Mining Time Series

Feb/13/2012

What is Time Series

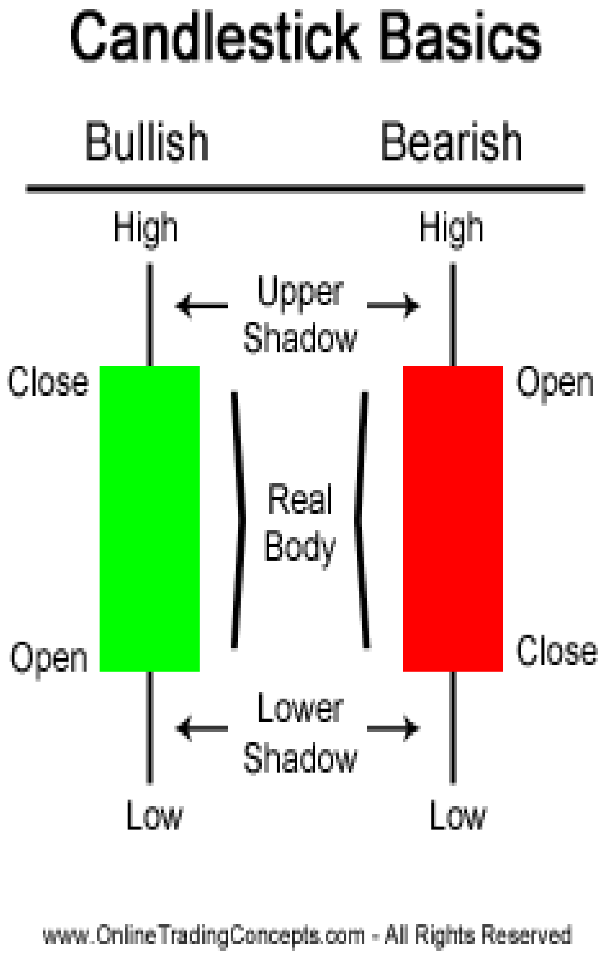
- Time series is a sequence of data points, measured typically at successive time points spaced at uniform time intervals.

- Time series mining comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.

- Time series are frequently plotted via line charts.

Example of Time Series Data

Candlestick Chart



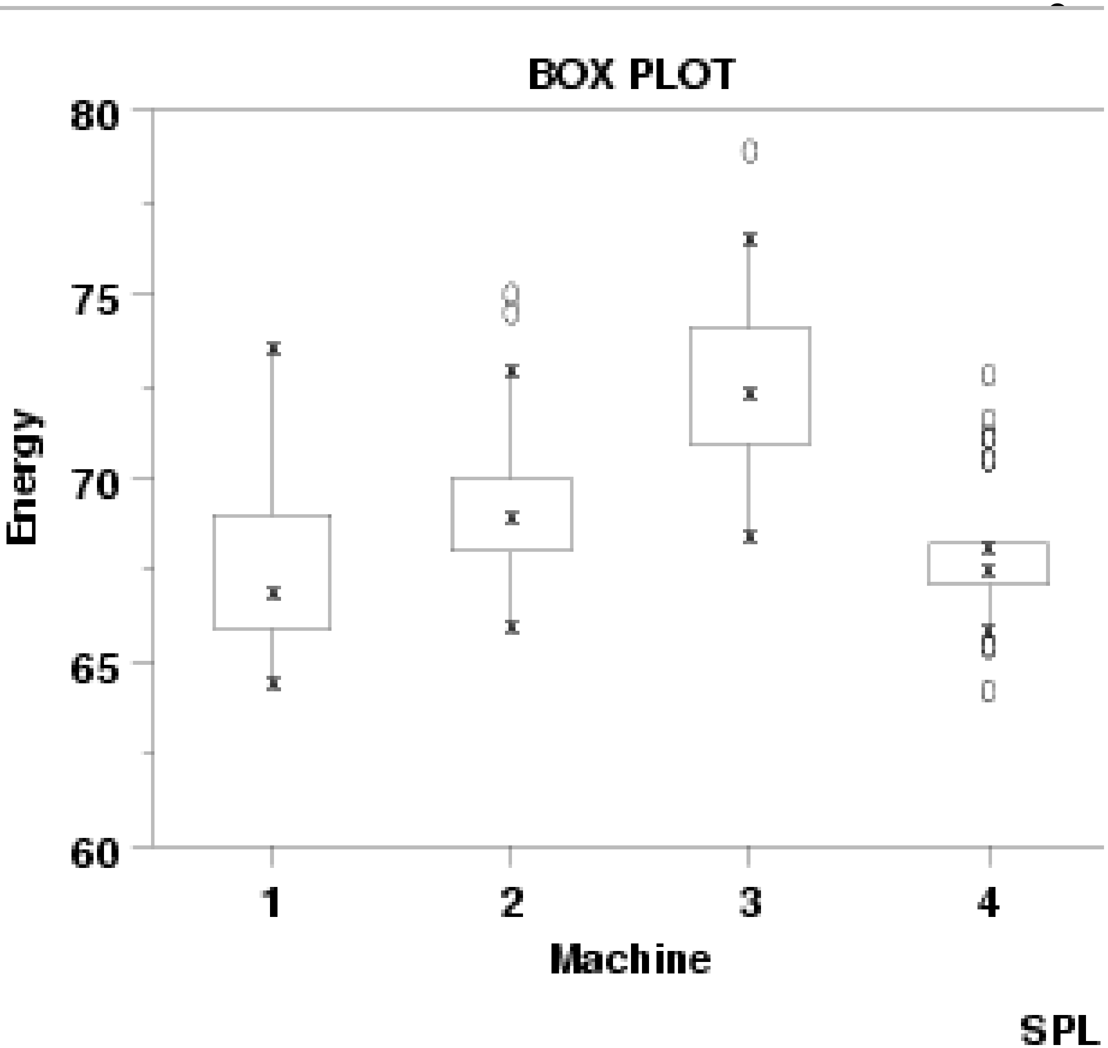
- A candlestick chart is a style of bar-chart used primarily to describe price movements of a security, derivative, or currency over time.

