COMPUTATIONAL FINANCE & RISK MANAGEMENT

UNIVERSITY of WASHINGTON

Department of Applied Mathematics

(high frequency) Data

CFRM 522 (005)

Introduction to Trading Systems

Lecture References

- Aldridge Ch 4: High-Frequency Data

- All graphs taken from this chapter are in the text

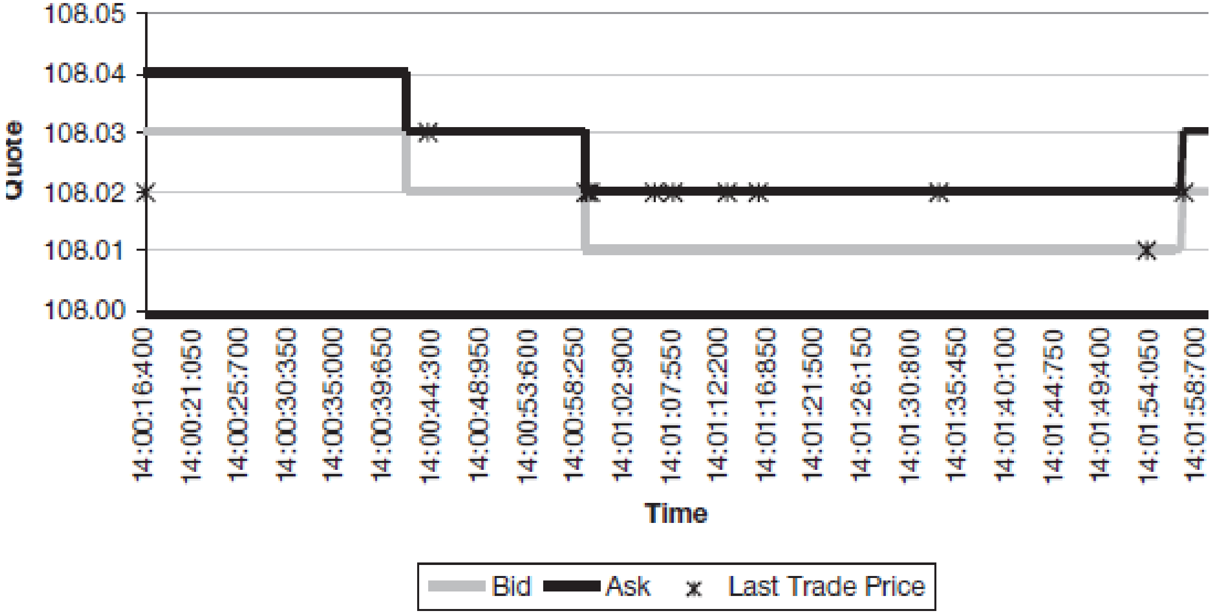
High Frequency Data

- In many ways not much different from other financial data, just more of it (contrary to author’s assertion)
- Two formats (Aldridge, p 53)
  - Level I: best bid price, best ask price, best bid size, best ask size, last trade price and, where available, size
  - Level II: all changes to the order book, including new limit order arrivals and cancellations at prices away from the market price.
- For working with strategies in this class, we will mainly be working with “bars” of data (eg, daily, 1 min, 30 min etc) that contain
  - Open/High/Low/Close price for each bar (“OHLC”)
  - Volume for each bar
  - Volume-Weighted Average Price (in some cases)

High Frequency Data

- Tick data: updated with each new highest bid price, lowest ask price etc, at the time *t* it arrives. Usually consists of:
  - Timestamp
  - Financial Security Information Code (SIC)
  - (Highest) bid/(Lowest) ask price
  - Available bid/ask size
  - Last trade price and size (aggregate order size for each)
  - May also contain security specific information such as option volatility, expiration date of futures or option contract etc

