigma^2 + \frac{S^2}{2}, \]

and the autocovariance remains unchanged

\[ \text{cov}(\Delta P_t, \Delta P_{t-j}) = \begin{cases} 
-\frac{sS^2}{4}, & \text{if } j = 1 \\
0, & \text{if } j > 1
\end{cases}. \]

Therefore,

\[ \rho_j = \begin{cases} 
-\frac{sS^2}{4 \sigma^2 + \gamma^2} (\leq 0), & \text{if } j = 1 \\
0, & \text{if } j > 1
\end{cases}. \]

**Empirical evidence of the bid–ask bounce**

- Figure 2 shows the IBM transaction price changes from November 1, 1990 to January 31, 1991. There are 63 trading days and 60,328 transactions.
Figure 2: *Time plot of transaction price changes in consecutive trades.*
- The following table provides a two-way classification of two consecutive IBM price movements, where

\[ \Delta P_t = \begin{cases} 
+ , & \text{if } \Delta P_t > 0 \\
0 , & \text{if } \Delta P_t = 0 \\
- , & \text{if } \Delta P_t < 0 
\end{cases} \]

<table>
<thead>
<tr>
<th>(t - 1)th trade</th>
<th>tth trade</th>
<th>+</th>
<th>0</th>
<th>−</th>
<th>Margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>456</td>
<td>5562</td>
<td>3991</td>
<td>10009</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>4913</td>
<td>29914</td>
<td>5539</td>
<td>40366</td>
<td></td>
</tr>
<tr>
<td>−</td>
<td>4640</td>
<td>4889</td>
<td>422</td>
<td>9951</td>
<td></td>
</tr>
<tr>
<td>Margin</td>
<td>10009</td>
<td>40365</td>
<td>9952</td>
<td>60326</td>
<td></td>
</tr>
</tbody>
</table>

The data show:

- a high tendency for prices to remain unchanged;

- the probability of moving up (+) and down (−) are about the same;

- the frequencies of price reversals (see, frequencies for (+, −) and (−, +) are larger than frequencies of consecutive increases (+, +) or decreases (−, −).

This is a clear demonstration of the bid–ask bounce. This is confirmed by a lag-1 autocorrelation of −0.408 for \( \Delta P_t \), which is highly significant for a sample of size \( T = 60,327 \).

### 4.2 Modelling transaction price changes

- A further important feature of transaction data is the

  - discrete-valued prices: they occur in multiples of a tick size;

  - at the NYSE before 1997, e.g., the tick size was $1/8. (Since 2001, the NYSE trades in decimals.)

- This discreteness is illustrated in Figure 3, which shows a section of the time series of the IBM transaction price changes from November 1, 1990 to January 31, 1991.
Figure 3: Section of the time series of the IBM transaction price changes from November 1, 1990 to January 31, 1991.
- The following table gives the frequencies in multiples of tick size for IBM stock:

<table>
<thead>
<tr>
<th>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>

(To2, p. 182, table 5.1)

It shows that

- 2/3 of transactions were without price changes;
- about 99% of changes are between $-3$ and 3 ticks.

- The discreteness and concentration of zero price changes make it difficult to use \textit{stochastic models} for \textit{continuous variables} as standard

  - linear regression models
  - or ARMA-type models.

- To address this problem, Hausman, Lo and MacKinlay (1992)$^1$ propose an ordered probit model that have the advantage of employing \textit{explanatory variables} to study intraday price moments.

\textbf{Ordered Probit Model}

- It has its origin in microeconometric studies of \textit{discrete dependent variables}:

  that take on a \textit{finite number of values} possessing a \textit{natural ordering}.

- Let


\( y_i^* : \) unobservable continuous price change, interpreted as the movement of the true virtual price \( P_t^* \):

\[
y_i^* = P_i - P_{t_{i-1}} \quad i : 1 \rightarrow n, \quad y_i^* \in \mathbb{R}
\]

sampled at times

\[
t_0, t_1, \ldots, t_{i-1}, t_i, \ldots, t_n,
\]

where

\( t_i : \) clock time of the \( i \)th transaction.

