High Frequency Data Analysis & Market Micro structure

T. V. Ramanathan

Department of Statistics
Savitribai Phule Pune University
Pune - 411007 (India).

ram@stats.unipune.ac.in

National Workshop on Financial Data Analytics

C R Rao AIMSCS, Hyderabad

December 27-30, 2014

Introduction

- Observations taken at finer time intervals
- Ultimate - Transaction-by-Transaction data - NYSE TAQ (Trades & quotes data base - time duration - second)
- HFD is important for studying issues related to trading process and market micro structure
  1. To compare the efficiency of different trading systems in price discovery
  2. To study the dynamics of bid-and-ask quotes of a particular stock
  3. To study the order dynamics or to investigate the question of who provides the market liquidity.
- Cho, Russel, Tiao and Tsay (2003) - Analyzed HFD from Taiwanese stock exchange.

Some important aspects that can be looked in to:

1. Non synchronous trading
2. Bid-ask spread
3. Transaction Data
4. Price movements - Modeling
5. Duration - Modeling

Non synchronous trading

- Different stocks have different trading frequencies, (even for same stock has different trading frequencies) - For daily returns, non synchornous trading can introduce the following
  1. lag-1 cross correlation between stock returns
  2. lag-1 serial correlation in a portfolio return
  3. negative serial correlations of the return series of a single stock.
- Illustrate with stock A (more frequently traded than) and stock B. News impact at the close of the day.
- see Campbell, Lo, and MacKinlay (1997) and the references therein.

Non synchronous trading

- Let r_t be the continuously compounded return of a security at the time index t, (iid assumption for simplicity) with $E(r_t) = \mu$ and $Var(r_t) = \sigma^2$.
- Let π be the probability that the security is not traded at a particular time period (assumed to be same for all time periods)
- Let r_t^0 be the observed return ($= 0$ when there is no trade at time t). Then we have

Non synchronous trading

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

Non synchronous trading

- It can be proved that

$$E(r_t^0) = \mu, \;\; Var(r_t^0) = \sigma^+ \frac{2\pi\mu^2}{1-\pi}, \;\; Cov(r_t^0, r_{t-j}^0) = -\mu^2\pi^j, \; j > 1$$

- Thus, when $\mu \neq 0$, the non synchronous trading induces negative autocorrelations in an observed security return series.
- The above discussion can be generalized to the return series of a portfolio that consists of N securities; see Campbell et al. (1997, Chapter 3).

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.
- Market makers are granted monopoly rights by the exchange to post different prices for purchases and sales of a security.
- Buy at the bid price P_b and sell at a higher ask price P_a.
- The difference $P_a - P_b$ is called as the bid-ask spread, which is the primary source of compensation for market makers.

Bid-Ask Spread

- Bid-ask spread introduces negative lag-1 serial correlation in an asset return.
- Model introduced by Roll (1984): $P_t = P_t^* + I_t \frac{S}{2}$
  P_t - Observed market price of an asset, $S = P_a - P_b$, the bid-ask spread, P_t* - fundamental value of the asset in a frictionless market, I_t, sequence of independent random variables taking values $+1$ and -1 with equal probabilities $1/2$.
- I_t can be interpreted as an order-type indicator, with 1 signifying buyer-initiated transaction and -1 seller-initiated transaction.
- Whenever there is no change in P_t^*, the observed price change is

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

Bid-Ask Spread

- It can be easily seen that

$$E(\Delta P_t) = 0, \;\; Var(\Delta P_t) = S^2/2,$$

$$Cov(\Delta P_t, \Delta P_{t-1}) = -S^2/4, \;\; Cov(\Delta P_t, \Delta P_{t-j}) = 0 \text{ for } j > 1$$

- The autocorrelation function of ΔP_t is

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

- Known as *bid-ask bounce.*

Bid-Ask Spread

- **Intuitive interpretation:**
- Assume $P_t^* = (P_a + P_b)/2$. Then P_t will be P_a or P_b. If previous obs. value is P_a (higher value), then the current observed value will be either 0 or P_b (lower value). Thus, ΔP_t is either 0 or $-S$. In the case of P_b, it will be 0 or S, and thus the lag-1 negative correlation in ΔP_t becomes apparent.