- It is a combination of a line-chart and a bar-chart,

Pre-Processing Steps

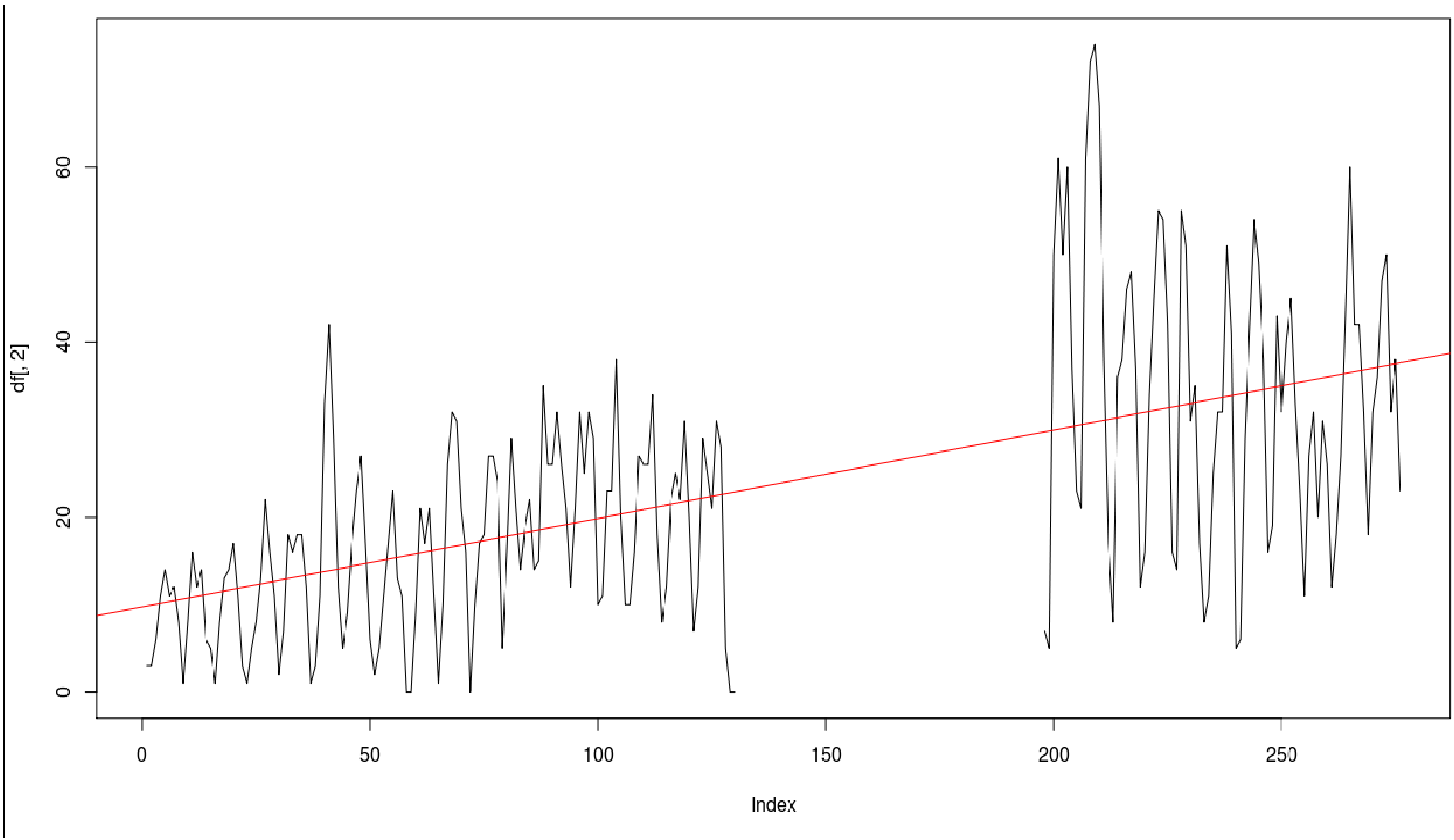
- Visualize the data
- Outlier removal – Box Plot
- Observe Linear Trend – lm(), abline() in R
- Compute Correlation
- Prepare for in-sample testing or back-testing

Box Plot and Outliers



Box plot graphically depicting groups of data through their five-number summaries: the *sample minimum*, *lower quartile*, *median*, *upper quartile*, and *sample maximum*. A boxplot indicates which observations might be considered as outliers.

Linear Trend Observation



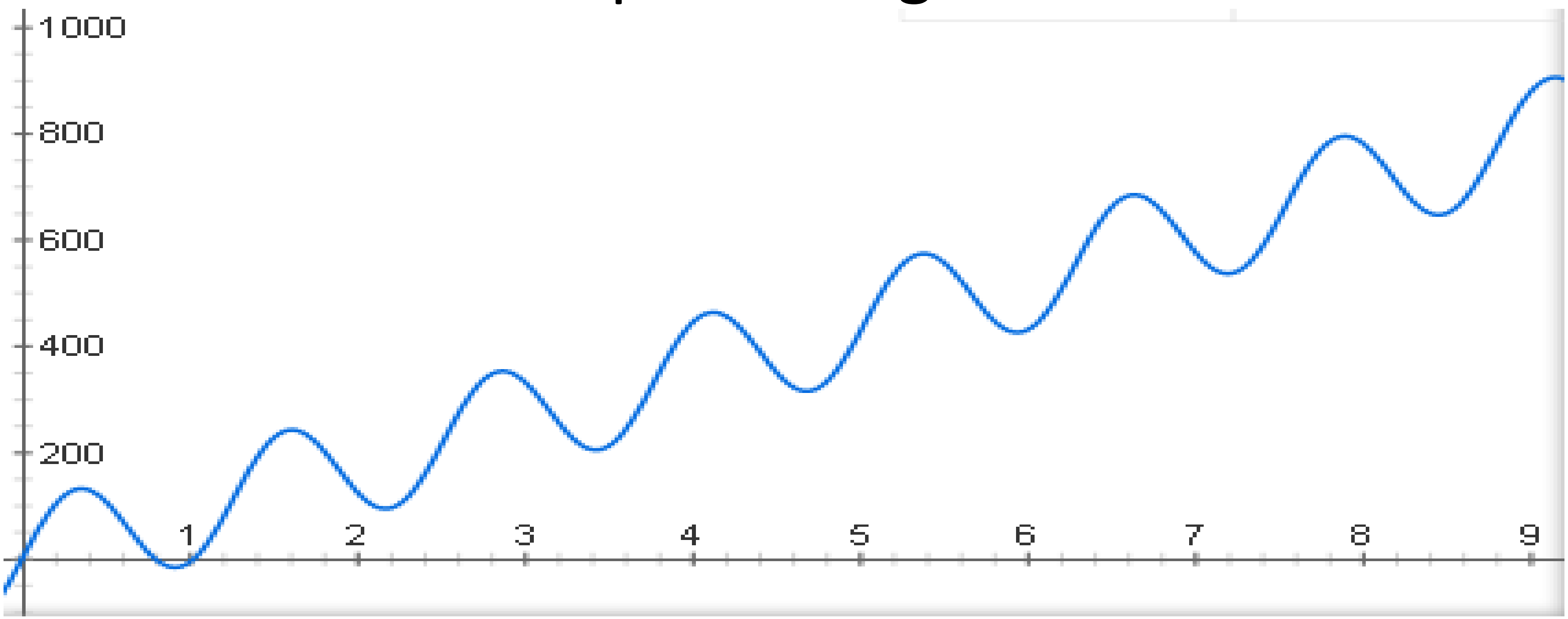
Lm(), abline() functions in R

```
> str(d)
'data.frame':   276 obs. of  1 variable:
 $ D: int  3 3 6 11 14 11 12 8 1 8 ...
> summary(d)
       D
 Min.   : 0.00
 1st Qu.:11.00
 Median :21.00
 Mean   :22.91
 3rd Qu.:32.00
 Max.   :74.00
 NA's   :67.00
>
>
>
> str(d)
'data.frame':   276 obs. of  1 variable:
 $ D: int  3 3 6 11 14 11 12 8 1 8 ...
> head(d)
   D
1  3
2  3
3  6
4 11
5 14
6 11
> d_data_frame<-data.frame(t=c(1:length(d[,1])),x=d[,1])
> str(d_data_frame)
'data.frame':   276 obs. of  2 variables:
 $ t: int  1 2 3 4 5 6 7 8 9 10 ...
 $ x: int  3 3 6 11 14 11 12 8 1 8 ...
> head(d_data_frame)
  t  x
1 1  3
2 2  3
3 3  6
4 4 11
5 5 14
6 6 11
```

```
> trend<-lm(d_data_frame$x~d_data_frame$t,na.action=na.exclude)
> trend$coefficients
  (Intercept) d_data_frame$t
    9.7372240      0.1010691
> trend_line<-predict(trend)
> str(trend_line)
 Named num [1:276] 9.84 9.94 10.04 10.14 10.24 ...
 - attr(*, "names")= chr [1:276] "1" "2" "3" "4" ...
> head(trend_line)
        1         2         3         4         5         6
 9.838293  9.939362 10.040431 10.141500 10.242569 10.343639
> # OR
> # After having the original plot of raw data
> abline(coef=trend$coefficients,col='red')
```

Remove Trend or Not?

