Análise de Dados de Alta Freqüência

1. Introdução

- Dados finaceiros de alta freqüência (DAF) são observações sobre variáveis financeiras ações, taxas de juros, taxas de câmbio, opções etc, tomadas diariamente ou em escala intra-diária, freqüentemente irregularmente espaçadas no tempo.

- São importantes em estudos empíricos da micro-estrutura do mercado: descoberta de preços, competições entre mercados relacionados, comportamento estratégico de participantes do mercado, modelagem da dinâmica de mercado em tempo real.

- Dados típicos: "'trade-by-trade"'em mercados acionários, "quotes" de taxas de câmbio; os tempos são geralmente dados em segundos; dados "tick-by-tick"'.

- DAF têm características únicas, que não aparecem em dados com freqüências mais baixas (semanais, mensais etc).

- Dados não-sincronizados

  - negociações de ações não aparecem de forma sincronizada; ações diferentes têm freqüências de negócios diferentes e mesmo para uma mesma ação, a intensidade de negociação varia de hora para hora do dia.

  - para retornos diários de ações, negociações não-sincronizadas podem introduzir:

  a) correlação cruzada de lag 1 entre retornos de ações;

  b) correlação de lag 1 no retorno de uma carteira;

  c) em alguma situações, correlação negativa na série de retornos de uma particular ação.

- O número de observações de uma série de DAF é usualmente enorme; por exemplo, o número diário de cotações para t.c. USD/EUR no mercado "spot" é da ordem de 20.000.

- Os DAF são geralmente registrados com erros e têm que ser corrigidos ("limpos") antes de serem analisados.

  Os dados são irregularmente espaçados, com número aleatório de observações por dia.

- DAF exibem padrões periódicos (intra-dia e intra-semana): as atividades de uma bolsa de valores são mais densas no começo e fechamento do que na hora do almoço.

2. Algumas Bases de Dados

- TAQ (Trades and Quotes): dados do NYSE, AMEX, NASDAQ; mercado acionário; desde 1992.

- Berkeley Options Data Base: dados de mercados de opções; 1976-1996.

- Olsen Associates: mercados de câmbio; desde 1980's.

3. Software e Surveys

- S-PLUS HF Library

  B. Yam and Eric Zivot

  http://faculty.washington.edu/ezivot/splus.htm

- S-Plus Library

  W. Breymann

http://www.math.ethz.edu/∼breymann

- Andersen (2000), Campbel et al. (1997), Dacorogna et al. (2001), Wood (2000), Goodhart and O'Hara (1997).

4. Manipulação de Dados

- Primeiramente, é necessário construir as variáveis de mercado, que incluem: variações de preços entre transações (ou entre cotações,n caso de dados de FX=t.c.), durações entre negócios ou cotações, e "spread" entre cotações bid e ask.

- Preços de ativos financeiros movem-se em incrementos mínimos, chamados "ticks" (ou "tick sizes"), logo variações de preços podem ser expressas em u.m. ou número de ticks. Por exemplo, em cotações

de USD/EUR, o tick é $ 0.0001, enquanto que na NYSE é de $ 0.01, e na BOVESPA R$ 0,01.

- variações de preços são obtidas tomando-se diferenças no nível de preço.

- Retornos de ações "overnight" diferem substancialmente dos demais retornos ao longo do dia, portanto é necessário que se tenha um critério para analisá-los, o mesmo ocorrendo com finais de semana. Mesma observação vale para durações.

- spread bid-ask: diferença entre

  bid: preços segundo os quais os "traders" compram ativos; (p_b)

  e

  ask: preços segundo os quais os " traders" vendem os ativos; (p_a)

"bid-ask spread" : $p_a - p_b$; também é dado em múltiplos de ticks (usualmente pequeno, um ou dois ticks);

Para o público: p_b= preço de venda; p_a= preço de compra.

- Direção da negociação: uma negociação é considerada

  "buy" - se preço da transação é maior do que a cotação média do bid-ask;

  "sell" -se for menor;

  " indeterminada" - se for igual.

- Volatilidade realizada: central para a teoria moderna de finanças; é usada em

  apreçamento de ativos, seleção de carteiras e administração de riscos; pode ser estimada de várias maneiras.

- Variações de preços de negócios e de cotações são v.a. discretas e múltiplas

do tick; uma proporção significativa de observações não apresenta variações de preços; a maioria limitadas por ± 3 ticks.

- Muitas negociações parecem ocorrer no mesmo tempo, logo há um número significativo de transações com durações nulas.

- Atividades financeiras podem exibir padrões periódicos devido ao efeito do calendário: preços de ativos, volumes, durações, spread de bid-ask, freqüência de ticks. Por exemplo, na NYSE, intensidade de transações têm a forma de um ''U'' invertido. Durações têm efeito oposto. Dados de taxas de câmbio têm um padrão intra-diário com 3 picos, correspondentes a horas de negócios nos três centros geográficos: Asia, Europa e USA.

