11ª Escola de Séries Temporais e Econometria

Analysis of High Frequency Financial Data: Methods, Models and Software

Eric Zivot
Associate Professor and Gary Waterman Distinguished Scholar, Department of Economics
Adjunct Associate Professor, Department of Finance
University of Washington
August 1, 2005

About Me

- PhD Economics, Yale University, 1992
  - Supervisors: Peter Phillips and Donald Andrews
  - Areas of emphasis: time series econometrics, financial and macro econometrics, Bayesian methods
- Current Research Topics
  - Analysis of high frequency time series
  - Simulation-based estimation of time series models
  - Nonstationary time series, structural change
  - State space models
  - GMM estimation and inference with weak instruments
- Software Development
  - Splus (S+FinMetrics) and R for time series

Agenda

- Lecture 1
  - Introduction to high frequency data
- Lecture 2
  - Realized variance measures: theory
- Lecture 3
  - Realized variance measures: empirical analysis

Lecture 1: Introduction to High Frequency Financial Data

- Introduction and Motivation
- High Frequency Data Sources
- Challenges to Statistical Modeling
- Using S-PLUS for Analyzing High Frequency Data
- Graphical Analysis
- Creating Market Variables
- Descriptive Analysis of High-Frequency Data
- Calendar Patterns in Market Activities
- Statistical Modeling of High Frequency Data

Introduction and Motivation

- What is High-Frequency Financial Data?
  - Ten years ago it was daily data
    - Large data sets consisted of 1000s of stocks over 20-30 years (e.g. Center for Research in Security Prices (CRSP) data.
    - 5 – 10 million observations
  - Now it is tick-by-tick or transaction level data on prices, quotes, volume, order book
    - Large data sets consist of 1000s of stocks over 10-15 years (e.g. New York Stock Exchange (NYSE) TAQ data
    - 1 – 2 billion observations or more

Introduction and Motivation

- Academic Research Topics
  - Market microstructure theory
  - Price discovery and market quality
  - Modeling and estimating liquidity
  - Strategic behavior of market participants
  - Event studies
  - Modeling real-time dynamics of trading process
  - Estimation of continuous-time models
  - Volatility modeling and estimation

Introduction and Motivation

- Finance Industry Applications
  - Short-term trading
    - Pairs trading
    - Arbitrage strategies
    - Event analysis
  - Transaction cost and price impact modeling
    - Order execution
    - Market making
  - Derivatives pricing
    - Continuous-time models
    - Volatility estimation
  - Risk Management

Sources for High Frequency Data

- Historical Data
  - Equity – NYSE TAQ
  - FX – Olsen & Associates
  - Options – Berkeley Options Database
- Commercial Redistributors
  - Wharton Data Services (`wrds.wharton.upenn.edu`)
  - QAI Fast-Tick (`www.qaisoftware.com`)

NYSE Trades and Quotes (TAQ) Database

- Released by NYSE and provides intraday information for stocks traded on NYSE, NASDAQ-AMEX and SmallCap issues starting in 1993. See www.nyse.com/taq.
- TAQ does not include transaction data that is reported *outside* of the Consolidated Tape hours of operation. As of August 2000, those hours are 8:00am to 6:30pm EST. As of March 4, 2004, the tape will open at 4:00am EST. Trading in NYSE-listed securities between 8:00am – 9:30am by other markets are also not in TAQ.

NYSE TAQ Data

- TAQ is available for purchase directly from the New York Stock Exchange. Individual months are available, as well as annual subscriptions. The product is currently delivered on multiple DVD's containing data for one month and is distributed approximately four weeks after the last trading day of each month.
- Substantial academic discounts are available
  - $100 per month for historical data.

