用R进行高频金融数据分析简介

李洪成

2013R用户会议

大纲

- 高频数据分析简介
- R的高频数据分析包
- 示例

高频数据

金融市场中， 逐笔交易数据(transaction by transaction data) 或逐秒记录数据 (tick by tick data)被称为高频数据。纽约股票交易所的交易行情数据库包含了综合磁带系统报告的所有证券的交易和报价记录(Trades and Quotes- NYSE TAQ), 另外WRDS TAQ, Reuters, Bloomberg等。

高频数据的特点

- 数据量大：一只股票一天中可以有几百万条交易。
- 交易间的时间间隔是不规则的，不是等间隔
- 保存的数据由于多种原因会包含错误
  - 不正确的交易量
  - 失时效的价格
  - 一秒钟的多重交易
  - 不准确的时间(innaccurate times)

某股票2010年10月4日到10月15日
相邻两个交易的价格变动频率

美分	<-2	[-2,-1)	[-1,0)	0	(0,1]	(1,2]	>2
频数	540	1794	55325	304067	54860	1711	558
百分比	0.128	0.428	13.209	72.595	13.098	0.408	0.132

- 高频金融数据用于研究与交易过程和市场微观结构相关的大量问题
  - 股票买卖报价的动态性
  - 市场的流动性
  - 算法交易
  - 收益的实际波动率

Data Structure of Trade data

- PRICE 交易价格
- SIZE 交易股数
- COND: 交易条件代码
- CORR: 修改标识, 交易为正常即没有经过校正、修改或者被标记为取消
- G127 Combined "G", Rule 127, and stopped stock trade

```
> head(sample_tdataraw)
                     SYMBOL EX   PRICE    SIZE      COND CR  G127
2008-01-04 17:30:26 "XXX"  "N" "193.76" "345050" "O"  "0" "0"
2008-01-04 17:30:27 "XXX"  "N" "193.82" "100"    "E"  "0" "0"
2008-01-04 17:30:27 "XXX"  "N" "193.82" "400"    "E"  "0" "0"
2008-01-04 17:30:27 "XXX"  "N" "193.82" "50"     "E"  "0" "0"
2008-01-04 17:30:27 "XXX"  "N" "193.82" "50"     "E"  "0" "0"
2008-01-04 17:30:27 "XXX"  "N" "193.82" "50"     "E"  "0" "0"
2008-01-04 17:30:27 "XXX"  "D" "193.76" "150"    "4"  "0" "0"
2008-01-04 17:30:27 "XXX"  "D" "193.76" "300"    "4"  "0" "0"
2008-01-04 17:30:27 "XXX"  "D" "193.76" "50"     "4"  "0" "0"
2008-01-04 17:30:27 "XXX"  "D" "193.76" "200"    "4"  "0" "0"
2008-01-04 17:30:27 "XXX"  "D" "193.76" "750"    "4"  "0" "0"
2008-01-04 17:30:28 "XXX"  "D" "193.76" "250"    "4"  "0" "0"
2008-01-04 17:30:28 "XXX"  "D" "193.76" "150"    "4"  "0" "0"
2008-01-04 17:30:28 "XXX"  "D" "193.76" "50"     "4"  "0" "0"
2008-01-04 17:30:28 "XXX"  "T" "193.3"  "100"    "F"  "0" "0"
2008-01-04 17:30:28 "XXX"  "D" "193.76" "50"     "4"  "0" "0"
2008-01-04 17:30:28 "XXX"  "D" "193.76" "50"     "4"  "0" "0"
2008-01-04 17:30:28 "XXX"  "N" "193.59" "50"     "E"  "0" "0"
2008-01-04 17:30:28 "XXX"  "N" "193.59" "50"     "E"  "0" "0"
2008-01-04 17:30:28 "XXX"  "N" "193.59" "50"     "E"  "0" "0"
2008-01-04 17:30:28 "XXX"  "D" "193.59" "50"     "@"  "0" "0"
2008-01-04 17:30:28 "XXX"  "T" "193.47" "100"    "F"  "0" "0"
2008-01-04 17:30:28 "XXX"  "N" "193.42" "50"     "E"  "0" "0"
2008-01-04 17:30:29 "XXX"  "N" "193.47" "50"     "E"  "0" "0"
2008-01-04 17:30:29 "XXX"  "N" "193.47" "50"     "E"  "0" "0"
2008-01-04 17:30:29 "XXX"  "D" "193.57" "50"     "@"  "0" "0"
2008-01-04 17:30:29 "XXX"  "N" "193.42" "50"     "E"  "0" "0"
2008-01-04 17:30:29 "XXX"  "N" "193.35" "50"     "E"  "0" "0"
```