5. Modelo de Roll (1984)

P_t: preço de mercado do ativo

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

onde :

$S = p_a - p_b$;

P_t^*: preço do ativo num mercado "sem fricção";

$I_t \sim$ i.i.d., binária,

$$I_t = \begin{cases} 1, & \text{com probabilidade } 1/2 \text{ (compra iniciada)}, \\ -1, & \text{com probabilidade } 1/2 \text{ (venda iniciada)}. \end{cases}$$

Logo,

$$P_t = P_t^* + \begin{cases} S/2, & \text{com prob. } 1/2, \\ -S/2, & \text{com prob. } 1/2 \end{cases}$$

Se não houver mudança em P_t^*, as variações de preço são

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

Dado que $E(I_t) = 0$ e $\text{Var}(I_t) = 1$, seguem-se que:

$$E(\Delta P_t) = 0,$$

$$\text{Var}(\Delta P_t) = \frac{S^2}{2},$$

$$\text{Cov}(\Delta P_t, \Delta P_{t-1}) = -\frac{S^2}{4},$$

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

do que resulta

$$\rho_j(\Delta P_t) = \begin{cases} -0,5, & \text{se } j = 1 \\ 0, & \text{se } j > 1. \end{cases}$$

Portanto, o "bid-ask spread" introduz correlação negativa de lag 1 na série de variações de preços. É o chamado "bid-ask bounce".

Suponha que : $P_t^* = \frac{p_a + p_b}{2}$.

Então, $P_t = p_a$ ou $P_t = p_b$, com probabilidade 1/2.

Se $P_t = p_a$, então $\Delta P_t = 0$ ou $\Delta P_t = -S$.

Se $P_t = p_b$, então $\Delta P_t = 0$ ou $\Delta P_t = S$.

Suposição: $\Delta P_t^* = P_t^* - P_{t-1}^* = \varepsilon_t \sim$ i.i.d.$(0, \sigma^2)$,

ou seja, um passeio aleatório (suponha ε_t independente de I_t).

Pode-se mostrar neste caso que

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

ou seja, a correlação é reduzida, mas o efeito permanece.

O efeito do spread continua a existir para carteiras e séries multivariadas.

6. Modelos para Variações de Preços

Sejam:

$y_i = \Delta P_{t_i} = P_{t_i} - P_{t_{i-1}}$ variação de preço,

$$\Delta t_i = t_i - t_{i-1} \quad \text{duração}$$

Como dados são discretos e concentrados (em "não há variação"), há dificuldades em modelar variações de preços intra-diários.

- Modelo "Probit" Ordenado

  y_i: k possíveis valores, $s_1, \ldots, s_k$.

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

  onde $y_i^* = P_{t_i}^* - P_{t_{i-1}}^*$ é a variação de preço do ativo *virtual*, não observada, suposta seguir o modelo

  $$y_i^* = \mathbf{x}_i \beta + \varepsilon_i.$$

  Hauseman, Lo and MacKinlay (1992)

- Modelo de Decomposição

  $$y_i = A_i D_i S_i,$$

  onde:

  $A_i = 1$, se existe variação de preço na negociação i,

$A_i = 0$, caso contrário;

$\{D_i|A_1 = 1\} = 1$, se preço aumenta na negociação i,

$\{D_i|A_1 = 1\} = -1$, se preço diminui na negociação i;

$S_i =$ tamanho da variação de preço (em ticks), se existe variação;

$S_i = 0$, se nao existe variação de preço.

$\mathcal{F}_i =$ informação até i-ésima transação

$$P(y_i|\mathcal{F}_i) = P(A_i D_i S_i|\mathcal{F}_i) =$$

$$= P(S_i|D_i, A_i, \mathcal{F}_i).P(D_i|A_i, \mathcal{F}_i).P(A_i|\mathcal{F}_i).$$

$$p_i = P(A_i = 1), \quad \delta_i = P(D_i = 1|A_i = 1),$$

$S_i|D_i, A_i = 1 \sim$ 1+ geométrica de parâmetro $\lambda_{j,i}$.

Classificar negociação i em 3 categorias:

(1) não há variação de preços; $A_i = 0$, com prob. $1 - p_i$;

(2) preço aumenta: $A_i = 1, D_i = 1$, com prob. $p_i\delta_i$; S_i cresce, com densidade $1 + g(\lambda_{u,i})$;

(3) preço decresce: $A_i = 1, D_i = -1$, com prob. $p_i(1 - \delta_i)$, S_i decresce com densidade $1 + g(\lambda_{d,i})$.