- For some techniques, percentage changes of time series data points ought to be calculated



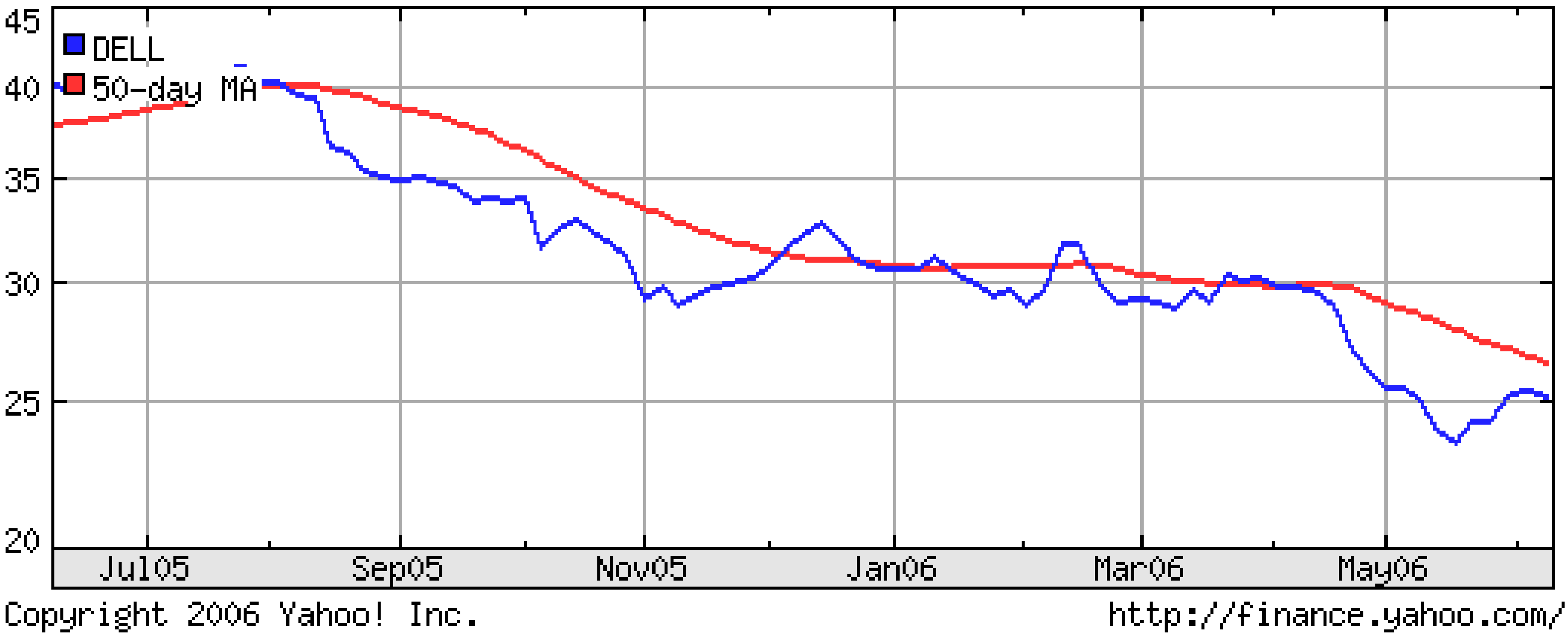
- Same variance over time -> Remove
- Different variance over time -> Further Analysis

Approaches Toward Time Series Mining

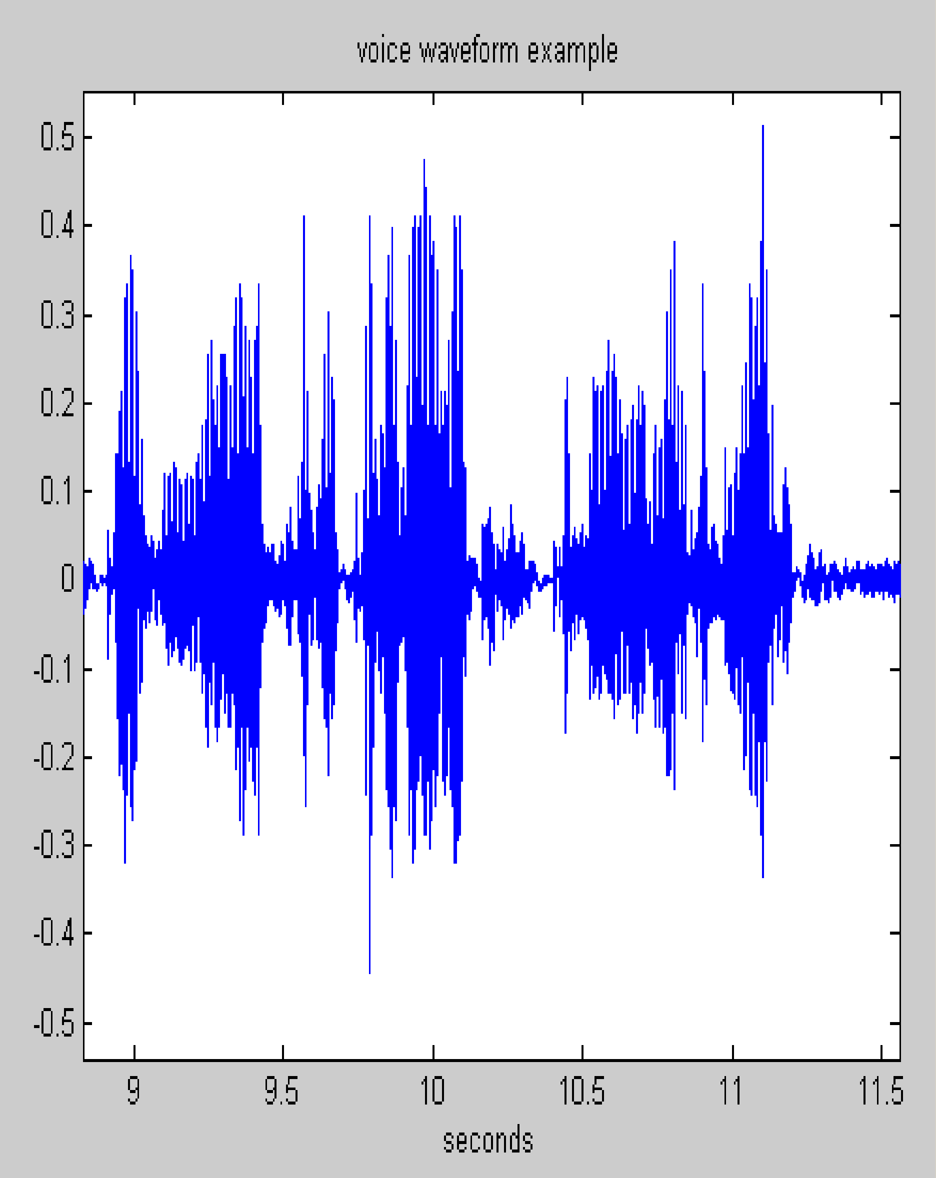
- Signal Processing Approaches (MA, MACD, etc.) – Technical Analysis for finance
- Model Based Approaches (AR, EMM, GARCH, etc.) – Quantitative Analysis for finance

