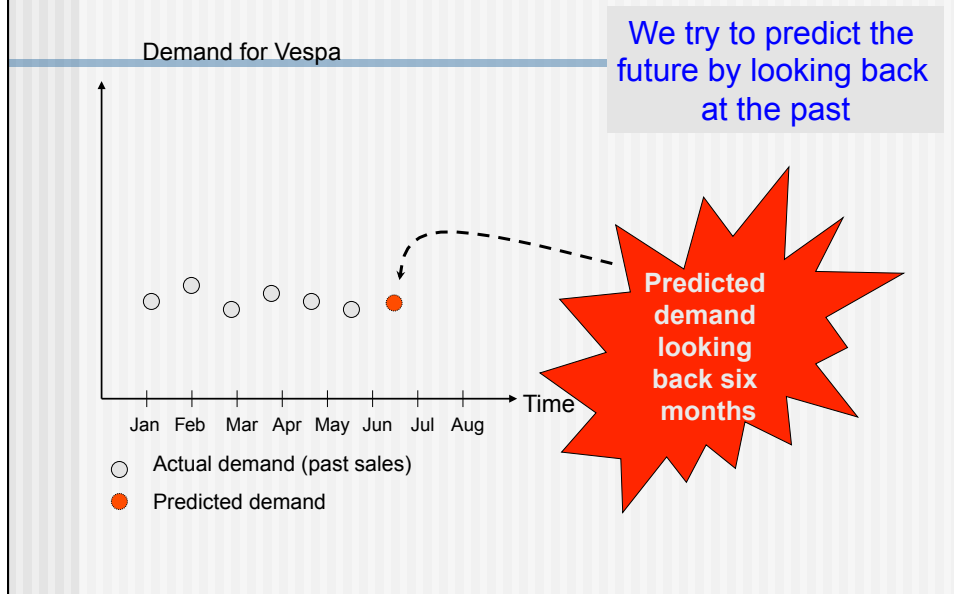


Time Series and Forecasting

Introduction to Forecasting

- What is forecasting?
 - Primary Function is to Predict the Future using (time series related or other) data we have in hand
- Why are we interested?
 - Affects the decisions we make today
- Where is forecasting used?
 - forecast demand for products and services
 - forecast availability/need for manpower
 - forecast inventory and materiel needs daily

What is forecasting all about?



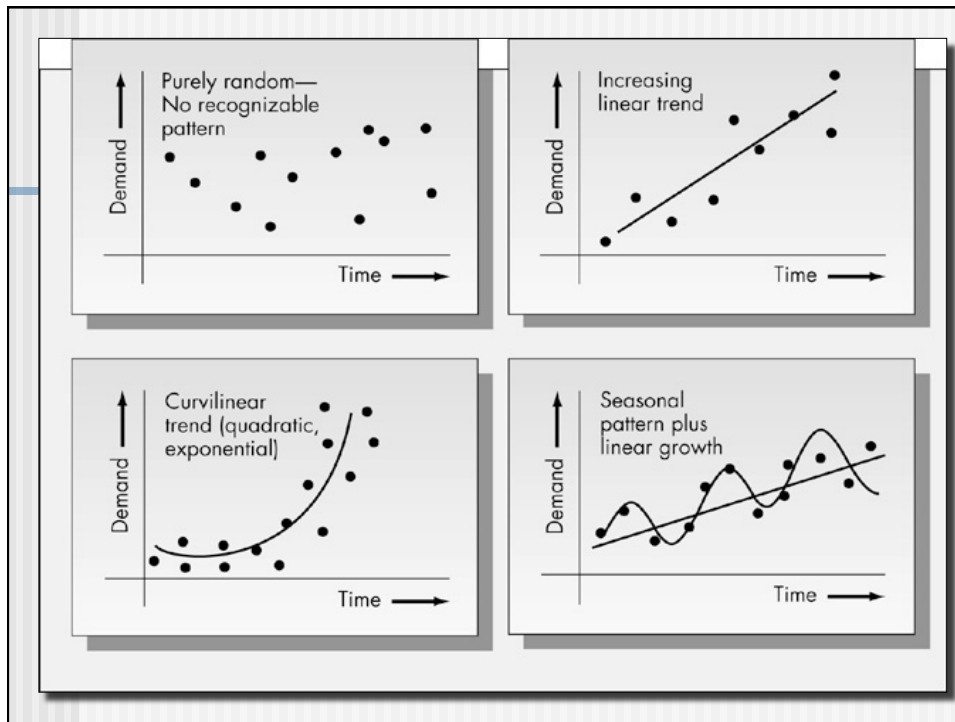
Characteristics of Forecasts

- They are usually wrong!
- A good forecast is more than a single number
 - Includes a mean value and standard deviation
 - Includes accuracy range (high and low)
- Aggregated forecasts are usually more accurate
- Accuracy erodes as we go further into the future.
- Forecasts *should not* be used to the exclusion of known information