Data Structure of Quote data

- BID：卖价
- BIDSIZ:卖出量，以100股为单位
- OFR：买价
- OFRSIZ：买入量
- MODE：报价条件标识

```
> head(sample_qdataraw)
                     SYMBOL EX  BID      BIDSIZ OFR      OFRSIZ MODE
2008-01-04 17:30:00 "XXX"  "T" "193.12" "0.5"  "193.94" "0.5"  "12"
2008-01-04 17:30:26 "XXX"  "P" "193.31" "0.5"  "193.96" "2.5"  "12"
2008-01-04 17:30:26 "XXX"  "N" "193.18" "2"    "193.82" "10"   "12"
2008-01-04 17:30:26 "XXX"  "T" "193.52" "0.5"  "193.97" "0.5"  "12"
2008-01-04 17:30:26 "XXX"  "T" "193.47" "2"    "193.97" "0.5"  "12"
2008-01-04 17:30:26 "XXX"  "N" "193.5"  "2.5"  "193.96" "1.5"  "12"
```

R的高频数据分析包

- R中针对高频数据的添加包：highfrequency
- 该包最新版本为0.2，基于R 2.12.0或者更高版本,依赖于 xts, zoo两个包。
- highfrequency是另外两个已有R包的更新版
  - RTAQ (Cornelissen and Boudt 2012) TradeAnalytics project
  - realized (Payseur 2008).

Highfrequency主要功能

- 组织高频数据
- 高频数据的清理、整理
- 高频数据的汇总
- 高频数据的相关模型：
  - 波动率模型
  - 流动性

输入数据

- 三类高频数据
  - NYSE TAQ数据库中的 .txt文件
  - WRDS数据库中的 .csv文件
  - Tickdata.com的.asc文件
- 函数convert()可以把上述三类数据转换为xts对象

```
convert(from, to, datasource, datadestination,
    trades=TRUE,quotes=FALSE,
    ticker=c("AA","AAPL"), dir=TRUE, extension="txt",
    header=FALSE,tradecolnames=NULL,
    quotecolnames=NULL, format="%Y%m%d
    %H:%M:%S");
```

把数据从硬盘载入R中

- 函数TAQLoad把数据载入R中

```
>   xts_data = TAQLoad( tickers="IBM", from="2011-12-01",
+              to="2011-12-02",trades=F,
+              quotes=TRUE, datasource=datadestination)
```

高频数据的处理

Function	Function Description
	All Data:
ExchangeHoursOnly	Restrict data to exchange hours
selectexchange	Restrict data to specific exchange
	Trade Data:
noZeroPrices	Delete entries with zero prices
autoSelectExchangeTrades	Restrict data to exchange with highest trade volume
salesCondition	Delete entries with abnormal Sale Condition
mergeTradesSameTimestamp	Delete entries with same time stamp and use median price
rmTradeOutliers	Delete entries with prices above/below ask/bid +/- bid/ask spread
	Quote Data:
noZeroQuotes	Delete entries with zero quotes
autoSelectExchangeQuotes	Restrict data to exchange with highest bidsize + offersize
mergeQuotesSameTimestamp	Delete entries with same time stamp and use median quotes
rmNegativeSpread	Delete entries with negative spreads
rmLargeSpread	Delete entries if spread > maxi*median daily spread
rmOutliers	Delete entries for which the mid-quote is outlying with respect to surrounding entries
Wrapper cleanup functions (perform sequentially the following for on-disk data)	
tradesCleanup	noZeroPrices, selectExchange, salesCondition, mergeTradesSameTimestamp.
quotesCleanup	noZeroQuotes, selectExchange, rmLargeSpread, mergeQuotesSameTimestamp rmOutliers
tradesCleanupFinal	rmTradeOutliers (based on cleaned quote data as well)

等间隔数据、数据同步

- aggregatets(data,on="minutes",k=1)
- refreshTime(list(stock1,stock2))

Realized volatility measures

Estimator	Univariate	Multivariate	Jump robust	Microstructure noise robust	Tick-by-tick returns as input	Positive semidefinite
medRV (Andersen *et al.* 2012)	x		x			/
minRV (Andersen *et al.* 2012)	x		x			/
rCov (Andersen *et al.* 2003)	x	x				x
rBPCov (Barndorff-Nielsen and Shephard 2004)	x	x	x			
rOWCov (Boudt *et al.* 2011a)	x	x	x			x
rThresholdCov (Gobbi and Mancini 2009)		x	x			x
rTSCov (Zhang 2011)	x	x		x	x	
rRTSCov (Boudt and Zhang 2010)	x	x	x	x	x	
rAVGCov (Ait-Sahalia *et al.* 2005)	x	x		x	x	x
rKernelCov (Barndorff-Nielsen *et al.* 2004)	x	x		x	x	x
rHYCov (Hayashi and Yoshida 2005)		x			x	