NYSE TAQ Data

- **Trade information**: All trades, time-stamped to the second, for all stocks traded on NYSE & regional affiliates, and the NASDAQ-AMEX
  - Do not know trading parties
  - Do not know if trade is buyer or seller initiated
- **Quote information**: all best bid-ask quotes posted by specialists (NYSE, AMEX) and by market makers (NASDAQ) for all stocks

Olsen & Associates FOREX Databases

- Company founded by Richard Olsen
  - Commercial providers of high quality intra-day foreign exchange data
  - Research institute for analysis of high frequency data
- Sponsored three international conferences on the analysis of high frequency financial time series
  - Made available historical data sets
  - www.olsendata.com

Olsen & Associates FOREX Databases

- Indicative (non-binding) dealer quotes on spot exchange rates for wide assortment of currency pairs published over the Reuters network
- 24 hour market
- No transaction or volume information
- Bid/Ask quotes by dealer/institution
- Data are “pre-filtered” using proprietary data cleaning technology (“magic” Olsen filter)

Challenges to Statistical Modeling

- Huge number of observations
  - Can be 20,000 quotes per day for US/EUR!
- Dirty data
- Irregularly spaced observations
- Multiple observations with same time stamp
- Heavy-tailed return distributions
- Long memory behavior
- Strong intra-day and intra-week periodicities
- Variables move in discrete increments
- Data for multiple assets seldom occur at the same time

Limitations of Typical Statistical Software

- Lack flexible time and date handling facilities
- Lack flexible time series graphics capabilities
- Lack functionality for data cleaning
- Lack proper statistical methods
- Lack custom programming capability
- Data set size limitations

Advantages of S-PLUS for High Frequency Data

- Advantages of S-PLUS
  - New big data capabilities in S-PLUS 7
  - Flexible data reading capabilities
  - Flexible and powerful date handling
  - Specialized graphics for time series and big data
  - Easy to create specialized functions
  - Advanced statistical models
- Advantages of S+FinMetrics
  - S-PLUS module with 500+ functions for the econometric modeling and prediction of economic and financial time series
  - Specialized functions for handling time series

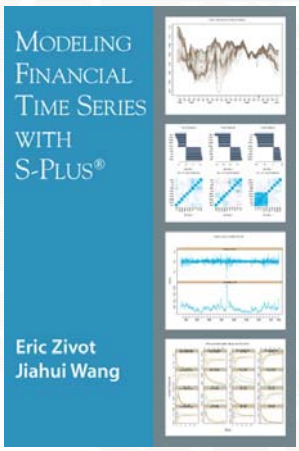
S-PLUS / S+FinMetrics™

Simple Descriptive Tools

Advanced Modeling Tools

- Smoothing & Filtering
- ACF & PACF
- Spectral Analysis
- Aggregation and Seasonal Adjustment
- Technical Analysis & **Fixed Income Analytics**
- ARIMA with Regressors and Long Memory
- Dynamic Time Series Regression
- Tests for **Unit Roots**, Cointegration, **Nonlinearity**
- Extreme Value Distributions and **Copulas**
- **Simulate Solutions to SDEs**
- **Nonlinear regime switching** and neural networks
- General Rolling Estimation
- Seemingly Unrelated Regression
- Vector Autoregression and Cointegration
- GARCH – Univariate and Multivariate
- **State Space Models** and Kalman Filter Tools
- Statistical Factor Models for Large Portfolios
- **Method of Moments Estimation – GMM & EMM**

Documentation for S+FinMetrics™



- New Chapters in *Second Edition*
  - Copulas
  - Nonlinear Models
  - Continuous-Time Models
  - Generalized Method of Moments
  - Semi-nonparametric Conditional Density Models
  - Efficient Method of Moments

HF: S-PLUS Library for Analysis of High Frequency Financial Data

- Yan, B. and E. Zivot (2004). “Analysis of High-Frequency Data with S-PLUS”, Working Paper, Department of Economics, University of Washington
- Paper and library available for download at http://faculty.washington.edu/ezivot
- HF Library is being incorporated into S+FinMetrics 2.1 and will make use of the big data capabilities of S-PLUS 7 Enterprise Developer

Time Series in S-PLUS

- S-PLUS 6.0 “timeSeries” Objects
  - Combines data with “timeDate” object
  - Flexible enough to describe essentially all types of financial time series data
    - Regularly spaced calendar data
    - Irregularly spaced tick-by-tick data
    - Allows time-zone specification
    - Easy event handling
      - Holidays, market closures, etc.
  - Powerful plotting functionality