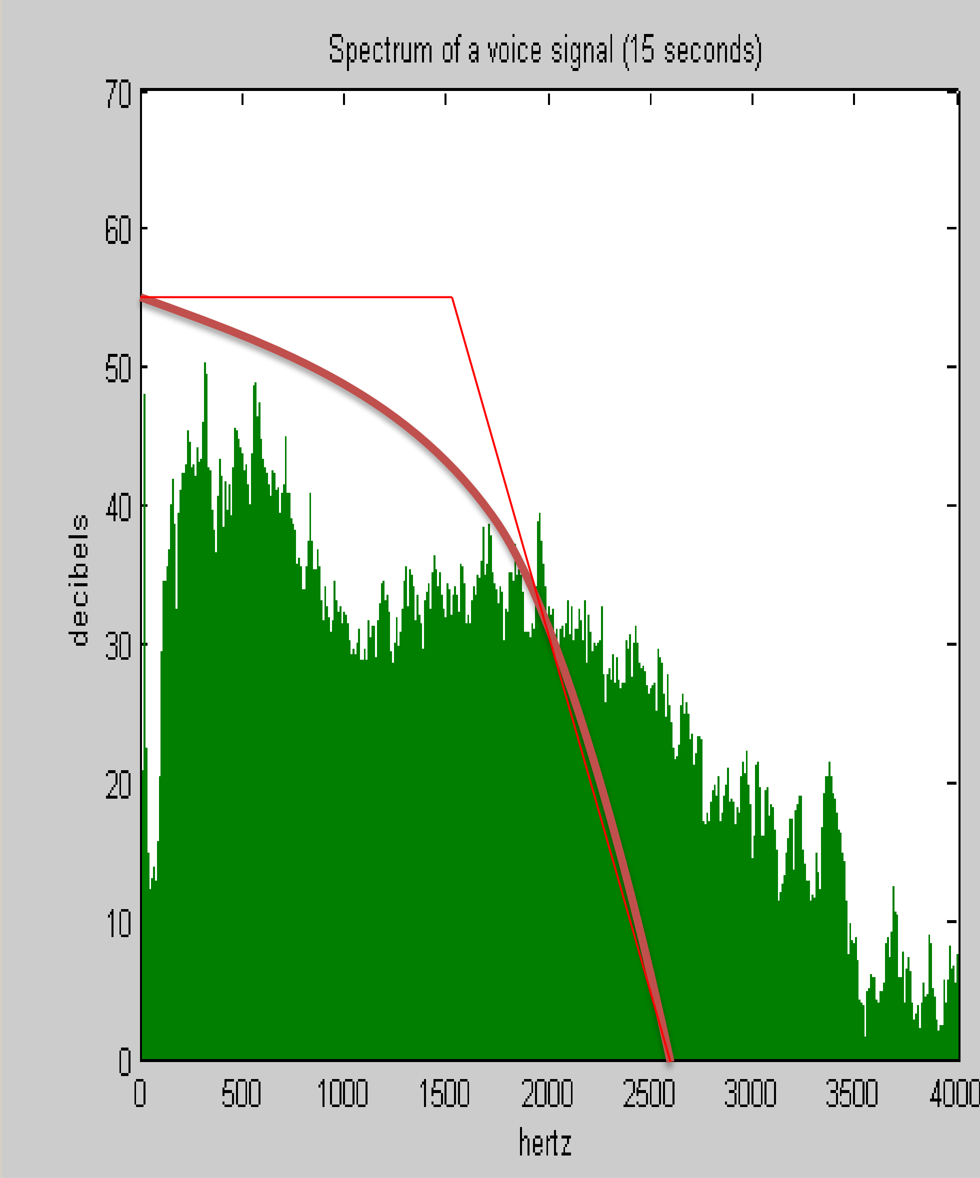
Moving Average (MA)

- Moving average is a type of low pass filter used to analyze a set of data points by creating a series of averages of different subsets of the full data set.



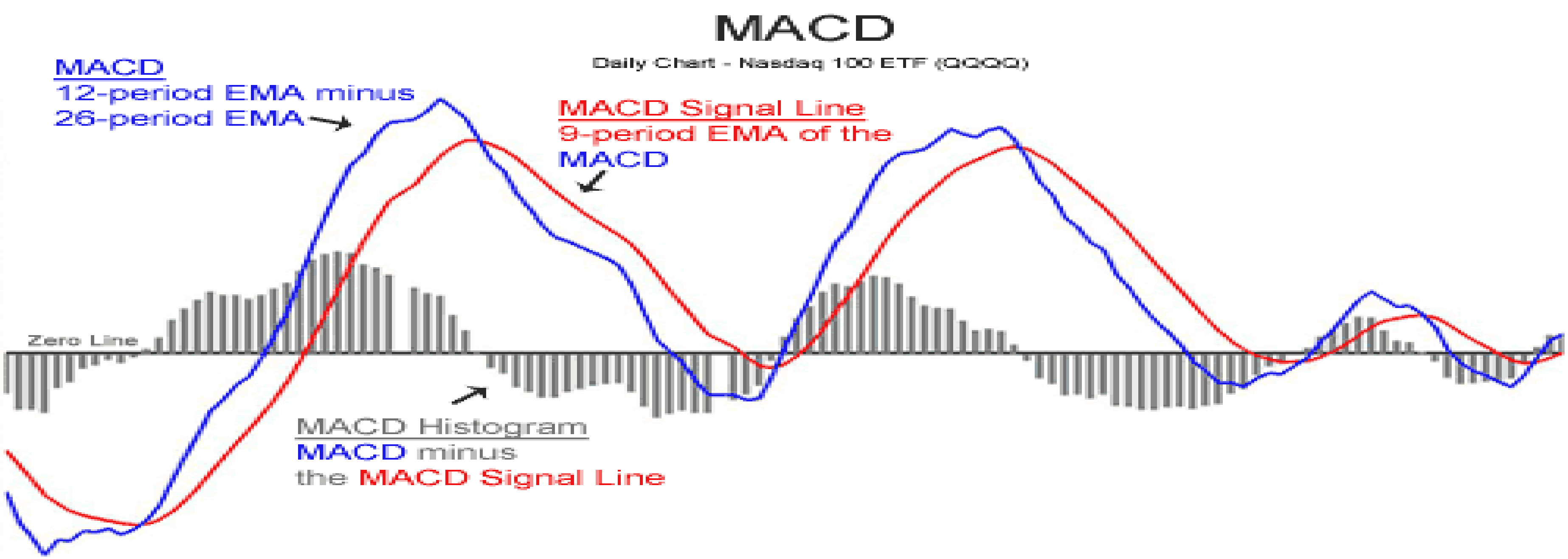
Low Pass Filter (in red)





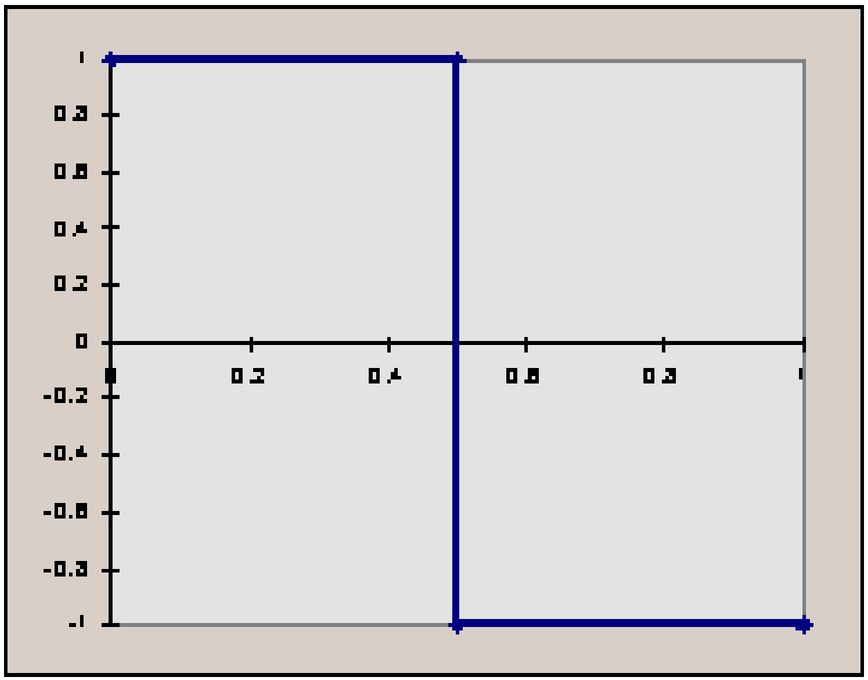
Moving Average Convergence-Divergence (MACD)

