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.
   - Impact of daily price limits (many foreign markets)

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?

It turns out r_t^o is given in Eq. (5.1),

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

One can use this relation to show that

$$\begin{aligned} \text{Var}(r_t^o) &= \sigma^2 + \frac{2\pi\mu^2}{1-\pi} \\ \text{Cov}(r_t^o, r_{t-j}^o) &= -\mu^2\pi^j, \quad j \geq 1. \end{aligned}$$

Bid-ask bounce

Bid and ask quotes introduce **negative** lag-1 serial correlation.

Setup: simplest case of Roll(1984)

True price P_t^* is unchnaged, i.e. $P_t^* = P_{t-1}^*$

$S = P_a - P_b$ is the bid-ask spread

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

Then,

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

where I_t and I_{t-1} are independent dummy variables with $P(I_t = 1) = 0.5$.

Once can show that

$$\begin{aligned} \text{Var}(\Delta P_t) &= S^2/2 \\ \text{Cov}(\Delta P_t, \Delta P_{t-1}) &= -S^2/4 \\ \text{Cov}(\Delta P_t, \Delta P_{t-j}) &= 0, \quad j > 1. \end{aligned}$$

The result continues to hold if P_t^* follows a random walk model.

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 FX
3. 1-minute returns on index futures and cash market

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.
6. Diurnal Pattern

An illustration

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%).
- Kurtosis of adj-duration: 44.23(.02)

Frequencies of price change

Number(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 Figure 5.1 and Figure 5.2 on page 183 of the text.

Another example: Transaction data of IBM stock in December 1999, from TAQ.

Two important changes

- Number of trades increased to sixfold (134,120 trades)
  - trades with zero time-duration became 23%
  - 42 trades in a single second in December 3, 1999.
- Tick size reduces to \$1/16; the percentage of trades without price change decreased from 67% to 46%.

Econometric models

1. Time Duration and Duration Models
2. Nonlinearity in Time Durations
3. A Model for Price Change and Duration
4. Hierarchical Models
5. Models for bid and ask quotes

Include statistical tools and methods useful in analyzing HF financial data

Data quality (need some cleaning)

- Trading hours (FX-24, 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)

Look at a simple ADS decomposition:

- Price $P_t = P_0 + \Sigma_i^{N(t)} C_i$
- Number of transactions in [0,t]: $N(t)$
- $C_i = A_i D_i S_i$
  - Action:
$$A_i = \begin{cases} 1 & \text{if } C_i \neq 0 \\ 0 & \text{otherwise} \end{cases}$$
  - Direction, given $A_i = 1$:
$$D_i = \begin{cases} 1 & \text{if } C_i > 0 \\ 0 & \text{if } C_i < 0 \end{cases}$$
  - Size, given $A_i = 1$ and D_i: multiple of tick size
- Can be estimated by logistic regression

Model specification:

- Action A_i: Governed by a logistic regression

$$P(A_i = 1|F_{i-1}) = \text{logit}(F_{i-1})$$

- Direction given $A_i = 1$:

$$P(D_i = 1|F_{i-1}, A_i = 1) = \text{logit}(A_i, F_{i-1})$$

- Size given $A_i = 1$ and D_i:

$$P(S_i = s|A_i = 1, D_i = 1, F_{i-1}) \sim 1 + g(\lambda_{u,i})$$
$$P(S_i = s|A_i = 1, D_i = -1, F_{i-1}) \sim 1 + g(\lambda_{d,i})$$

  where $g(.)$ denotes a Geometric distribution and $\lambda_{j,i}$ is governed by a logistic equation:

$$\ln(\frac{\lambda_{j,i}}{1-\lambda_{j,i}}) = \text{linear function of } F_{i-1}, A_i = 1, D_i.$$

Likelihood function:

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

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

A simple ADS model: IBM data 59,838 observations.

- Predictors: $\{A_{i-1}, D_{i-1}, S_{i-1}, V_{i-1}, x_{i-1}, BA_i\}$
  1. V_{i-1}: volume of the previous trade (divided by 1000)
  2. x_{i-1}: previous duration
  3. BA_i: the prevailing bid-ask spread
- Model:
  1. Action: $P(A_i|F_{i-1}) = p_i$, $\text{logit}(p_i) = \beta_0 + \beta_1 A_{i-1}$
  2. Direction: $P(D_i = 1|A_i = 1, F_{i-1}) = \gamma_i$, $\text{logit}(\gamma_i) = \delta_0 + \delta_1 D_{i-1}$
  3. Size: $\text{logit}(\lambda_{j,i}) = \theta_{j,0} + \theta_{j,1} S_{i-1}$.
- 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	$\lambda_{u,0}$	$\lambda_{u,1}$	$\lambda_{d,0}$	$\lambda_{d,1}$
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:

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

Bid-ask bounce

3. Weak evidence of price change cluster: price increases

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

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:

$$\frac{x_i}{\psi_i} \sim \epsilon_i, \quad \epsilon_i \sim i.i.d. \quad g(\theta), \quad E(\epsilon_i) = 1$$

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

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

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

Some properties of ACD model are easily available.

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

Focus on the first 5 days of November, 1990.

- Dates: November 1, 2, 5, 6, 7 of 1990
- Sample size: 3534 data points

A simple model: WACD(1,1)

$$x_i = \psi_i \epsilon_i, \quad \psi_i = 0.291 + 0.077 x_{i-1} + 0.836 \psi_{i-1}$$

where $\{\epsilon_i\}$ iid Weibull with $\hat{\alpha} = 0.878(0.011)$.

Implication:

- $E(x_i) = 3.34$ (vs 3.29, sample)
- Not strongly persistent: $\hat{\gamma}_1 + \hat{\omega}_1 \approx 0.91$
- Decaying intensity function.

Model diagnostics:

- ACF of $\hat{\epsilon}_i = \frac{x_i}{\hat{\psi}_i}$: $Q(12) = 5.7$ and $Q(24) = 19.9$
- ACF of $\hat{\epsilon}_i^2$: $Q(12) = 6.5$ and $Q(24) = 15.1$