TAQ Data in ASCII Form

- MSFT: 5/1/97 – 5/15/97 (2 weeks)
  - 98,724 trades; 20,656 quotes
  - Extracted from TAQ DVD to ASCII file

```
cond |ex |symbol |corr |g127 |price |siz |tdate |tseq |ttim |
T  |T  |MSFT  |0  |0  |121.125  |1500  |01MAY1997  |0  |28862  |
T  |T  |MSFT  |0  |0  |121.5625  |500  |01MAY1997  |0  |28944  |
T  |T  |MSFT  |0  |0  |121.5625  |1000  |01MAY1997  |0  |29000  |
T  |T  |MSFT  |0  |0  |121.5625  |1200  |01MAY1997  |0  |29002  |
T  |T  |MSFT  |0  |0  |121.625  |1000  |01MAY1997  |0  |31095  |
```

- ASCII data is imported to S-PLUS data.frame and then converted to S-PLUS timeSeries object using constructor function `timeSeries()`

TAQ Data in S-PLUS

- Representation as `timeSeries` object in S-PLUS

```
> msftt.ts[1:5,]
     Positions Cond  Ex Symbol Corr G127   Price   Size Seq
5/1/1997 8:01:02 T    T   MSFT    0    0    121.1250 1500 0
5/1/1997 8:02:24 T    T   MSFT    0    0    121.5625  500 0
5/1/1997 8:03:20 T    T   MSFT    0    0    121.5625 1000 0
5/1/1997 8:03:22 T    T   MSFT    0    0    121.5625 1200 0
5/1/1997 8:38:15 T    T   MSFT    0    0    121.6250 1000 0
```

Dates are in timeDate object

Data is in a data frame

Olsen Data in S-PLUS

- USD/EUR spot rate quotes: 3/11/2001-3/17/2001 (2 weeks)
  - 126,988 quotes

```
> eurusd.ts[1:5,]
                Positions    Bid    Ask Institution
3/11/2001 22:01:35 0.9326 0.9330 ONEC
3/11/2001 22:01:37 0.9326 0.9331 AREX
3/11/2001 22:09:34 0.9326 0.9331 NWHK
3/11/2001 22:09:36 0.9327 0.9332 AREX
3/11/2001 22:11:08 0.9322 0.9327 NWHK
```

Aligning Time Series

```
> msftt.ts[1:5,"Price"]
       Positions   Price
 5/1/1997 9:30:02 122.000
 5/1/1997 9:30:06 122.125
 5/1/1997 9:30:09 122.000
 5/1/1997 9:30:10 122.000
 5/1/1997 9:30:14 122.125
> msftq.ts[1:5,"Bid"]
       Positions     Bid
 5/1/1997 9:30:01 122.000
 5/1/1997 9:30:06 122.000
 5/1/1997 9:30:13 122.000
 5/1/1997 9:30:14 121.875
 5/1/1997 9:30:17 121.875
```

```
> align.ts =
   align(msftq.ts[,"Bid"],
+ pos = positions(msftt.ts),
+ how = "nearest")

> align.ts[1:5]
        Positions     Bid
 5/1/1997 9:30:02 122.000
 5/1/1997 9:30:06 122.000
 5/1/1997 9:30:09 122.000
 5/1/1997 9:30:10 122.000
 5/1/1997 9:30:14 121.875
```

Other align options: `drop, before, after, interep`

Merging Time Series

```
> msftt.ts[1:5,"Price"]
        Positions   Price
5/1/1997 9:30:02 122.000
5/1/1997 9:30:06 122.125
5/1/1997 9:30:09 122.000
5/1/1997 9:30:10 122.000
5/1/1997 9:30:14 122.125
> msftq.ts[1:5,"Bid"]
        Positions     Bid
5/1/1997 9:30:01 122.000
5/1/1997 9:30:06 122.000
5/1/1997 9:30:13 122.000
5/1/1997 9:30:14 121.875
5/1/1997 9:30:17 121.875
```

```
> merge.ts =
  seriesMerge(msftt.ts[,"Price"],m
  sftq.ts[,"Bid"], how="nearest")

> merge.ts[1:5,]
        Positions   Price     Bid
5/1/1997 9:30:02 122.000 122.000
5/1/1997 9:30:06 122.125 122.000
5/1/1997 9:30:09 122.000 122.000
5/1/1997 9:30:10 122.000 122.000
5/1/1997 9:30:14 122.125 121.875
```