- The MACD is a computation of the difference between two moving averages. This difference is charted over time, alongside a moving average as a trigger. The divergence between the two is shown as a histogram or bar graph



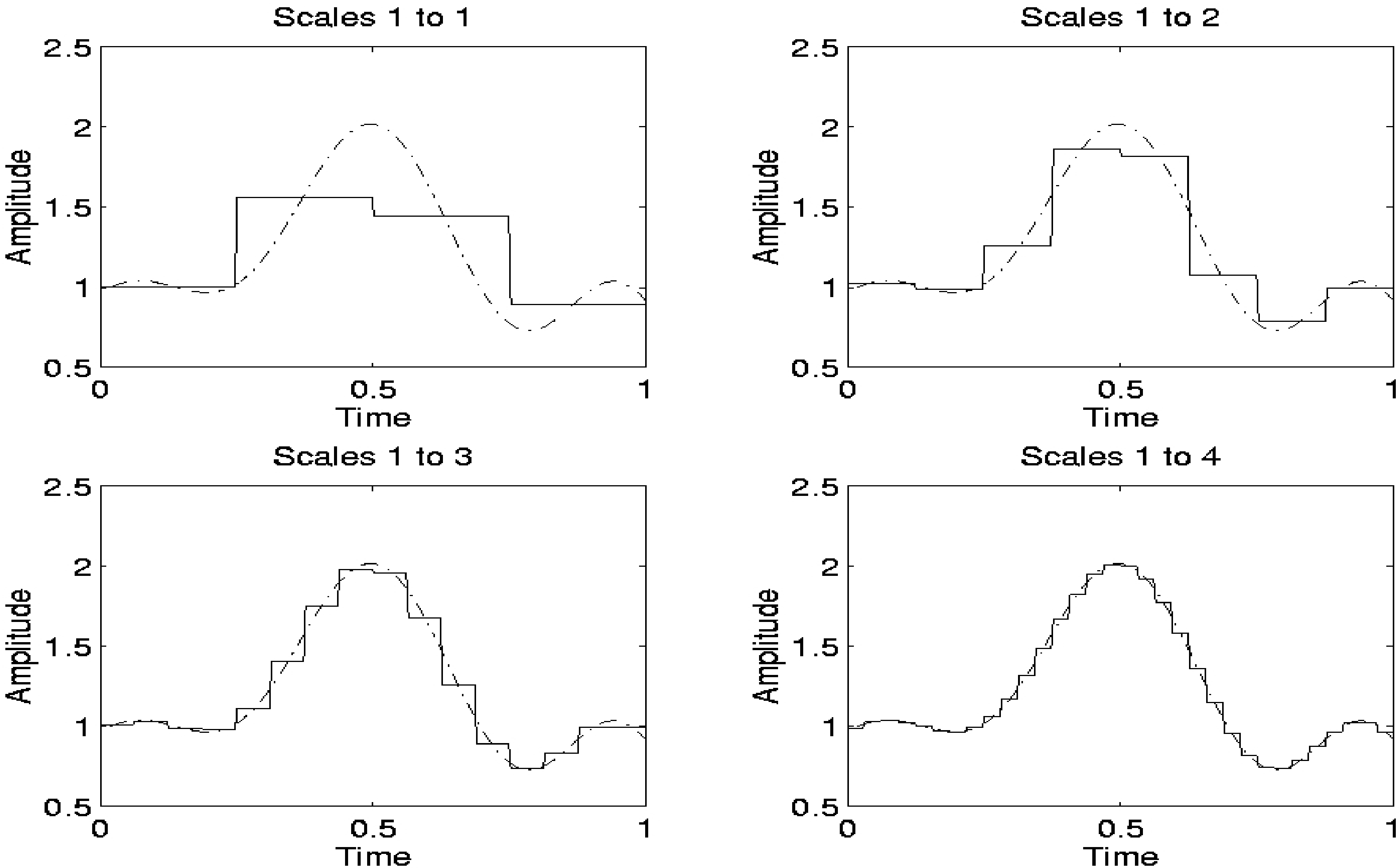
Haar Wavelet Analysis

- Haar Building Block



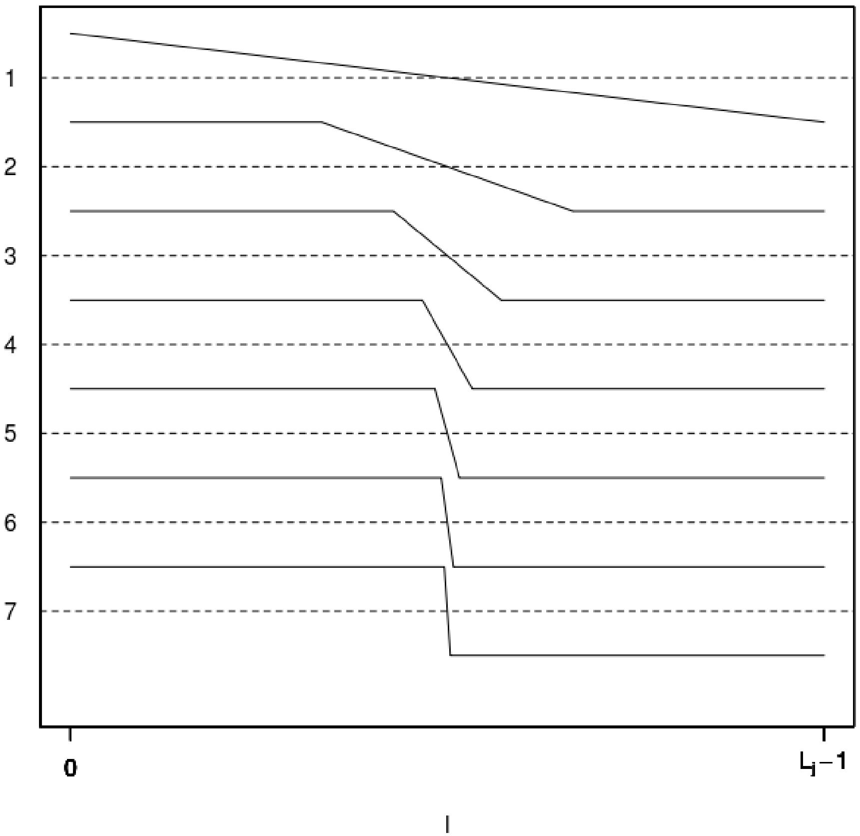
- Try to decompose the a continues signal, and assign each piece with a constant.