Obter verossimilhança e estimadores dos parâmetros associados a modelos logísticos usados para p_i e δ_i.

McCulloch and Tsay (2000)

7. Modelos para Durações

São modelos propostos para intervalos de tempo entre negociações. Lembremos que

durações longas indicam falta de atividade de "trades", ou períodos sem informações novas.

- Modelo ACD

  Engle e Russel (1998) propuseram o mo-

  delo ACD (de "autoregressive conditional duration"). Zhang, Russel e Tsay (2001) estenderam o modelo ACD para considerar não-linearidades e quebras estruturais nos dados.

  Como vimos, transações intradiárias exibem padrões periódicos. Chamemos de x_i a $i-$ésima duração ajustada, ou seja,

$$x_i = \frac{\Delta t_i}{f(t_i)},$$

  onde $f(t_i)$ é uma função determinística que estima a componente periódica de

Δt_i. Tsay (2002) sugere usar "splines" cúbicas, funções indicadoras e quadráticas.

- Seja $\psi_i = E(x_i|\mathcal{F}_{i-1})$ a média condicional da duração entre os negócios $i-1$ e i e $\mathcal{F}_{i-1}$ toda a informação disponível até a negociação $i-1$.

  O modelo ACD é dado por

$$x_i = \psi_i \varepsilon_i, \quad \varepsilon_i \geq 0, \sim \text{iid}, \tag{2}$$

$$\psi_i = \omega + \sum_{j=1}^{r} \gamma_j x_{i-j} + \sum_{j=1}^{s} \omega_j \psi_{i-j}, \tag{3}$$

  $\omega > 0, \ \gamma_j \geq 0, \omega_j \geq 0.$

  O modelo é indicado ACD(r, s).

- A seqüência ε_i usualmente é considerada uma v.a. exponencial com média 1 ou Weibull, com média 1.

- Se $\eta_i = x_i - \psi_i$ (uma diferença martingale) podemos escrever ($q = \max(r, s)$)

$$x_i = \omega + \sum_{j=1}^{q} (\gamma_j + \omega_j) x_{i-j} - \sum_{j=1}^{s} \omega_j \eta_{i-j} + \eta_j,$$

ou seja um ARMA(q,s).

- Supondo-se modelos estacionários,

$$E(x_i) = \frac{\omega}{1 - \sum_{j=1}^{q} (\gamma_j + \omega_j)}. \qquad (4)$$

Logo, devemos ter $\sum_j (\gamma_j + \omega_j) < 1$.

- EACD (1,1)

$$\begin{aligned} x_i &= \psi_i \varepsilon_i, \quad \varepsilon_i \sim \mathrm{E}(1), & (5) \\ \psi_i &= \omega + \gamma_1 x_{i-1} + \omega_1 \psi_{i-1}. & (6) \end{aligned}$$

- $E(x_i) = E(\psi_i)$;

Se x_i estacionário, $E(x_i) = E(\psi_i) = \frac{\omega}{1-\gamma_1-\omega_1} = \mu_x$.

$\text{Var}(x_i) = \mu_x^2 \frac{1-\omega_1^2-2\gamma_1\omega_1}{1-\omega_1^2-2\gamma_1\omega_1-2\gamma_1^2}$.

Portanto, variância incodicional constante se $2\gamma_1^2 + \omega_1^2 + 2\gamma_1\omega_1 < 1$.

- Estimação: máxima verossimilhança.

- Volatilidade Realizada
  - Modelos da família ARCH e MVE estimam volatilidade, considerada uma variável não-observada.
  - Resultados não satisfatórios, previsões não precisas.
  - Modelos multivariados podem ser não-factíveis para dimensões altas.
  - Estimar, modelar e prever volatilidade e correlação usando dados de alta freqüência intra-diários.

    Medidas de volatilidade e correlação baseadas em retornos ao quadrado

e produtos de retornos. Depois modelar com modelos mais simples (tipo ARIMA, ARFIMA).

- $p_{i,t}$: log-preço do ativo i, no instante t, alinhados a um intervalo de tempo comum igualmente espaçado (e.g., 5, 15 ou 30 minutos).

  m: número de "trades" durante uma sessão de negócios

  $m = 72$ intervalos de 5 min, se sessão tem 6 horas por dia.

  Δt: intervalo de amostragem (e.g, 5 min).

  $r_i(t, j\Delta t)$: log-retorno do ativo i, $i = 1, \ldots, n$, no dia t, $t = 1, \ldots, T$, no instante $j\Delta t$, $j = 1, \ldots, m$.