Other merge options: `drop`, `before`, `after`, `interep`, `union`

Aggregating Time Series

```
# compute non-overlapping 5-minute average price

> mean.5min = aggregateSeries(msftt.ts[,"Price"],
+ by="minutes",k.by=5,FUN=mean)

> mean.5min[1:5,]
        Positions     Price
5/1/1997 9:30:00 121.8950
5/1/1997 9:35:00 121.3145
5/1/1997 9:40:00 121.5339
5/1/1997 9:45:00 121.6914
5/1/1997 9:50:00 122.2734
```

Average price between 9:30 and 9:35

How Much Data Can You Analyze in S-PLUS?

- On 32 bit operating systems theoretical limit is 4GB of addressable memory
- On Windows, practical limit is closer to 1.5GB
- S-PLUS memory requirements
  - # of bytes required for data = r*c*8*4.5
  - r = rows, c = columns, 8 = bytes for numeric data, 4.5 = avg # of data copies for modeling functions
  - Ex: Data set with 98,672 rows and 507 columns requires about 1.8 GB memory

Overview of S-PLUS Library HF (Bingchen Yan and Eric Zivot)

- Access data from TAQ and Olsen FxFx databases
- Perform data cleaning and graphical diagnostics
- Define exchange and market time
- Construct market variables
  - Price change, B/A spread, duration, trade direction, realized volatility
- Enhancements to S-PLUS functions `align` and `aggregateSeries` to better handle HF financial data
- Construction of realized variance measures
- Nonparametric estimation of intra-day periodicities

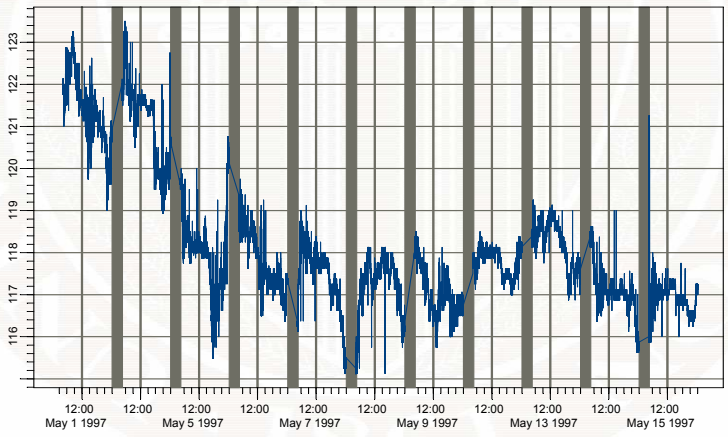
HF Functions

`TAQLoad()`	`tsBW()`
`OlsenLoad()`	`Genr.RealVol()`
`reorderTS()`	`DurationInInterv()`
`plotByDays()`	`PriceChgInInterv()`
`ExchangeHoursOnly()`	`getSpread()`
`FxBizWeekOnly()`	`SmoothAcrossIntervs()`
`align.withinDay()`	`tableSmoother()`
`align.withinWeek()`	`rbindtimeSeries()`
`diff.withinDay()`	`aggregateSeriesHF()`
`diff.withinWeek()`	`tradeDirec()`