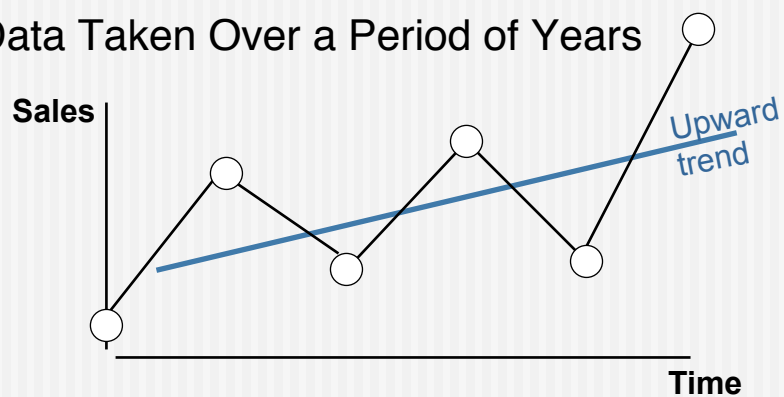
Time Series Methods

- A time series is just collection of past values of the variable being predicted. Also known as naïve methods. Goal is to isolate patterns in past data. (See Figures on following pages)
 - Trend
 - Seasonality
 - Cycles
 - Randomness



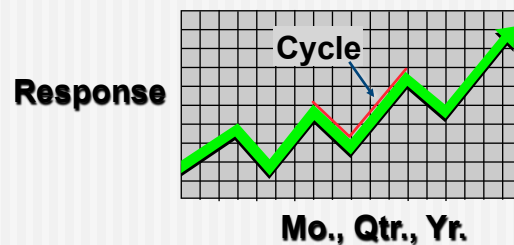
Trend Component

- Overall Upward or Downward Movement
- Data Taken Over a Period of Years



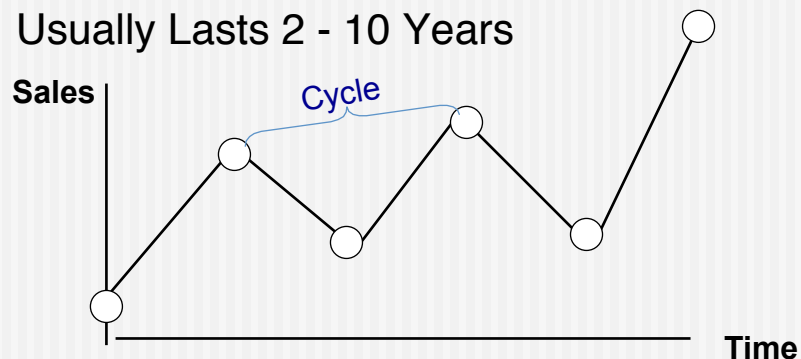
Cyclical Component

- Repeating up & down movements
- Due to interactions of factors influencing economy
- Usually 2-10 years duration



Cyclical Component

- Upward or Downward Swings
- May Vary in Length
- Usually Lasts 2 - 10 Years

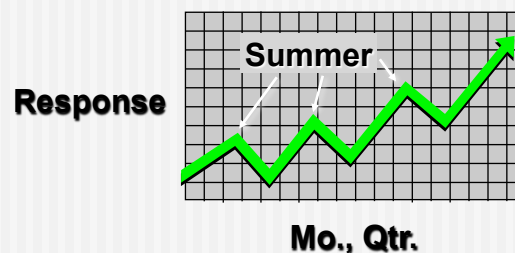


Cyclostationary process

- Signal having statistical properties that vary cyclically with time
- For example, the maximum daily temperature in New York City can be modeled as a cyclostationary process
 - the maximum temperature on July 21 is statistically different from the temperature on December 20; however,
 - it is a reasonable approximation that the temperature on December 20 of different years has identical statistics.

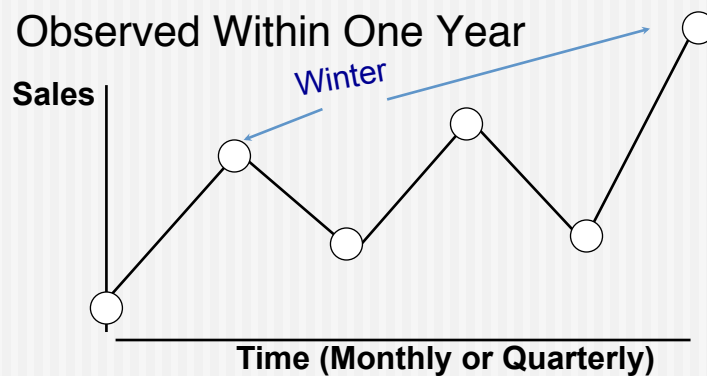
Seasonal Component

- Regular pattern of up & down fluctuations
- Due to weather, customs etc.
- Occurs within one year



