Analysis of High-Frequency Financial Data & Market Microstructure

Market microstructure:

Why is it important?

1. Important in market design & operation, e.g. to compare different markets (NYSE vs NASDAQ)

2. To study price discovery, liquidity, volatility, etc.

3. To understand costs of trading

4. Important in learning the consequences of institutional arrangements on observed processes, e.g.

Nonsynchronous trading

Bid-ask bounce

Impact of changes in tick size, after-hour trading, etc.

Nonsynchronous trading:

Key implication: may induce serial correlations even when the underlying returns are iid.

Setup: log returns $\{r_t\}$ are iid (μ, σ^2)

For each time index t, P(no trade)$= \pi$.

Cannot observe r_t if there is no trade.

What is the observed log return series r_t^o ?

We consider a simple model proposed by Lo and MacKinlay (1990).

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

When there is a trade at time index t, we define r_t^0 as the cumulative return from the previous trade. Let k_t be the largest non-negative integer such that no trade occurred in the periods $t-k_t, t-k_t+1, \cdots, t-1$. Then

$$r_t^0 = r_t + r_{t-1} + \cdots + r_{t-k_t}.$$

The relationship between r_t and r_t^0 is as follows:

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

Some basic properties:

$$\begin{aligned} & Er_t^0 = \mu, \\ & E(r_t^0)^2 = \sigma^2 + \mu^2[\frac{2}{1-\pi} - 1], \\ & Var(r_t^0) = \sigma^2 + \mu^2[\frac{2}{1-\pi} - 1] - \mu, \\ & E(r_t^0 r_{t-1}^0) = (1-\pi)\mu^2, \\ & Cov(r_t^0, r_{t-1}^0) = -\pi\mu^2, \\ & \rho_1(r_t^0) = \frac{-(1-\pi)\pi\mu^2}{(1-\pi)\pi\sigma^2 + 2\pi\mu^2}, \\ & Cov(r_t^0, r_{t-j}^0) = -\pi^j\mu^2, \ j \geq 1. \end{aligned}$$

Bid-ask bounce

In some stock exchange, market makers provide market liquidity by standing ready to buy or sell whenever the public wishes to buy or sell.

They buy at the bid price P_b and sell at a higher ask price P_a. The difference $P_a - P_b$ is called the $bid-ask\ spread$, which is the compensation for the market makers.

Assume the the fundamental value of an asset is P_t^* as time t. Let P_t be the observed market price of this asset at time t.

Consider the simple model of Roll (1984):

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

$S = P_a - P_b$, I_t is i.i.d random variable with $P(I_t = 1) = P(I_t = -1) = 0.5$.

I_t can be interpreted as an order-type indicator (binary variable).

If there is no change in P_t^*, then

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

$E(I_t) = 0$ and $Var(I_t) = 1$ for all t.

$E(\Delta P_t) = 0$,

$Var(\Delta P_t) = S^2/2$,

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

$Cov(\Delta P_t, \Delta P_{t-j}) = 0,\ j > 1$

The ACF of ΔP_t is

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

Thus, the bid-ask spread introduces a negative lag-1 serial correlation in the series of the observed price changes. It is referred to the $bid - ask\ bounce$ in the finance literature.

If we assume that P_t^* follows a random walk, then $\Delta P_t^* = \epsilon_t$, where ϵ is i.i.d. with mean 0 and variance σ^2.

$E(\Delta P_t) = 0$,

$Var(\Delta P_t) = \sigma^2 + S^2/2$,

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

$Cov(\Delta P_t, \Delta P_{t-j}) = 0,\ j > 1$

The ACF of ΔP_t is

$$\rho_j(\Delta P_t) = \begin{cases} -\frac{S^2/4}{S^2/2+\sigma^2} & \text{if } \ j = 1 \\ 0 & \text{if } \ j > 1 \end{cases}$$

The effect of bid-ask spread continues to exist in the portfolio returns and in multivariate financial time series.

High-Frequency Financial Data

Observations taken with time intervals 24 hours or less

Some example:

1. Transaction (or tick-by-tick) data

2. 5-minute returns in Hong Kong Exchange Market

3. 1-minute returns on index futures.

Some Basic Features of the Data:

1. Irregular time intervals

2. Leptokurtic or Heavy tails

3. Discrete values, e.g. price in multiples of tick size

4. Large sample size

5. Multi-dimensional variables, e.g. price, volume, quotes, etc.

An Example:

IBM stock transaction data from 11/01/1990 to 1/31/1991

Source: Trades, Orders Reports and Quotes (TORQ)

Trading days: 63

Sample size: 60,328

Intraday trades: 60,265.

Data available: bid, ask, transaction prices, volume, time, etc.

Zero durations: 6531 (about 11%).

Frequencies of price change

$No.(tick)$	≤ -3	-2	-1	0	1	2	≥ 3
$Percentage$	0.66	1.33	14.53	67.06	14.53	1.27	0.63

Number of trades in 5-minute intervals.

See Figures 5.1 and 5.2 on page 214 of the text.

Econometric models

1. Time Duration and Duration Models

2. Nonlinearity in Time Durations

3. A Model for Price Change and Duration

4. Models for bid and ask quotes

Statistical tools and methods useful in analyzing HF financial data

Data quality (need some cleaning)

Trading hours (HK-market: 4 hours, US-market: 6.5 hours, but ...)