波动率预测

- HAR-模型

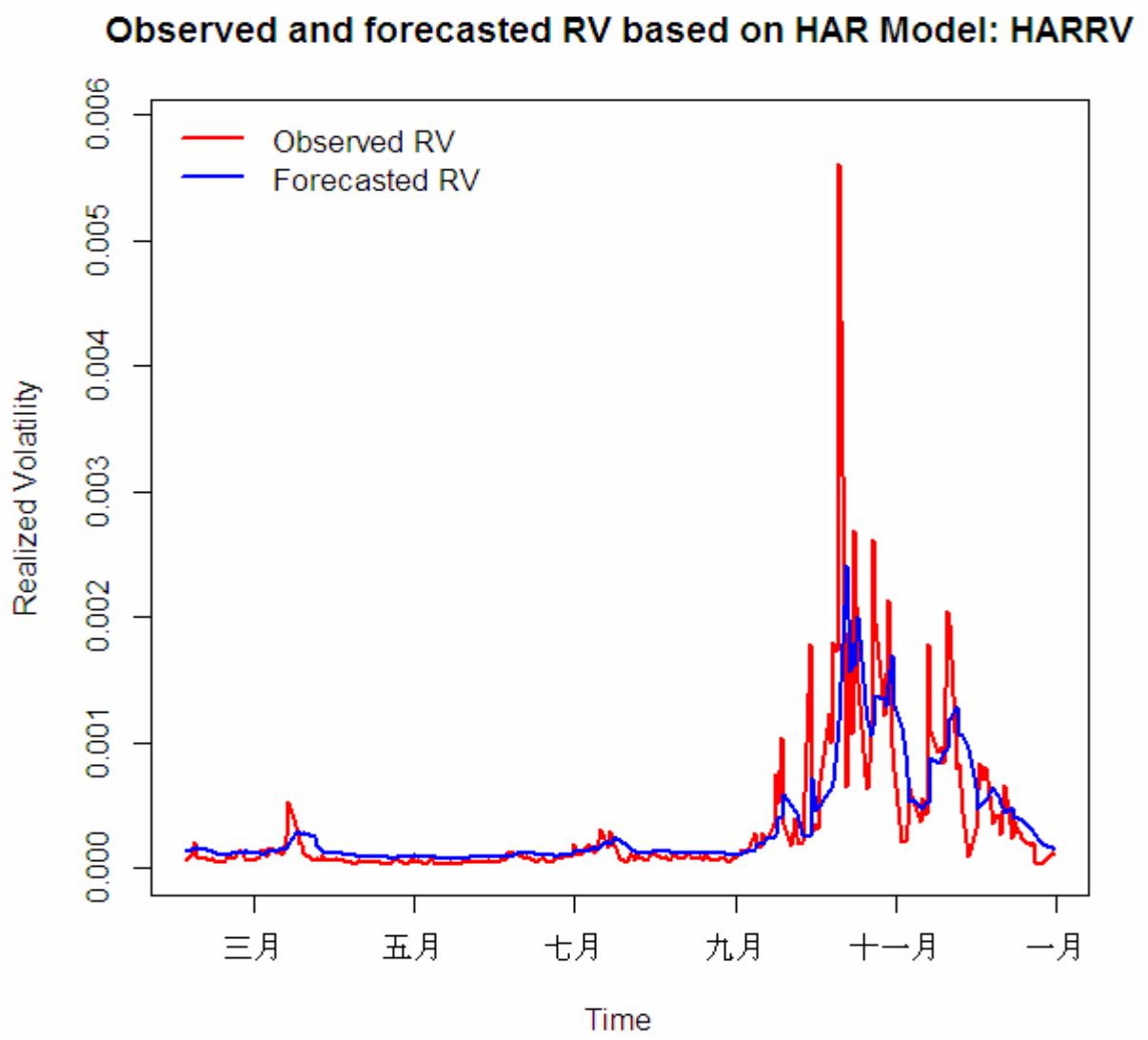
  Heterogeneous Autoregressive

它实现了三种类型的HAR模型：

- HAR-RV：$RV_{t+1} = \beta_0 + \beta_D RV_t + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \epsilon_{t+1}.$
- HAR-RV-J：$RV_{t,t+h} = \beta_0 + \beta_D RV_t + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + J_t + \epsilon_{t,t+h}.$
- HAR-RV-CJ：Jump + Continuous Sample path Component Variation

```
harModel(data, periods = c(1, 5, 22),
    periodsJ = c(1,5,22), leverage=NULL,
    RVest = c("rCov", "rBPCov"), type =
    "HARRV", jumptest = "ABDJumptest",
    alpha = 0.05, h = 1, transform =
    NULL, ...)
```

HAR-RV: Dow Jones Industrial Average in 2008



- HEAVY 模型

$$\mathrm{Var}(r_t|\mathcal{F}_{t-1}^{HF}) = h_t = \omega + \alpha RM_{t-1} + \beta h_{t-1}, \quad \omega, \alpha \geq 0 \text{ and } \beta \in [0,1],$$
$$\mathrm{E}(RM_t|\mathcal{F}_{t-1}^{HF}) = \theta_t = \omega_R + \alpha_R RM_{t-1} + \beta_R \theta_{t-1}, \quad \omega_R, \alpha_R, \beta_R \geq 0 \text{ and } \alpha_R + \beta_R \in [0,1].$$

```
heavyModel(data,
    p=matrix( c(0,0,1,1),ncol=2 ),
    q=matrix( c(1,0,0,1),ncol=2 ),
    startingvalues = NULL, LB = NULL, UB
    = NULL, backcast = NULL, compconst
    = FALSE);
```

高频数据分析的其他方法

- 价格变化模型
  - 顺序概率值模型
  - 分解模型
- 持续期模型
  - 日模式的成分
  - ACD模型
  - 估计