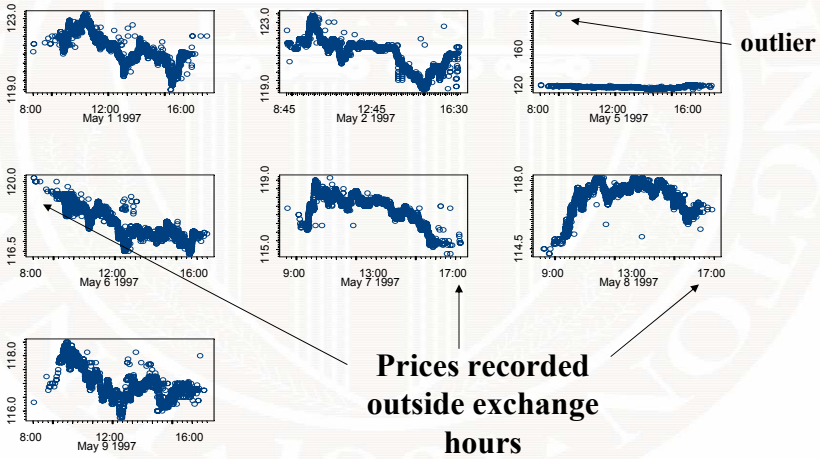
Data Cleaning and Graphics

- Common Data Errors
  - Mis-ordered time-stamps
  - Data recording errors
  - Missing or partial data
  - Time stamps outside of trading hours
- Graphical Diagnositics are Essential!!!
  - Must be careful because large amount of HF data may overwhelm plotting functions
  - HF function `plotByDays()`

MSFT Trade Price: 5/1/97 – 5/15/97



Trade Price by Day



Creating Market Variables

- Price/Quote Changes
  - Price impact analysis
  - Price Discovery
- Durations – time between events
  - Many types of duration
    - Transaction, quote, price, volume
  - Liquidity modeling
- Spreads (Bid/Ask)
  - Market maker behavior
- Trade Direction – Buy/Sell Indicators
  - Demand modeling
- Volatility Measures
  - Derivatives pricing, Value-at-Risk

Complications

- Must separate overnight from intra-day changes
- Restrict data to exchange hours (Equity) or business week (FX)
- Need to deal with holidays, daylight savings times (DST), market closures
- Remove intraday seasonalities (diurnal effects) prior to modeling

Compute Price Changes

```
> msftt.ts = ExchangeHoursOnly(ts = msftt.ts,
+           exch.hours = c("9:30", "16:00"),
+           start.include = T, close.include = T)

> pcTicks.msft = PriceChgInInterv(msftt.ts[, "Price"],
+                ticksize = 1/8,
+                interv.type = "daily",
+                bound.hours = c("9:30", "16:00"))

> pcTicks.msft[1:3]
        Positions Price
5/1/1997 9:30:06  1
5/1/1997 9:30:09 -1
5/1/1997 9:30:10  0
```

Compute Duration Between Trades

```
> duration.msftt = DurationInInterv(x = msftt.ts,
+                  units = "seconds",
+                  interv.type = "daily",
+                  bound.hours = c("9:30", "16:00"))
> duration.msftt[1:5, ]
Positions Duration.in.seconds
5/1/1997 9:30:06 4
5/1/1997 9:30:09 3
5/1/1997 9:30:10 1
5/1/1997 9:30:14 4
5/1/1997 9:30:14 0
```

Compute Bid/Ask Spread

```
> spread.msft = getSpread(ask = msftq.ts[, "Ask"],
+                         bid = msftq.ts[, "Bid"],
+                         ticksize = 1/8)
> spread.msft[1:5, ]
Positions Spread
5/1/1997 9:30:14 1
5/1/1997 9:30:17 2
5/1/1997 9:30:17 1
5/1/1997 9:30:21 1
5/1/1997 9:30:57 1
```

Trade Direction – Buy or Sell Indicator

- TAQ Consolidated Tape does not indicate if transaction is “buyer” or “seller” initiated
- Use Lee-Ready rule to infer trade direction
  - Trade is “buy” if price > mid-quote lagged 5 seconds
  - Trade is “sell” if price < mid-quote lagged 5 seconds
  - Trade is “indeterminate” if price = mid-quote lagged 5 seconds
- Requires merge of Trade and Quote data

Compute Trade Direction

```
> mq.msft = getMidQuote(ask = msftq.ts[,"Ask"],
+                       bid = msftq.ts[, "Bid"])
> trade.direc.msft =
+ tradeDirec(trade = msftt.ts[, "Price"],
+             mq = mq.msft,
+             timeLag = "5s")

> trade.direc.msft[1:5,]
         Positions BuySellDirec
5/1/1997 9:30:02 0
5/1/1997 9:30:06 1
5/1/1997 9:30:09 0
5/1/1997 9:30:10 0
5/1/1997 9:30:14 1
```