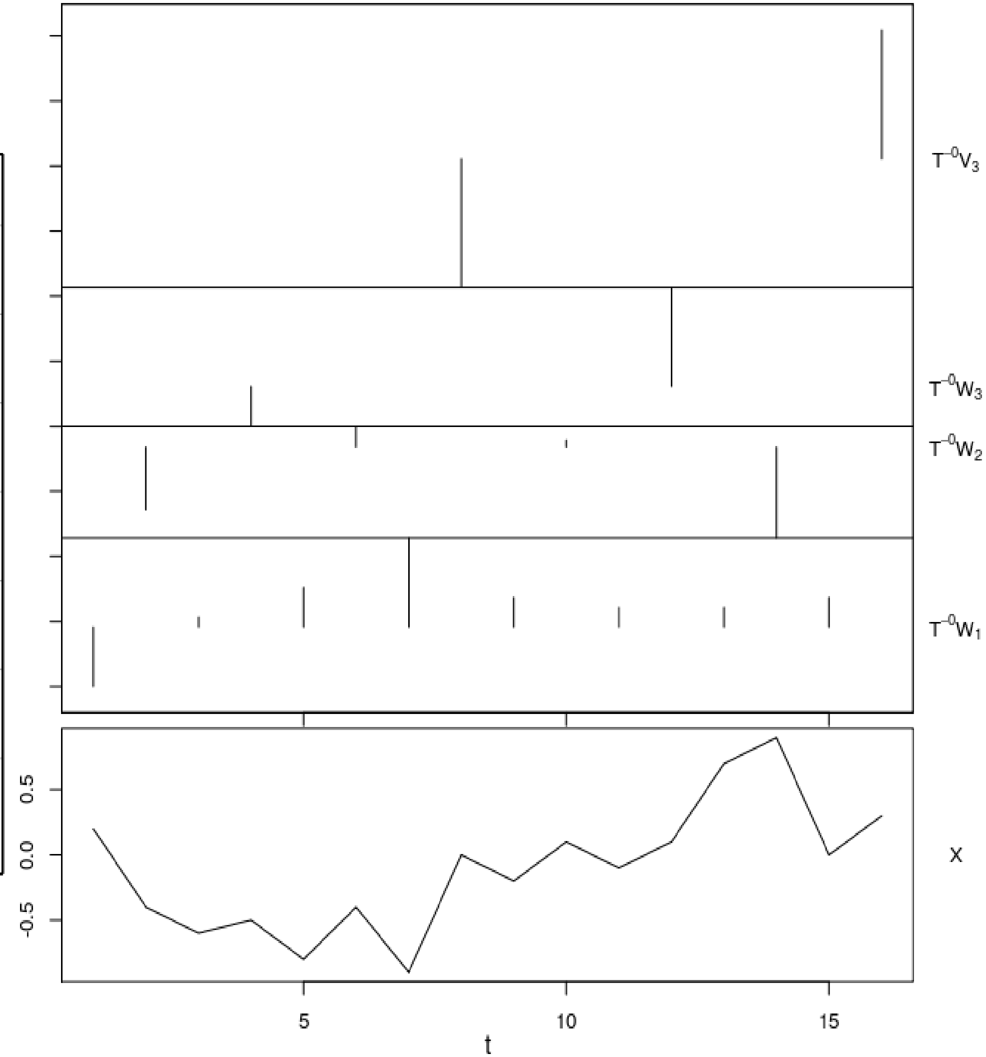
Haar Wavelet Analysis



Haar Wavelet in R

```
library('wavelets')
figure98.wt.filter("haar")
```






```
> X1 <- c(.2,-.4,-.6,-.5,-.8,-.4,-.9,0,-.2,.1,-.1,.1,.7,.9,0,.3)
> wt <- dwt(X1, filter = "haar", n.levels=3)
```

More Techniques

- Fisher Transform – getting a bell shaped PDF of time series data
- CG Oscillator – Obtain time series signal trend by observing the center of gravity of the signal
- Relative Vigor Index, etc.
- Smoothing, Averaging, Filtering and Normalizing…

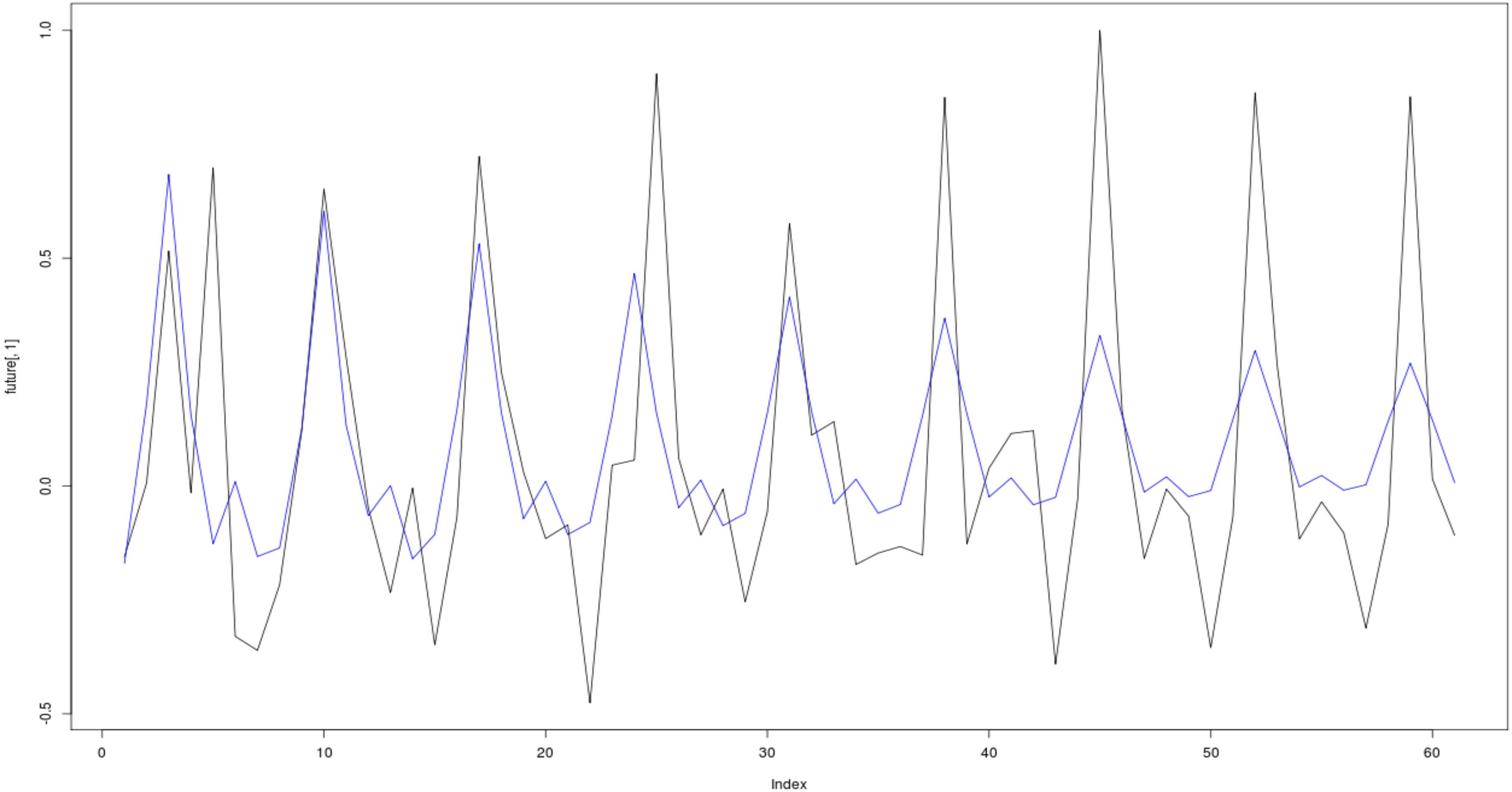
Discussion: How much sense does Technical Analysis (Signal Processing Approaches for Time Series Forecasting) make?