A. HF Data for S&P 500 ETF Recorded from 14:00:16:400 to 14:02:00:000 GMT: Best Bid, Best Ask, and Last Trade Data

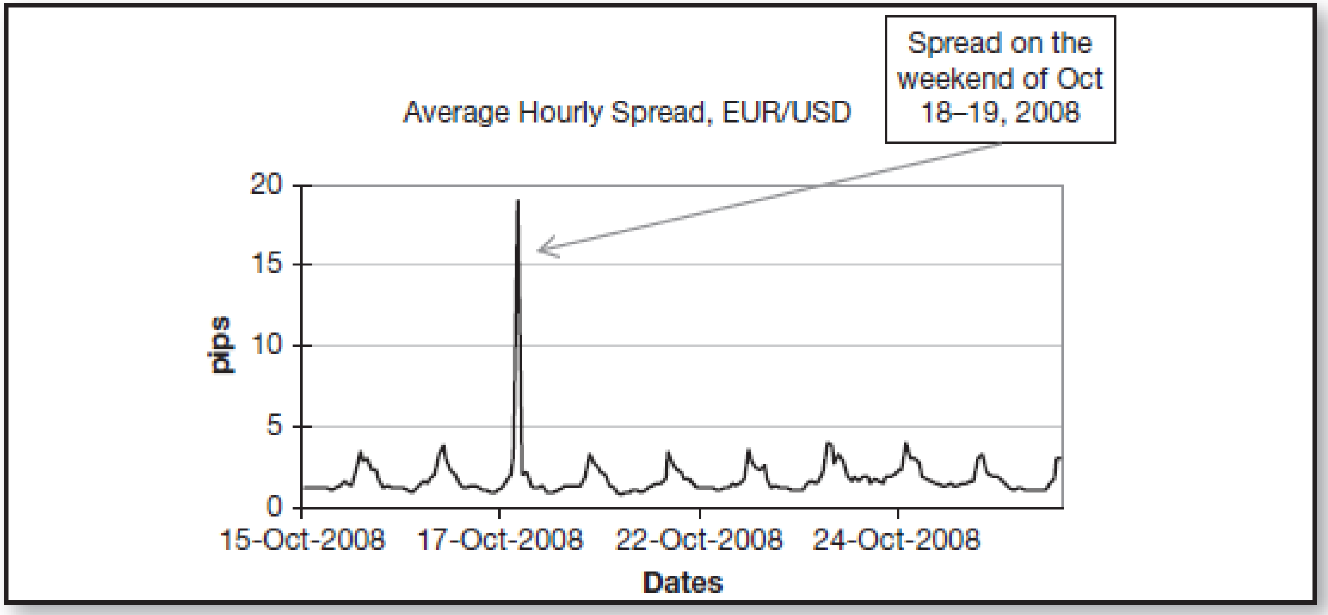


Properties of High-Frequency Data

- Voluminous (no doubt)
- Bid/Ask bounce: carries information about market movement
- Not normally or lognormally distributed (same for most other financial data, although the assumption is often imposed in traditional models)
- Irregularly spaced in time: Durations between data arrival also carry market information
- Does not include buy or sell trade direction information
  - Also same for a lot of other market data
  - However, with high-frequency data, predictive analytics can be employed to assess the probability of an up or down next move

Bid-Ask Spread

- EUR/USD hourly spreads around the 2008 Lehman Crisis (fig 4.2, p 59, Aldridge):



- Can indicate (certainly in hindsight) something's happening
- Forecasting method for bid-ask spread (Roll, 1984), p 61
- A "pip" means "percentage in point", used in FX trading
- For most currencies, 1 pip = 0.0001 ($\frac{1}{100}$ of one cent for USD, CAD)
- Japanese Yen, 1 pip = 0.01

More Properties of HF Data

- Normality/Lognormality assumption
  - Goes out the window for HF data
  - However, a strong assumption in non-HFT models as well
  - Q-Q plots in Ch 4 demonstrate this
- HF data irregularly spaced in time
  - Not an issue with, say, monthly returns in portfolio management
  - The pattern of observations and irregularities themselves can contain useful market information in HFT or algorithmic trading
  - Durations are often modeled using Poisson processes (pp 68-69)