- The Ordered Probit model assumes a linear regression model for \( y_i^* \):

\[
y_i^* = x_i' \beta + \varepsilon_i,
\]

where \( \varepsilon_i \) is Gaussian with \(^2\)

\[
E[\varepsilon_i | x_i] = 0, \quad E[\varepsilon_i^2 | x_i] = \sigma_i^2, \quad \text{cov}(\varepsilon_j, \varepsilon_j) = 0 \forall i \neq j
\]

and

\[
x_i = (x_{i1}, \ldots, x_{ik})' : \text{explanatory variables available at time } t_{i-1}.
\]

- The conditional variance is assumed to be

\[
\sigma_i^2 = g(w_i' \gamma) : \text{positive function, e.g., } \sigma_i^2 = (w_i' \gamma)^2.
\]

with

\[
w_i : \text{vector of explanatory variables, e.g., } t_i - t_{i-1} : \text{(time between transaction)}
\]

- Now, let

\[
y_i : \text{observed price change associated with the } i\text{th transaction.}
\]

In the ordered probit \( y_i \) may assume \( k \) possible discrete values:

\[
y_i \in \{s_1, \ldots, s_k\},
\]

\(^2\)The normality assumption is imposed purely for convenience. A logistic distribution, e.g., generally give similar results in practice.
e.g., based upon the following 7 categories:

\[
\begin{align*}
    s_1 &= -3 \iff y_i \in (-\infty; -3] \\
    s_2 &= -2 \iff y_i = -2 \\
    \vdots \\
    s_6 &= 2 \iff y_i = 2 \\
    s_7 &= 3 \iff y_i = [3; +\infty).
\end{align*}
\]

Note, that this combines several categories of \( y_i \) into \( s_1 \) and \( s_7 \), while \( s_2, \ldots, s_6 \) represent a single value.

- The heart of the ordered probit is the assumption about the relationship between \( y_i \) and \( y_i^* \). In particular, it is assumed that

\[
y_i = \begin{cases} 
  s_1, & \text{if } \alpha_0 < y_i^* \leq \alpha_1 \\
  s_2, & \text{if } \alpha_1 < y_i^* \leq \alpha_2 \\
  \vdots \\
  s_k, & \text{if } \alpha_{k-1} < y_i^* \leq \alpha_k
\end{cases},
\]

where the unknown partition boundaries (tresholds) \( \{\alpha_j\} \) satisfy:

\[-\infty = \alpha_0 < \alpha_1 \cdots < \alpha_{k-1} < \alpha_k = \infty.\]

- Observe the trade-off when selecting the number of categories \( k \):

the more categories \( k \), the higher the price resolution, but also the more parameters to be estimated.

- Under the Gaussian assumption for \( \varepsilon_i \), we obtain the following conditional pdf for \( y_i \mid (x_j, w_j) \):
\[ P(\frac{yi}{y_i} = s_j \mid x_i, w_i) = P(\alpha_{j-1} < \frac{x_i' \beta + \varepsilon_i}{y_i} \leq \alpha_j \mid x_i, w_i) \]

\[ = \begin{cases} 
P(\alpha_{j-1} < \frac{x_i' \beta + \varepsilon_i}{y_i} \leq \frac{x_i' \beta + \varepsilon_i}{y_i} \mid x_i, w_i), & \text{if } j = 1 \\
P(\alpha_{j-1} < \frac{x_i' \beta + \varepsilon_i}{y_i} < \alpha_j \mid x_i, w_i), & \text{if } j = 2, \ldots, k-1 \\
P(\alpha_{k-1} < \frac{x_i' \beta + \varepsilon_i}{y_i} \mid x_i, w_i), & \text{if } j = k \\
\end{cases} \]