Autoregressive Model (AR)

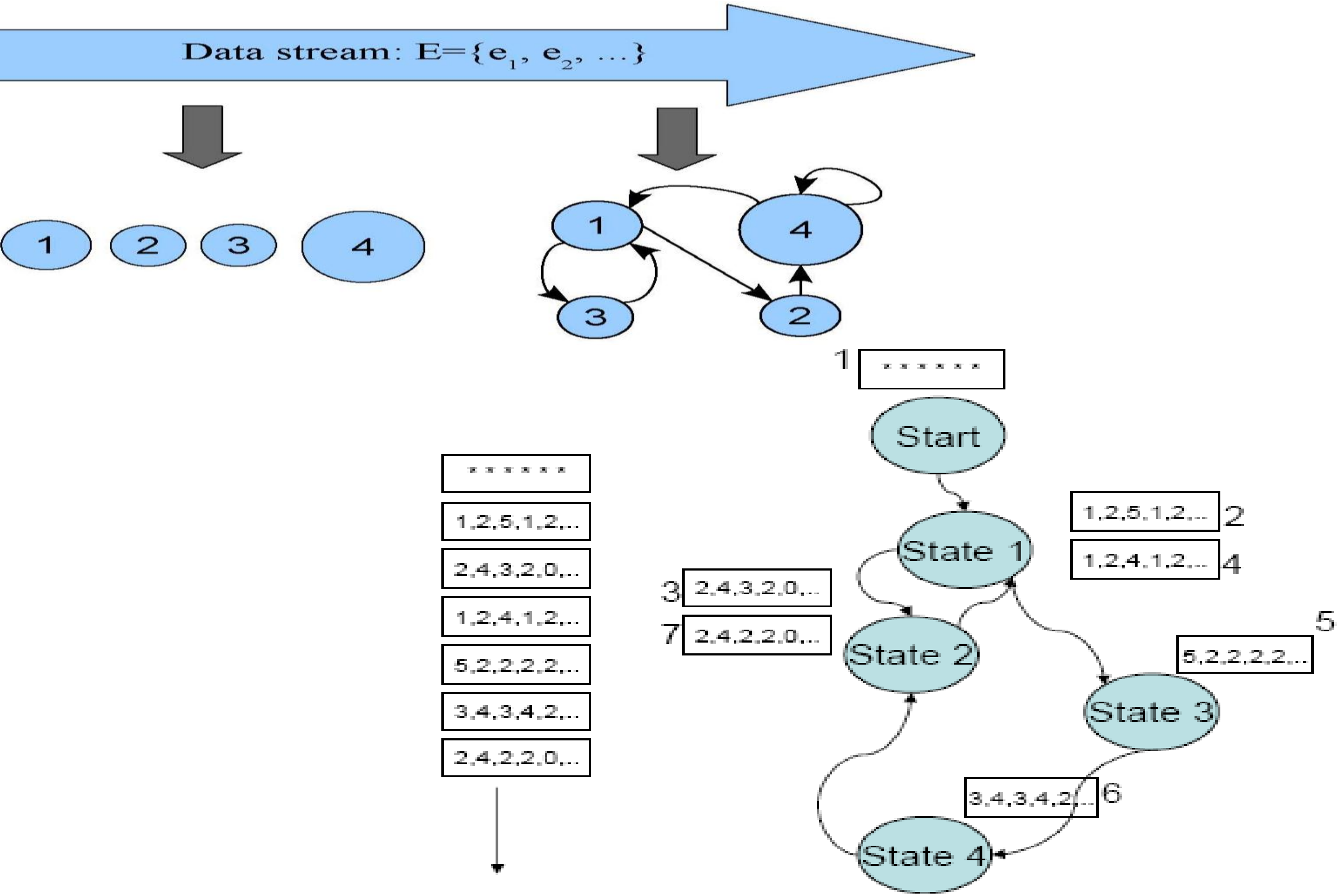
- $X_t = C + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$

```
> str(d)
'data.frame':   276 obs. of  1 variable:
 $ D: int  3 3 6 11 14 11 12 8 1 8 ...
> AR_obj<-ar(d[,1],na.action=na.exclude)
> str(AR_obj)
List of 14
 $ order       : int 13
 $ ar          : num [1:13] 0.8383 -0.2178 0.0792 -0.1655 0.209 ...
 $ var.pred    : num 65.6
 $ x.mean      : num 22.9
 $ aic         : Named num [1:24] 260.8 116.5 95.5 95.5 97.4 ...
  ..- attr(*, "names")= chr [1:24] "0" "1" "2" "3" ...
 $ n.used      : int 209
 $ order.max   : num 23
 $ partialacf  : num [1:23, 1, 1] 0.7095 -0.3234 0.0964 0.0226 0.3508 ...
 $ resid       : num [1:209] NA NA NA NA NA NA NA NA NA NA ...
 $ method      : chr "Yule-Walker"
 $ series      : chr "d[, 1]"
 $ frequency   : num 1
 $ call        : language ar(x = d[, 1], na.action = na.exclude)
 $ asy.var.coef: num [1:13, 1:13] 0.004709 -0.003815 0.00103 -0.000611 0.0010
93 ...
 - attr(*, "class")= chr "ar"
```

Performance of AR



Extensible Markov Model (EMM)

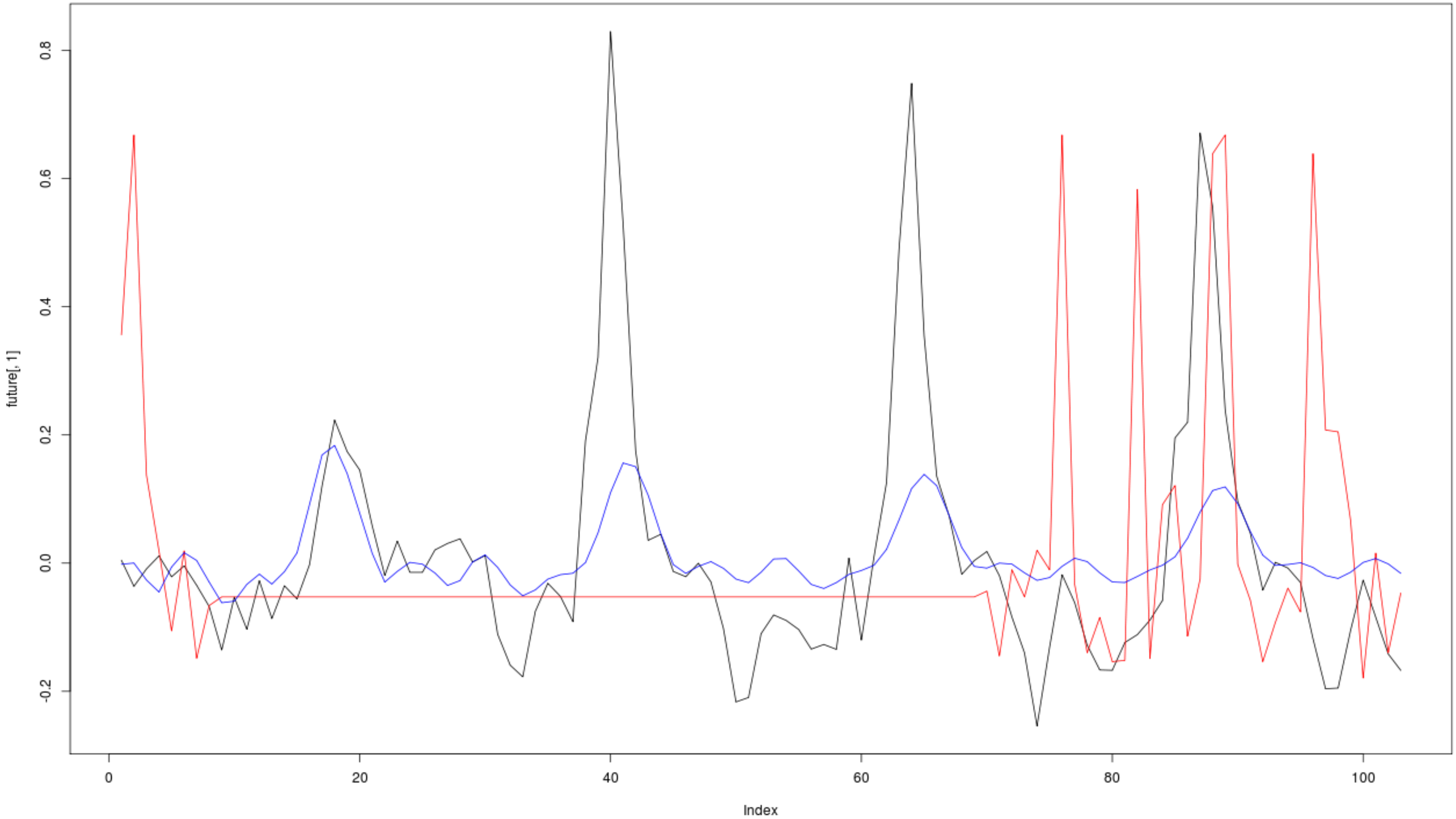


Discussion: The AR-EMM innovation

- AR model takes the weighted sum of previous p values to get the next data point

- How about use p as the size of each vector used by EMM for clustering? Will these p historical values recommend by AR model improve the performance of EMM prediction?