Compute Realized Volatility

- p_t= log-price of asset at time t (aligned to common clock)
- Δ= fraction of a trading session associated with the implied sampling frequency,
- m=1/Δ= number of sampled observations per trading session
- Intra-day continuously compounded (cc) returns from time t to t+Δ

$$r_{t+\Delta} = p_{t+\Delta} - p_t$$

Compute Realized Volatility

- Daily Realized Variance

$$RV_t = \sum_{j=1}^{m} r_{t-1+j\Delta}^2$$

- Daily Realized Volatility

$$RVOL_t = \sqrt{RV_t}$$

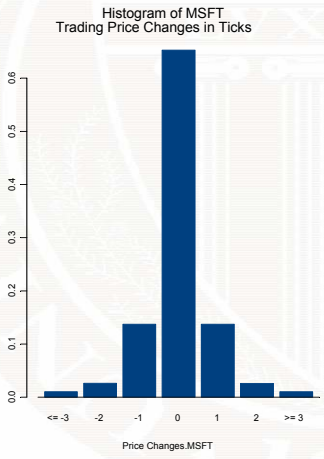
Compute Daily Realized Volatility from 5-Minute Equity Returns

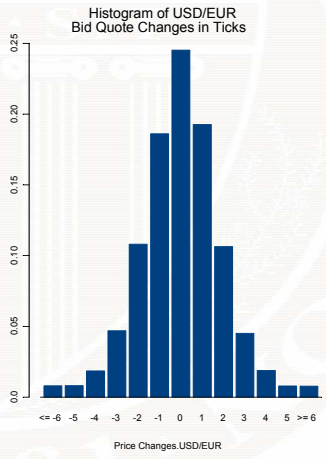
```
> rvDaily.msft =
+ Genr.RealVol(ts = log(msftt.ts[, "Price"])*100,
                interv.type = "daily",
+               bound.hours = c("9:30", "16:00"),
+               rv.span = timeSpan("6h30m"),
+               rt.span = timeSpan("5m"))

> rvDaily.msft[1:5,]
         Positions RealizedVol
 5/1/1997 16:00:00 2.149662
 5/2/1997 16:00:00 1.869500
 5/5/1997 16:00:00 2.357502
 5/6/1997 16:00:00 2.232159
 5/7/1997 16:00:00 2.215329
```

Descriptive Analysis of High Frequency Data

- Price changes of transaction prices and quotes are discrete valued variables, only taking values in multiples of tick sizes.
- There is tendency for price reversal, or bid-ask bounce in transaction price changes.
- Typically during active trading periods, several trades or quotes may appear to occur at the “same” time and share the same time stamp. Consequently, there may be a significant fraction of transactions with zero durations.
- Prices are often recorded at regular intervals (e.g. every 5 minutes) but not all assets trade at the same time or with the same frequency. This may cause cross correlation between returns, serial correlation in portfolio returns and negative serial correlation in individual returns.

Descriptive Analysis: Price Change





Serial Correlation and Bid-Ask Bounce

- Result: Bid-Ask spread introduces negative lag-1 serial correlation in an asset return
- Intuition comes from Roll's (1984) model

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

$P_t^* =$ constant fundamental value independent of S

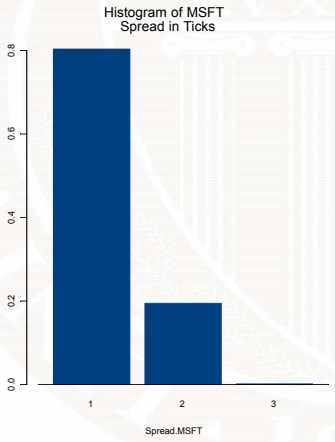
$$S = P_{Ask} - P_{Bid}$$

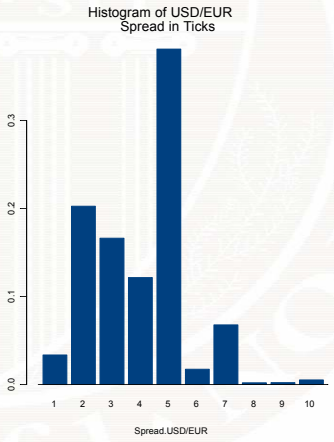
$$I_t = \begin{cases} 1 \text{ with probability } 0.5 \\ -1 \text{ with probability } 0.5 \end{cases}$$