Seasonal Component

- Upward or Downward Swings
- Regular Patterns
- Observed Within One Year

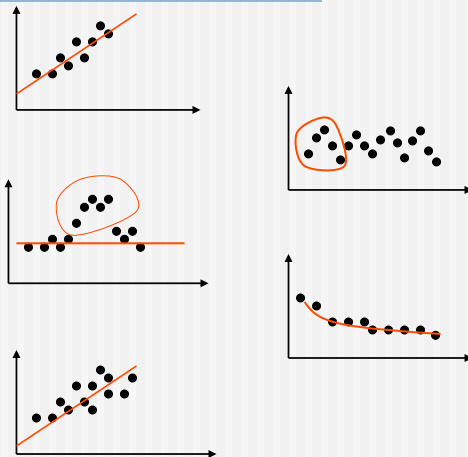


Irregular Component

- Erratic, unsystematic, 'residual' fluctuations
- Due to random variation or unforeseen events
 - Union strike
 - War
- Short duration & nonrepeating

What should we consider when looking at past demand data?

- Trends
- Seasonality
- Cyclical elements
- Autocorrelation
- Random variation

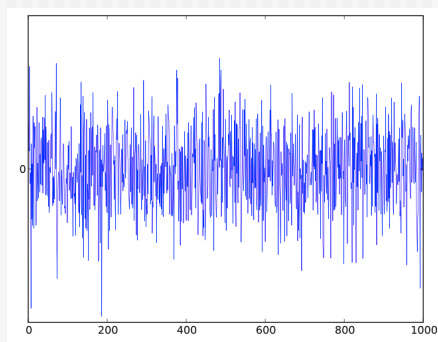


Stationary stochastic process

- Stochastic process whose joint probability distribution does not change when shifted in time
- Mean and variance do not change over time and do not follow any trends

White Noise is stationary

- In signal processing, white noise is a random signal with a constant power spectral density



Linear Systems

- It's possible that P , the process whose output we are trying to predict, is governed by linear dynamics.
- Example: stationary stochastic process

Causal Models

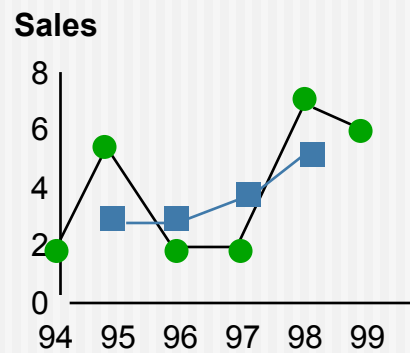
- Let Y be the quantity to be forecasted
- (X_1, X_2, \dots, X_n) are n variables that have predictive power for Y .
- A causal model is $Y = f(X_1, X_2, \dots, X_n)$.
- A typical relationship is a linear one:
$$Y = a_0 + a_1 X_1 + \dots + a_n X_n$$

Moving Average

<u>Year</u>	<u>Sales</u>	<u>MA(3) in 1,000</u>
1995	20,000	NA (20+24+22)/3 = 22 (24+22+26)/3 = 24
1996	24,000	
1997	22,000	
1998	26,000	(22+26+25)/3 = 24
1999	25,000	NA

Moving Average

Year	Response ●	Moving Ave ■
1994	2	NA
1995	5	3
1996	2	3
1997	2	3.67
1998	7	5
1999	6	NA



In the simple moving average models the forecast value is

$$F_{t+1} = \frac{A_t + A_{t-1} + \dots + A_{t-n}}{n}$$

t is the current period.

F_{t+1} is the forecast for next period

n is the forecasting horizon (how far back we look),

A is the actual sales figure from each period.

Example:

Kroger sells (among other stuff) bottled spring water

Month	Bottles
<i>Jan</i>	<i>1,325</i>
<i>Feb</i>	<i>1,353</i>
<i>Mar</i>	<i>1,305</i>
<i>Apr</i>	<i>1,275</i>
<i>May</i>	<i>1,210</i>
<i>Jun</i>	<i>1,195</i>
<i>Jul</i>	<i>?</i>

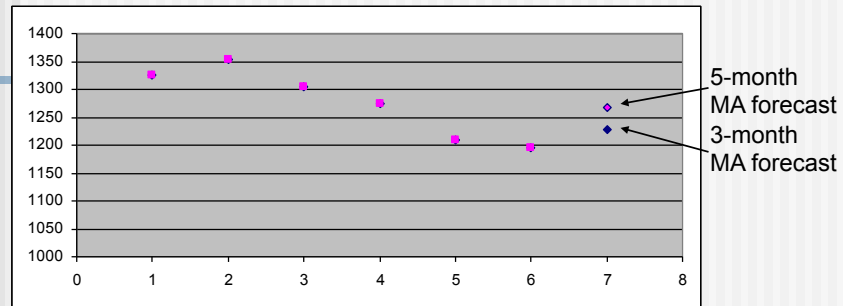


What if we use a 3-month simple moving average?

$$F_{Jul} = \frac{A_{Jun} + A_{May} + A_{Apr}}{3} = 1,227$$

What if we use a 5-month simple moving average?