Performance of AR-EMM



GARCH Model

- Generalized Autoregressive Conditional Heteroskedasticity Model

The simplest and most commonly used GARCH model designed by Bollerslev is the GARCH (1,1) and is defined as

$$\sigma_i^2 = \omega + \alpha r_{i-1}^2 + \beta \sigma_{i-1}^2$$

$$\omega > 0 \text{ and } \alpha, \beta \geq 0$$

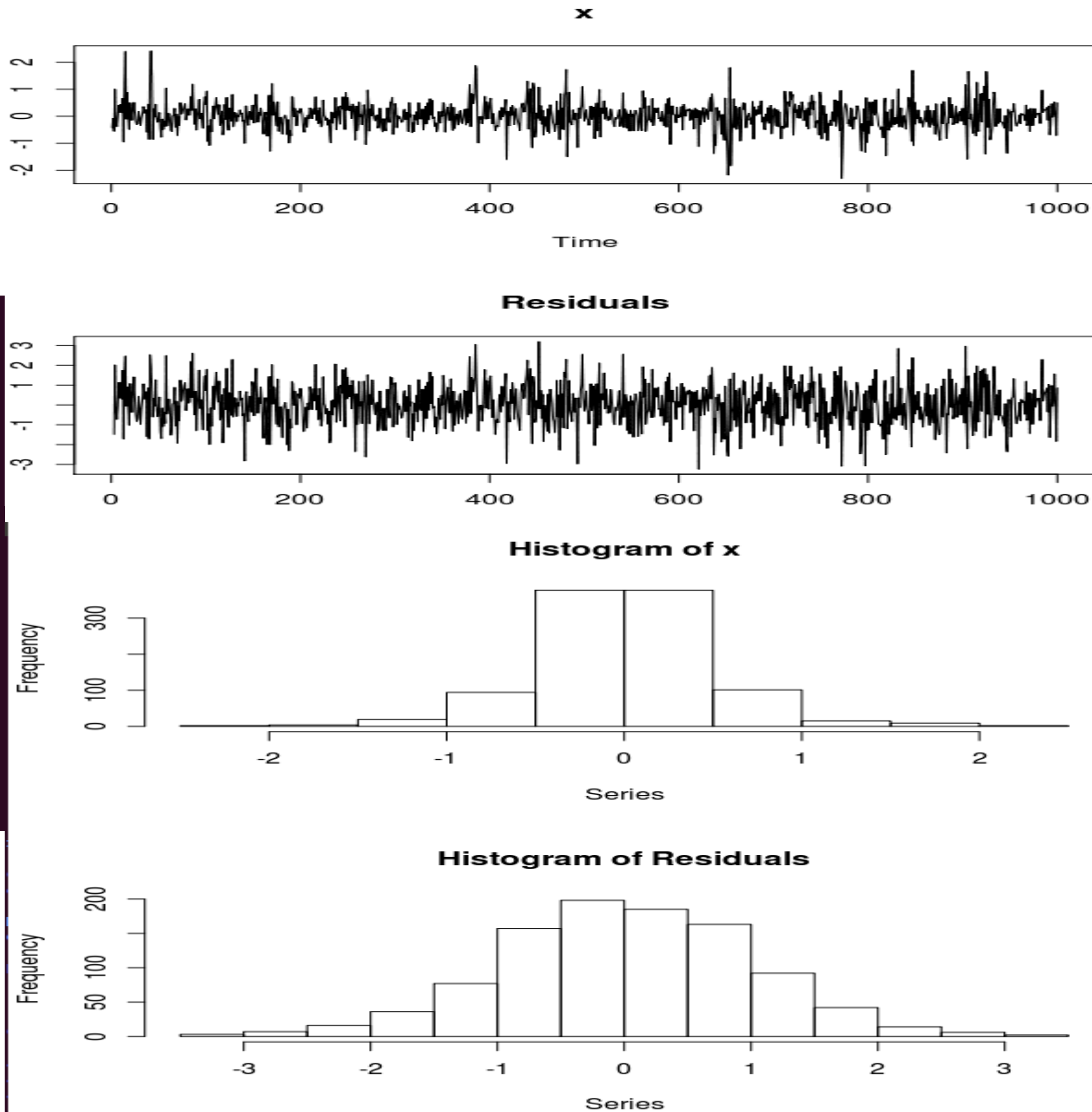
where:
α is the weight assigned to the lagged squared returns
β is the weight assigned to the lagged variances
ω is a constant equal to γ x V_L where V_L is the long run variance rate and γ is its weight .

This model estimates the volatility on a given day based on a linear combination of the squared returns and volatilities of the previous days plus a constant. Indeed the "(1,1)" term in GARCH (1,1) indicates that the current variance is based on the squared return and variance of the previous day (1 lag for each). We use the squared returns because they also exhibit strong recognizable patterns.

GARCH in R

```
   See 'library(help="tseries")' for details.

> n <- 1100
> a <- c(0.1, 0.5, 0.2) # ARCH(2) coefficients
> e <- rnorm(n)
> x <- double(n)
> x[1:2] <- rnorm(2, sd = sqrt(a[1]/(1.0-a[2]-a[3])))
> for(i in 3:n) # Generate ARCH(2) process
+ {
+ x[i] <- e[i]*sqrt(a[1]+a[2]*x[i-1]^2+a[3]*x[i-2]^2)
+ }
> x <- ts(x[101:1100])
> x.arch <- garch(x, order = c(0,2)) # Fit ARCH(2)
```



Other Models in the future

- SABR model – for volatility analysis, designated to find the volatility smile.
- Black-Schole – for financial derivative pricing.
- Much more quantitative models for financial time series mining/analysis

References

- EmausBot. (2011, 12 30). Backtesting. Retrieved 2 8, 2012, from Wikipedia: http://en.wikipedia.org/wiki/Backtesting
- popularlibros. (2011, 11 14). 1st International Competition of Time Series Forecasting. Retrieved 2 8, 2012, from ICTSF-Motivation: http://www.caos.inf.uc3m.es/~jperalta/ICTSF/motivation.html
- Rob J. Hyndman, A. Z. (2011, 12 28). CRAN Task View: Time Series Analysis. Retrieved 2 8, 2011, from CRAN: http://cran.r-project.org/web/views/TimeSeries.html
- ZéroBot. (2012, 1 28). Autoregressive model. Retrieved 2 8, 2012, from Wikipedia: http://en.wikipedia.org/wiki/Autoregressive_model
- rEMM: Extensible Markov Model for Data Stream Clustering in R – Dr. M. F. Hahsler, Dr. M. H. Dunham
- Temporal Structure Learning for Clustering Massive Data Streams in Real-Time – Dr. M. F. Hahsler, Dr. M. H. Dunham
- The Magnificent EMM -- Michael Hahsler, Mallik Kotamarti, Charlie Isaksson – Presentation Slides from BYU
- Cybernetic Analysis for Stocks and Futures: Cutting-Edge DSP Technology to Improve Your Trading (Wiley Trading), John Ehlers
- All About Derivatives Second Edition (All About Series), Michael Durbin
- Modeling Univariate Volatility by Romain Berry, JP Morgan Investment Analytics & Consulting (http://www.jpmorgan.com/tss/General/Modeling_Univariate_Volatility/1267140581322)

Thank You

Questions?