Descriptive Analysis: Price Change

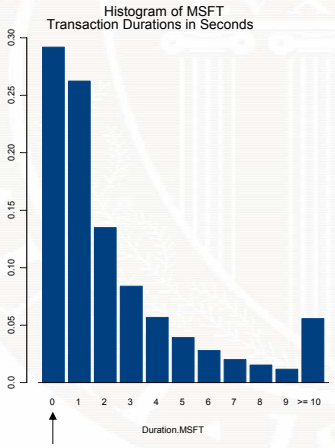
MSFT Price Changes	*i*th Trade		
(*i*-1)th Trade	"+"	"0"	"-"
"+"	787	8058	8020
"0"	8449	46869	8077
"-"	7630	8468	757

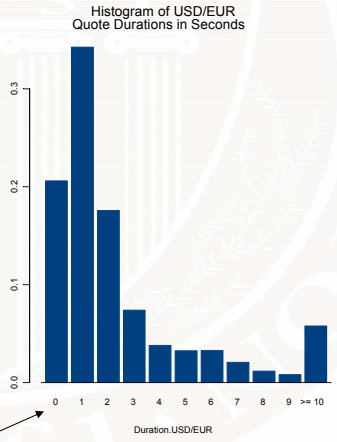
Descriptive Analysis: Spread





Descriptive Analysis: Duration





Note frequency of zero durations!

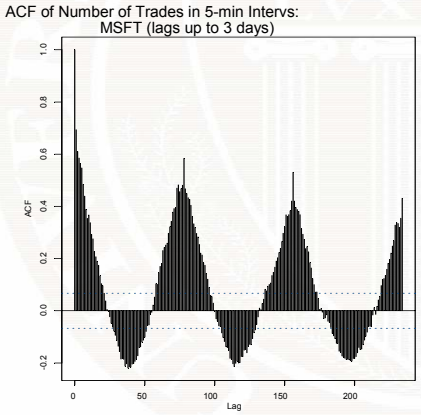
Calendar Patterns in High Frequency Data

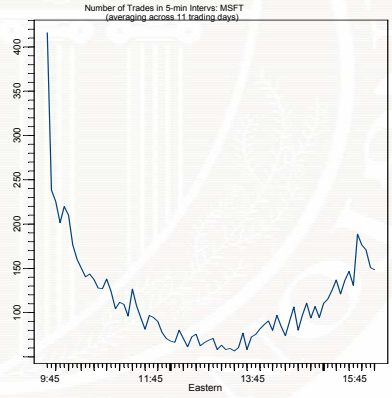
- Intraday calendar patterns (diurnal effects) have been found in the volatility of asset prices, transaction volumes,tick frequency, duration between ticks, and bid/ask spreads
- Equity activity variables, except duration, follow a reserve J-shaped pattern over trading hours. Duration follows an inverted U shape
- FX trading activities also follow an intra-day calendar pattern with three peaks corresponding to the business hours of three geographical trading centers (i.e. Asian, European, and American).

Nonparametric Estimation of Diurnal Effects

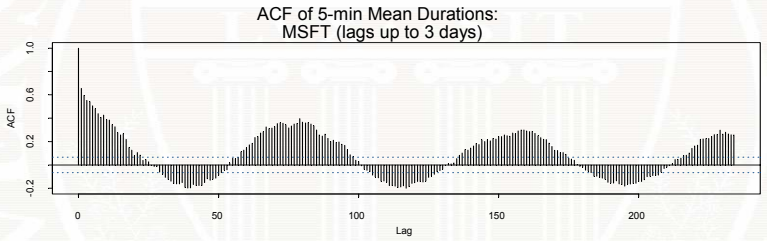
- Deterministic diurnal effects can be estimated by smoothing or averaging the variable in question across trading days.
- For example, the volatility measures at 9:35 for all of the observed trading days can be averaged to get a smoothed measure of volatility at 9:35. This can be done for all intraday time intervals.
- Alternatively one can use splines or trigonometric polynomials to capture diurnal effects