- Variância Realizada (VR)

$$VR_{it} = \sum_{j=1}^{m} r_i^2(t, j\Delta t), \quad t = 1, \ldots, T.$$

– Volatilidade Realizada (VOLR)

$$VOLR_{i,t} = \sqrt{VR_{i,t}}.$$

– log-volatildade realizada (LVOLR)

$$LVOLR_{i,t} = \ln(VOLR_{i,t}).$$

– $\mathbf{r}_{i,t} = (r_i(t, \Delta t), \ldots, r_i(t, m\Delta t))'$: vetor de log-retornos do ativo i, $i = 1, \ldots, n$, no dia t, em intervalos de Δt minutos.

$\mathbf{r}_t = (\mathbf{r}_{1,t}, \ldots, \mathbf{r}_{n,t})'$.

– Matriz $(n \times n)$ de covariâncias realizadas:

$$COVR_t = \mathbf{r}_t \mathbf{r}_t^{'}.$$

– A correlação realizada entre os ativos i e j é dada por

$$CORR_{i,j,t} = \frac{COVR_{i,j,t}}{VOLR_{i,t}VOLR_{j,t}},$$

$i, j = 1, \ldots, n,\ t = 1, \ldots, T.$

- Problemas práticos: escolha de m ou Δt; propriedades (consistência e normalidade assintótica) dependem de $\Delta t \to 0$ (ou $m \to \infty$).

- Duas questões importantes:

  Q_1: qual parâmetro VR estima?

  Q_2: Estimativas de VR são economicamente importantes?

- Andersen, Bollerslev, Diebold e Labys (2000 a,b, 2001, 2003) (ABDL)

  Barndorff-Nielsen e Shephard (2002 a,b, 2004 a,b) BNS

  Desenvolveram teoria rigorosa ligando VR com processos de tempo contínuo de variações de retornos quadráticos.

- Referências

Andersen, T. G. (2000). Some reflections on analysis of high-frequency data. *Journal of Business and Economic Statistics*

Dacorogna, M.M., Gençay, R., Müller, U.A., Olsen, R.B., and Pictet, O.V. (2001). *An Introduction to High- Frequency Finance*. Academic Press.

Goodhart, C.A.E. and O' Hara, M. (1997). High-frequency data in financial markets: Issues and applications. *Journal of Empirical Finance*, **4**, 73-114.

Wood, R.A. (2000). Market microstructure research databases: History and projections. *Journal of Business and Economic Statistics*, **18**, 14-145.

Tsay, R.S. (2002) *Analysis of Financial Time Series*. Wiley.

Engle, R.F. and Russel, J.R. (1998). Autoregressive conditional duration: A new model for irregularly space transaction data. *Econometrica*, **66**, 1127-1162.

Andersen, T., Bollerslev, T., Diebold, F.X. and Labys, P. (2000a). Exchange rate returns standardized by realized volatilidade are (nearly) Gaussian. *Multinational Finance Journal*, **4**, 159–179.

Andersen, T., Bollerslev, T., Diebold, F.X. and Labys, P. (2000b) Great realizations. *Risk*, **13**, 105–108.

Andersen, T., Bollerslev, T., Diebold, F.X. and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, **96**, 42–55.

Andersen, T., Bollerslev, T., Diebold, F.X. and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, **71**, 579–626.

Barndorff-Nielsen, O. and Shephard, N. (2002 a). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics*, **17**, 457–477.

Barndorff-Nielsen, O. and Shephard, N. (2002 b). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society*, Series B, **64**, 253–280.

Barndorff-Nielsen, O. and Shephard, N. (2004 a). Econometric analysis of realized covariation: High-frequency based covariance, regression and correlation in

financial economics. *Econometrica*, **73**, 885–926.

Barndorff-Nielsen, O. and Shephard, N. (2004 b). How accurate is the asymptotic approximation to the distribution of realized volatility? In *Identification and Inference for Econometric Models*. A Festschrift in Honour of T.J. Rothenberg, ed. by D.W.K. Andrews, J. Powell, P.A. Ruud and J.H. Stock. Cambridge: Cambridge University Press.

Campbell, J.Y., Lo, A.W. and MacKinlay, A.C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Hauseman, J., Lo, A. and MacKinlay, C. (1992). An ordered probit analysis of transaction stock prices. *Journal of Financial Economics*, **31**, 319–379.

McCulloch, R.E. and Tsay, R.S. (2000). Nonlinearity in high-frequency data and hierarchical models. Working Paper, GSB, University of Chicago.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance*, **39**, 1127–1140.

Zhang, M.Y., Russell, J.R. and Tsay, R.S. (2001). A nonlinear autoregressive conditional duration model with application to financial transaction data. *Journal of Econometrics*.

Zivot, E. (2005). *Analysis of High Frequency Financial Data: Methods, Models and Software*. 11th School of Time Series and Econometrics. Vila Velha, August 2005. Brazilian Statistical Association.