Bid-Ask Spread

- More realistic formulation: Assume P_t^* as a random walk,

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

- Then,

$$Var(\Delta P_t) = \sigma^2 + S^2/2, \;\; Cov(\Delta P_t, \Delta P_{t-j}) = 0 \text{ for } j > 1$$

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

- To know more about components of bid-ask spread, refer Campbell et al. (1997).

Transaction Data

- Transaction data: t_i - time measured in seconds from midnight, at which the i-th transaction of an asset, transaction price, the transaction volume, the prevailing bid and ask quotes etc. constitute the transactions data.
- Unequally Spaced Time Intervals
- Discrete-Valued Prices
- Existence of a Daily Periodic or Diurnal Pattern
- Multiple Transactions within a Single Second.

Ordered Probit Model: (Hauseman, Lo and MacKinlay (1992))

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

- y_t^* - unobservable price change of the asset under study, $y_t^* = P_{t_i}^* - P_{t_{i-1}}^*$, P_t^* is the virtual price of the asset at time t, x_i is the p-dimensional vector of explanatory variables available at time t_{i-1}, $E(\epsilon_i|x_i) = 0$, $Var(\epsilon_i|x_i) = \sigma_i^2$ and $Cov(\epsilon_i, \epsilon_j) = 0$ *for* $i \neq j$.
- Let the observed price change y_i assumes k possible values $s_1, s_2, ..., s_k$
- The ordered probit model postulates the relationship between y_i and y_* as

$$y_i = s_j, \text{ if } \alpha_{j-1} < y_i^* \leq \alpha_j, \ j = 1, 2, ..., k; -\infty < \alpha_1, ..., \alpha_k < \infty$$

Some Basics of Duration Data

- Three types of durations: trade, price and volume - proxies for trading intensity, trading volatility and liquidity.
- Trade durations (intensity): Time interval between consecutive trades.
- Price durations (volatility): Minimum duration that is required to observe a price change not less than a given amount.
- Volume durations (liquidity): The time spells such that the total traded volume is not smaller than (lets say) 25000 shares.
- For trade durations zero duration are common.
- Intra-day seasonality is observed - To be removed - (Cubic spline function suggested by Engle and Russell (1998))

Duration Models - ACD

- Why duration models? Time series cannot accommodate irregularly spaced data.
- High frequency data - Transaction data - Irregularly spaced
- Engle and Russell (1998) developed autoregressive conditional duration (ACD) model
- $\{t_0, t_1, ..., t_n, ...\}$ - sequence of arrival times of events
  $0 = t_0 \leq t_1 \leq t_2 \leq ... \leq t_n \leq .....$
- Duration: $x_i = t_i - t_{i-1}$
- $x_i = \psi_i \ \epsilon_i, \quad \psi_i = E(x_i|\mathcal{F}_{i-1})$ where ϵ_i are iid, $E(\epsilon_i) = 1$.

$$\psi_i = \omega + \sum_{j=1}^{p} \alpha_j x_{i-j} + \sum_{j=1}^{q} \beta_j \psi_{i-j}$$

- For a survey on ACD: Pacurar (2008)