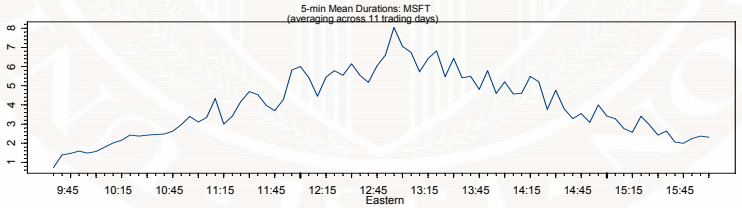
Diurnal Effects in Trading Activity: MSFT Stock





Diurnal Effects in Duration: MSFT Transactions

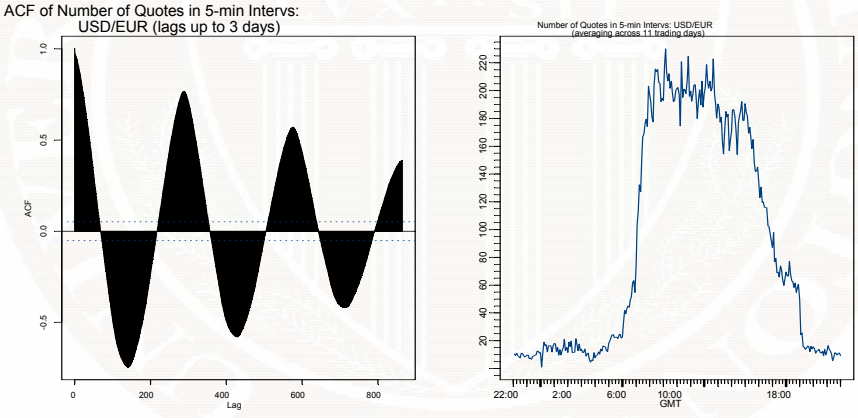




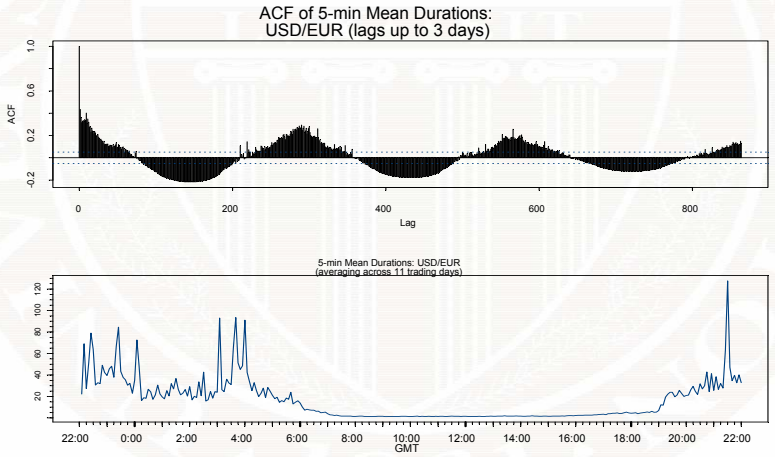
Intraday Trading Sessions for 24 Hour FX Market

	Asian	European	American	Post-American
Hours in GMT	22:00 - 06:00	06:00 - 12:00	12:00 - 18:00	18:00 - 22:00

Diurnal Effects in Quote Activity: USD/EUR



Diurnal Effects in Quote Duration: USD/EUR



Statistical Modeling of High Frequency Data

- Ordered probit model for price/quote changes
- Engle's ACD models for duration
- State space models for fair price extraction
- Cointegration models for pairs trading and price discovery
- Extreme value copula analysis for risk management
- Long memory, structural change and regime switching models for realized volatility

Texbook and Monograph References

- Campbell, J., A. Lo, and C. MacKinlay. ***The Econometrics of Financial Markets***, Princeton University Press, 1997.
- Tsay, R. ***Analysis of Financial Time Series***, John Wiley & Sons, 2002.
- Gourerioux, C., J. Jasiak. ***Financial Econometrics***, Princeton University Press, 2001.
- Dacarogna, M., M. Gencay, U.A. Muller, R. Olsen, O.V. Pictet. ***An Introduction to High Frequency Finance***, Academic Press, 2001.
- Bauwens, L., P. Giot. ***Econometric Modeling of Stock Market Intraday Activity***. Kluwer, 2001.
- Hasbrouck, J. ***Empirical Analysis of Market Micro-Structure***, Lecture notes, New York University, 2004.