\[ = \begin{cases} 
\Phi\left(\frac{\alpha_{j-1}-x_i' \beta}{\sigma_i}\right), & \text{if } j = 1 \\
\Phi\left(\frac{\alpha_{j-1}-x_i' \beta}{\sigma_i}\right) - \Phi\left(\frac{\alpha_{j-1}-x_i' \beta}{\sigma_i}\right), & \text{if } j = 2, \ldots, k-1 \\
1 - \Phi\left(\frac{\alpha_{k-1}-x_i' \beta}{\sigma_i}\right), & \text{if } j = k \\
\end{cases} \]

where

\( \Phi(\cdot) : \text{ cdf of a N}(0, 1)\)-distribution.

- The implication of the structure imposed on the relationship between observed \( (y_i) \) and latent \( (y^*_i) \) price moment is illustrated in Figure 4.
Figure 4: Probabilities in the ordered probit model
- Interpretation:

- Given the partition boundaries, a higher conditional mean $x_i/\beta$ implies a higher probability of observing a more extreme positive state. It shifts the pdf to the right.

- Hence, the regressors allow us to separate the effect of economic factors that influence the likelihood of the state versus the other.

- However, the marginal effects of regressors on probabilities are not equal to the coefficients in $\beta$ (in contrast to the standard linear regression model).

To illustrate this, consider a 3-category case with one regressor with:

$$
\pi_1 = P(y_i = s_1 \mid \cdot) = \Phi\left(\frac{\alpha_1 - x_i/\beta}{\sigma_i}\right)
$$

$$
\pi_2 = P(y_i = s_2 \mid \cdot) = \Phi\left(\frac{\alpha_2 - x_i/\beta}{\sigma_i}\right) - \Phi\left(\frac{\alpha_1 - x_i/\beta}{\sigma_i}\right)
$$

$$
\pi_3 = P(y_i = s_3 \mid \cdot) = 1 - \Phi\left(\frac{\alpha_2 - x_i/\beta}{\sigma_i}\right).
$$

The marginal effects of changes in the regressor are:

$$
\frac{\partial \pi_1}{\partial x_i} = -\phi\left(\frac{\alpha_1 - x_i/\beta}{\sigma_i}\right) \frac{\beta}{\sigma_i}
$$

$$
\frac{\partial \pi_2}{\partial x_i} = \left[ \phi\left(\frac{\alpha_1 - x_i/\beta}{\sigma_i}\right) - \phi\left(\frac{\alpha_2 - x_i/\beta}{\sigma_i}\right) \right] \frac{\beta}{\sigma_i}
$$

$$
\frac{\partial \pi_3}{\partial x_i} = \phi\left(\frac{\alpha_2 - x_i/\beta}{\sigma_i}\right) \frac{\beta}{\sigma_i}
$$

where $\phi(\cdot) =$ pdf of a $N(0, 1)$-distribution.

- Figure 5 illustrates the marginal effects for $\beta > 0$:
  increasing $x_i$, implies c.p. a shift of the pdf to the right, lowering $\pi_1$ and increasing $\pi_3$. The effect on $\pi_2$ is typically ambiguous. It depends on the form of the two pdfs.

The upshot is that we must be careful in interpreting the coefficients.
Figure 5: Effects of change in $x_i$ on probabilities ($\beta > 0$)
\textbf{ML-estimation}

- The ordered probit contains parameters
  \[ \theta = (\beta, \alpha_1, \ldots, \alpha_{k-1}, \gamma)'; \]

  they can be estimated by ML or by MCMC.

- Let
  \[ \delta_{ij} = \begin{cases} 1, & \text{if } y_i = s_j, \\ 0, & \text{else} \end{cases}, \quad j : 1 \rightarrow k, \]

  indicating the category of the \( i \)th price change \( y_i \).