ACD - Some Properties

- $ARMA(max(p,q),q)$ formulation possible.
- $\sum_{i=1}^{p} \alpha_i + \sum_{i=1}^{q} \beta_i < 1$, condition for cov. stationarity
- Stationarity and invertibility conditions need the roots of $1 - \alpha(L) - \beta(L)$ and $1 - \beta(L)$ respectively to lie outside the unit circle.
- $$E(x_i|\mathcal{F}_{i-1}) = \psi_i, \quad E(x_i) = \frac{\omega}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{i=1}^{q} \beta_i}$$
- $$Var(x_i|\mathcal{F}_{i-1}) = \psi_i^2 Var(\epsilon_i)$$

$$Var(x_i) = [E(x_i)]^2 \left(\frac{1 - 2(\sum_{i=1}^{p} \alpha_i)(\sum_{i=1}^{q} \beta_i) - (\sum_{i=1}^{q} \beta_i)^2}{1 - 2(\sum_{i=1}^{p} \alpha_i)^2 - 2(\sum_{i=1}^{p} \alpha_i)(\sum_{i=1}^{q} \beta_i) - (\sum_{i=1}^{p} \alpha_i)^2} \right).$$

SCD Models

- Let $\{\tau_i\}$ be the occurrence time of a certain event and $d_i = \tau_i - \tau_{i-1},\ i = 1, 2, ..., n$ be the durations.
- In the case of a SCD model, these observed durations d_i are modeled as the product of a latent variable $\mathbf{\Psi}_i$ and a positive random variable ϵ_i.
- That is,

$$d_i = \mathbf{\Psi}_i\ \epsilon_i, \quad \mathbf{\Psi}_i = e^{\psi_i},\ \psi_i = \alpha + \beta\psi_{i-1} + u_i,$$

with $\epsilon_i|\mathcal{F}_{i-1}$ independent and identically distributed (i.i.d.) random variables having a positive support, u_i i.i.d. with support on the real line R and ϵ_i is independent of u_j for all i and j, where $\mathcal{F}_{i-1}$ denotes the information set available at the end of duration d_{i-1}.

- It is assumed that the initial value ψ_0 is drawn from the stationary distribution of ψ.

Ongoing work : DST-SERB Project

- Developing a method of estimation for tv-ACD model and establishing the probabilistic properties of the estimators.
- Proposing a method to remove the intra-day seasonality in high frequency data using wavelets
- Generalized class of ACD model incorporating structural breaks which accommodates other successful ACD models as well
- Proposing non/semi parametric SCD models and estimating the parameters using different approaches and comparing all the available estimation methods
- Modeling durations using point process theory
- Developing duration models using new statistical tools which capture the dependence in the data very well like copulas and wavelets

Ongoing work : DST-SERB Project

- Employ the Bayesian non-parametric techniques of Ghosh and Ramamoorthi (2002)for modeling and estimation in almost all of our research works. The advances in Bayesian non-parametric inference methods have not received a full critical and comparative analysis of their scope and limitations in financial modelling;
- Ghosh et. al. (2011) studied Bayesian inference for non-parametric state-space model. These techniques will be useful for us while estimating non-parametric SCD models.
- Adams (2009) studied the Bayesian inference for point processes which we will be using while working on intensity modelling of durations.

References

1. Adams (2009)
2. Bauwens, L. and Veredas, D. (2004) The stochastic conditional duration model: a latent factor model for the analysis of fnancial durations, *Journal of Econometrics*, 119, 381-41
3. Ghosh, J. K. and R. V. Ramamoorthi. (2002). *Bayesian Nonparametrics*. Springer, 2002.
4. Campbell, Lo, and MacKinlay (1997)
5. Cho, Russel, Tiao and Tsay (2003)
6. Engle, R.F. and Russell, J.R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 66, 1127-1162.

References

1. Ghosh, A., Mukhopadhyay, S., Roy, S., Bhattacharya, S. (2011). Bayesian Inference in Non-parametric Dynamic State-Space Models. *arXiv preprint arXiv:1108.3262*
2. Hauseman, Lo and MacKinlay (1992)
3. Pacurar, M. (2008). Autoregressive conditional duration models in finance: A survey in the theoretical and empirical literature. *Journal of Economic Surveys* 22, 711-751.

THANK YOU