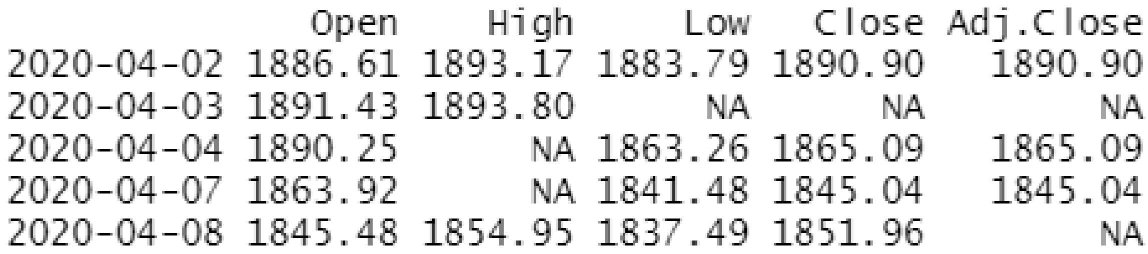
More Properties of HF Data

- Most HF data do not contain buy/sell (direction) information
  - However, estimation methods exist to predict whether a given trade was a buy or sell
  - Four are mentioned in Aldridge:
    - ➢ Tick rule
    - ➢ Quote rule
    - ➢ Lee-Ready rule
    - ➢ Bulk volume classification
  - Described on pp 70-73
  - Essentially predictive analytics problems
- Active area of research

Missing Data

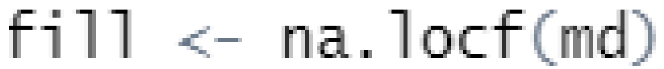
- We will eventually be using xts objects to bring data into backtesting

- Suppose we have missing data:

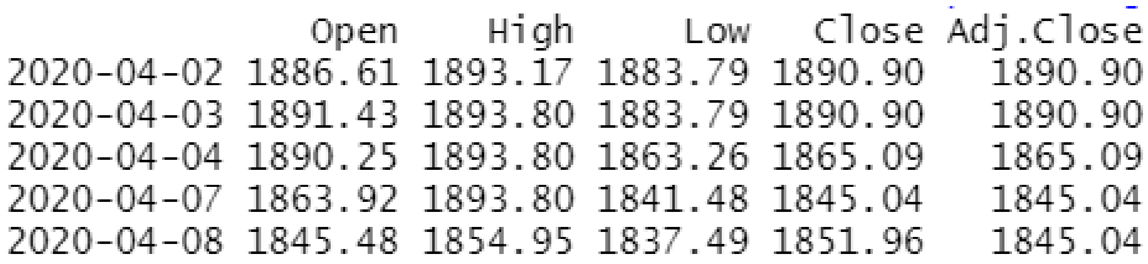


```
              Open    High     Low   Close Adj.Close
2020-04-02 1886.61 1893.17 1883.79 1890.90   1890.90
2020-04-03 1891.43 1893.80      NA      NA        NA
2020-04-04 1890.25      NA 1863.26 1865.09   1865.09
2020-04-07 1863.92      NA 1841.48 1845.04   1845.04
2020-04-08 1845.48 1854.95 1837.49 1851.96        NA
```

- One remedy is to carry forward the preceding data value, using the **`na.locf(.)`** function (overloaded for xts and zoo objects):



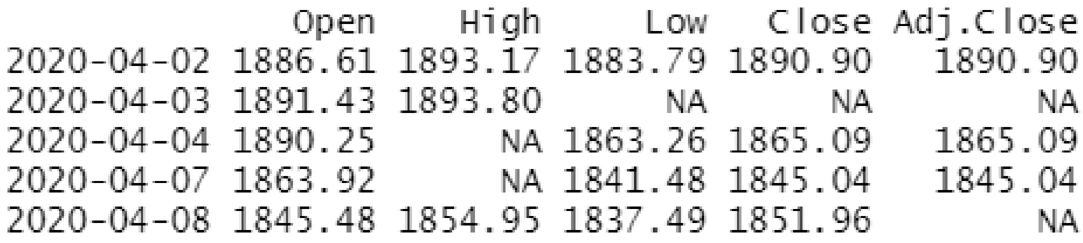
```
fill <- na.locf(md)
```



```
              Open    High     Low   Close Adj.Close
2020-04-02 1886.61 1893.17 1883.79 1890.90   1890.90
2020-04-03 1891.43 1893.80 1883.79 1890.90   1890.90
2020-04-04 1890.25 1893.80 1863.26 1865.09   1865.09
2020-04-07 1863.92 1893.80 1841.48 1845.04   1845.04
2020-04-08 1845.48 1854.95 1837.49 1851.96   1845.04
```