Time stamp vs transaction time

Missing values

Order types (market or limit orders)

Important statistical issues:

1. Stationarity

2. Nonlinearity

3. Structural breaks

Price Change: Discrete values

Ordered probit model: Hauseman, Lo, & MacKinlay (1992)

ADS model: Rydberg & Shephard (1998), McCulloch & Tsay (2000)

A simple ADS decomposition:

The price change at the ith transaction can be written as

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

where A_i is the a binary variable defined as

$$A_i = \begin{cases} 1 & \text{if there is a price change at } i\text{−th trade,} \\ 0 & \text{if there is not a price change at } i\text{−th trade.} \end{cases}$$

D_i denotes the direction of price change, i.e.

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

S_i is the the size of price change in ticks.

Our target is $P(y_i = s|F_{i-1})$.

$P(y_i = s|F_{i-1}) =$

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

Assume

$p_i = P(A_i = 1|F_{i-1})$ over time and

$$\ln\left(\frac{p_i}{1-p_i}\right) = x_i\beta', \;\; p_i = \frac{e^{x_i\beta'}}{1+e^{x_i\beta'}},$$

$\delta_i = P(D_i = 1|A_i = 1, F_{i-1})$ over time and

$$\ln\left(\frac{\delta_i}{1-\delta_i}\right) = z_i\gamma', \;\; \delta_i = \frac{e^{z_i\gamma'}}{1+e^{z_i\gamma'}},$$

where x_i and z_i are finite vectors consisting of F_{i-1}, e.g. $x_i = (\Delta t_{i-1}, A_{i-1}, D_{i-1}, S_{i-1})$, etc..

β and γ are constant parameters.

$$\begin{aligned} S_i &= s+1|\{D_i, A_i = 1\} \\ &\sim \begin{cases} \lambda_{1i}(1-\lambda_{1i})^s & \text{if } D_i = 1,\ A_i = 1, \\ \lambda_{2i}(1-\lambda_{2i})^s & \text{if } D_i = -1,\ A_i = 1. \end{cases} \end{aligned}$$

$\lambda_{1i}(1-\lambda_{1i})^s$ is a geometric distribution.

$$\ln\left(\frac{\lambda_{ji}}{1-\lambda_{ji}}\right) = w_{ji}\theta_j', \;\; \lambda_{ji} = \frac{e^{w_{ji}\theta_j'}}{1+e^{w_{ji}\theta_j'}}, \;\; j = 1, 2$$

where w_{ji} and θ_j are defined as x_i and β.

Log-likelihood function:

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

An Example: IBM data 59,838 observations.

$F_{i-1} = \{A_{i-1}, D_{i-1}, S_{i-1}, V_{i-1}, x_{i-1}, BA_i, \cdots\}$

V_{i-1}: volume of the previous trade (divided by 1000)
x_{i-1}: previous duration
BA_i: the prevailing bid-ask spread

Model:

$x_i\beta' = \beta_0 + \beta_1 A_{i-1}$

$z_i\gamma' = \delta_0 + \delta_1 D_{i-1}$

$w_{ji}\theta_j' = \theta_{j0} + \theta_{j1} S_{i-1}.$

Results:

$Parameter$	β_0	β_1	δ_0	δ_1
$Estimate$	-1.057	0.962	-0.067	-2.307
$Std.Err.$	0.104	0.044	0.023	0.056

$Parameter$	θ_{10}	θ_{11}	θ_{20}	θ_{22}
$Estimate$	2.235	-0.670	2.085	-0.509
$Std.Err.$	0.029	0.050	0.187	0.139

Implication

1. Prob of price change:

$P(A_i = 1|A_{i-1} = 0) = 0.258$

$P(A_i = 1|A_{i-1} = 1) = 0.476.$

2. Direction of price change:

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

3. Weak evidence of price change cluster: price increases

$S_i - 1|(D_i = 1) \sim \lambda_{1i}(1 - \lambda_{1i})^s,$

where

$$\lambda_{1i} = \frac{\exp(2.235 - 0.670 S_{i-1})}{1 + \exp(2.235 - 0.670 S_{i-1})}.$$

Duration and Duration Models

Focus on intraday time duration between transactions

Autoregressive conditional duration (ACD) model:

Engle and Russell (1998)

Intraday durations between trades, in seconds.

Use ideas of GARCH models

Define

1. t_i: time of the i-th trade, starting at midnight, measured in seconds.

2. $X_i = t_i - t_{i-1}$

3. $f(t)$: Diurnal pattern of daily trading.

4. $x_i = X_i / f(t_i)$: adjusted time duration of i-th trade

5. F_i: information set available at t_i (inclusive)

6. ψ_i: Expected duration, $E(x_i|F_{i-1})$.

ACD(r, s) model:

$$\frac{x_i}{\psi_i} \sim \epsilon_i, \epsilon_i \geq 0 \; and \; \sim i.i.d. \; with \; \epsilon_i = 1.$$

$$\psi_i = \omega_0 + \sum_{j=1}^{r} \gamma_j x_{i-j} + \sum_{j=1}^{s} \omega_j \psi_{i-j}.$$

The distribution of ϵ_i is either Standard Exponential or Standardized Weibull.

Refer to as an EACD or WACD model, respectively.

Let $\eta_i = x_i - \psi_i$.

$\{\eta_i\}$ is a martingale difference sequence.

ACD(r, s) model becomes

$$x_i = \omega_0 + \sum_{j=1}^{\max\{r,s\}} (\gamma_j + \omega_j) x_{i-j} - \sum_{j=1}^{s} \omega_j \eta_{i-j} + \eta_i.$$

Some properties of ACD model are easily available.

Remark: Duration models can be estimated via programs similar to GARCH models.