Then, the log-likelihood for a sample of \( n \) price changes \( Y = (y_1, \ldots, y_n)' \) is given by:

\[
\ell(\beta, \gamma, \alpha_1, \ldots, \alpha_k) = \sum_{i=1}^{n} \left\{ \delta_{i1} \ln \Phi \left( \frac{\alpha_1 - x_i'\beta}{\sigma_i} \right) \right. \\
+ \sum_{j=2}^{k-1} \delta_{ij} \ln \left[ \Phi \left( \frac{\alpha_j - x_i'\beta}{\sigma_i} \right) - \Phi \left( \frac{\alpha_{j-1} - x_i'\beta}{\sigma_i} \right) \right] \\
+ \delta_{ik} \ln \left[ 1 - \Phi \left( \frac{\alpha_{k-1} - x_i'\beta}{\sigma_i} \right) \right] \left. \right\}. 
\]

- To produce ML-estimates, \( \ell(\cdot) \) must be maximized using numerical iterative procedures, since there exists no closed-form solution for ML-estimates of an ordered probit.

- Note, that we typically have an \underline{identification problem} when estimating ordered probits:

  Based on data, we cannot identify all parameters, e.g., doubling the \( \alpha \)'s, \( \beta \)'s and \( \sigma_i \)'s leaves the likelihood unchanged.

  A typical \underline{identification assumption} is to set the constant term in \( \sigma_i^2 = w_i'\gamma : \gamma_0 \equiv 1. \)

18
- Hausman, Lo and MacKinlay (1992) applied the ordered probit to 1988 transaction data of several stocks.

Here, I present their results for IBM with 206,714 trades.

- They used nine categories for price changes; the conditional mean of the latent model is

\[ x'_i \beta = \beta_1 \Delta t_i / 100 + \sum_{\tau=1}^{3} \beta_{i-\tau} y_{i-\tau} + \cdots, \]

and the conditional variance

\[ \sigma_i^2 = 1.0 + \gamma_1^2 \Delta t_i / 100 + \gamma_2^2 A_{B_i - 1} \]

where

\[ \Delta t_i = t_i - t_{i-1} : \text{ time elapsed between trade } i - 1 \text{ and } i, \]
\[ \text{measured in seconds (duration)} \]

\[ A_{B_i - 1} : \text{ bid–ask spread prevailing at time } t_{i-1} \text{ in ticks} \]

\[ y_{i-\tau} : 3 \text{ lags } (\tau = 1, 2, 3) \text{ of the observed price change.} \]

- The parameter estimates and their t-ratios are given in the following table:

<table>
<thead>
<tr>
<th>(a) Boundary partitions of the probit model</th>
<th>Par.</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>Est.</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>
<td></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>
<td></td>
</tr>
</tbody>
</table>

(b) Equation parameters of the probit model

<table>
<thead>
<tr>
<th>Par.</th>
<th>( \gamma_1 )</th>
<th>( \gamma_2 )</th>
<th>( \beta_1 : \Delta t_i )</th>
<th>( \beta_2 : y_{i-1} )</th>
<th>( \beta_3 )</th>
<th>( \beta_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Est.</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>
</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>
</tr>
</tbody>
</table>

(see, TO2, p. 190)

The results show:
• All $t$-ratios are large, indicating that the estimates are highly significant.

• The boundary partitions are not equally spaced, but are almost symmetric around zero.

• The duration $\Delta t_i$ affects both the conditional mean and variance of price changes.

• The coefficient of lagged price changes $\beta_2 - \beta_4$ are negative, indicating price reversals.

• The bid–ask spread at time $t_{i-1}$ has a positive effect on the conditional variance.

- Hausman, Lo and MacKinlay (1992) used this specification to examine three specific aspects of transactions data:

  (1) Does trade size affects price changes?
      Accordingly, measures of trade size are included in the mean equation.

  (2) Does price discreteness matter?
      In particular, they compare the ordered probit with a simple linear regression model.

  (3) Does the particular sequence of trades affect the conditional distribution of price changes?
4.3 Duration models

(to be added, see TO2, chap. 5.5.)