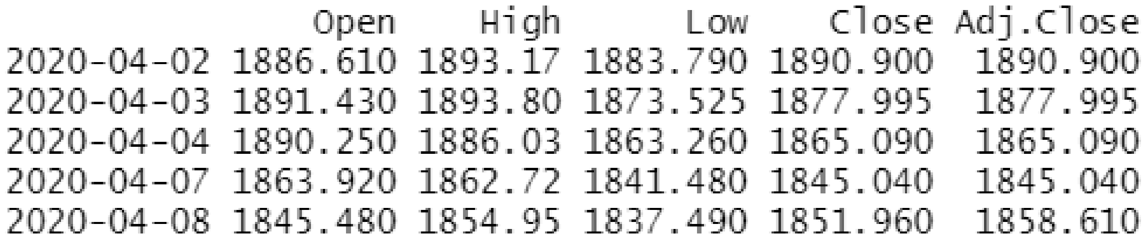
Missing Data

- The missing data:

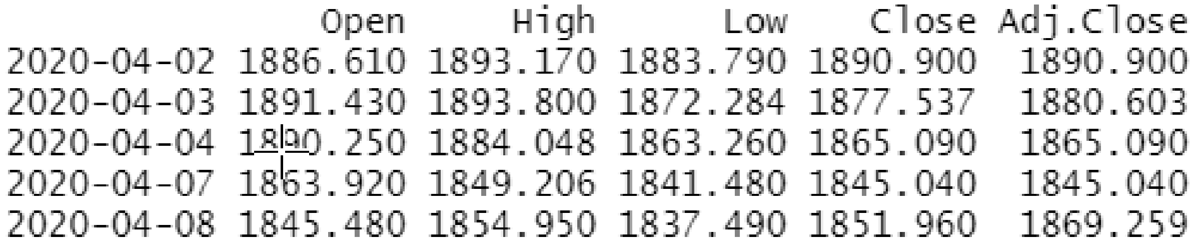


```
             Open    High     Low   Close Adj.Close
2020-04-02 1886.61 1893.17 1883.79 1890.90   1890.90
2020-04-03 1891.43 1893.80      NA      NA        NA
2020-04-04 1890.25      NA 1863.26 1865.09   1865.09
2020-04-07 1863.92      NA 1841.48 1845.04   1845.04
2020-04-08 1845.48 1854.95 1837.49 1851.96        NA
```

- Other remedies include
  - linear interpolation: **`na.approx.(.)`**
  - cubic spline interpolation: **`na.spline(.)`**



```
               Open    High      Low    Close Adj.Close
2020-04-02 1886.610 1893.17 1883.790 1890.900  1890.900
2020-04-03 1891.430 1893.80 1873.525 1877.995  1877.995
2020-04-04 1890.250 1886.03 1863.260 1865.090  1865.090
2020-04-07 1863.920 1862.72 1841.480 1845.040  1845.040
2020-04-08 1845.480 1854.95 1837.490 1851.960  1858.610
```



```
               Open     High      Low    Close Adj.Close
2020-04-02 1886.610 1893.170 1883.790 1890.900  1890.900
2020-04-03 1891.430 1893.800 1872.284 1877.537  1880.603
2020-04-04 1890.250 1884.048 1863.260 1865.090  1865.090
2020-04-07 1863.920 1849.206 1841.480 1845.040  1845.040
2020-04-08 1845.480 1854.950 1837.490 1851.960  1869.259
```

Mr Data

“Only one of us is in the correct time continuum”
Star Trek The Next Generation: *The Manheim Effect*