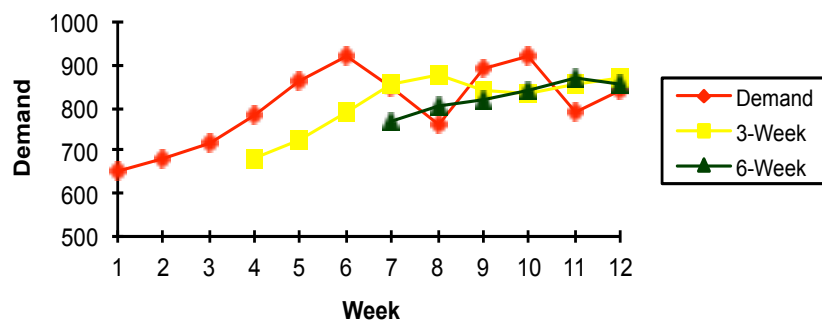
$$F_{Jul} = \frac{A_{Jun} + A_{May} + A_{Apr} + A_{Mar} + A_{Feb}}{5} = 1,268$$



What do we observe?

5-month average smoothes data more;
3-month average more responsive

Stability versus responsiveness in moving averages



Time series: weighted moving average

We may want to give more importance to some of the data

$$F_{t+1} = w_t A_t + w_{t-1} A_{t-1} + \dots + w_{t-n} A_{t-n}$$

$$w_t + w_{t-1} + \dots + w_{t-n} = 1$$

t is the current period.

F_{t+1} is the forecast for next period

n is the forecasting horizon (how far back we look),

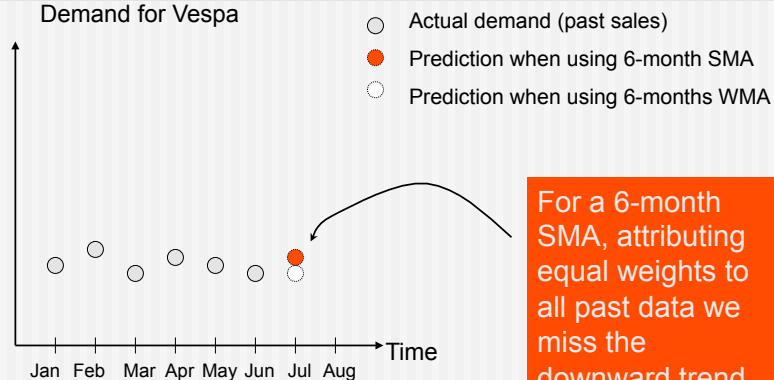
A is the actual sales figure from each period.

w is the importance (weight) we give to each period

Why do we need the WMA models?

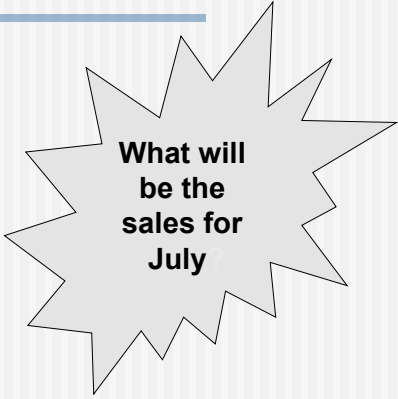
Because of the ability to give more importance to what happened recently, without losing the impact of the past.

Demand for Vespa



Example: Kroger sales of bottled water

Month	Bottles
<i>Jan</i>	1,325
<i>Feb</i>	1,353
<i>Mar</i>	1,305
<i>Apr</i>	1,275
<i>May</i>	1,210
<i>Jun</i>	1,195
<i>Jul</i>	?



**What will
be the
sales for
July**

6-month simple moving average...

$$F_{Jul} = \frac{A_{Jun} + A_{May} + A_{Apr} + A_{Mar} + A_{Feb} + A_{Jan}}{6} = 1,277$$

In other words, because we used equal weights, a slight downward trend that actually exists is not observed...

What if we use a weighted moving average?

Make the weights for the last three months more than the first three months...

	6-month SMA	WMA 40% / 60%	WMA 30% / 70%	WMA 20% / 80%
July Forecast	1,277	1,267	1,257	1,247

The higher the importance we give to recent data, the more we pick up the declining trend in our forecast.

How do we choose weights?

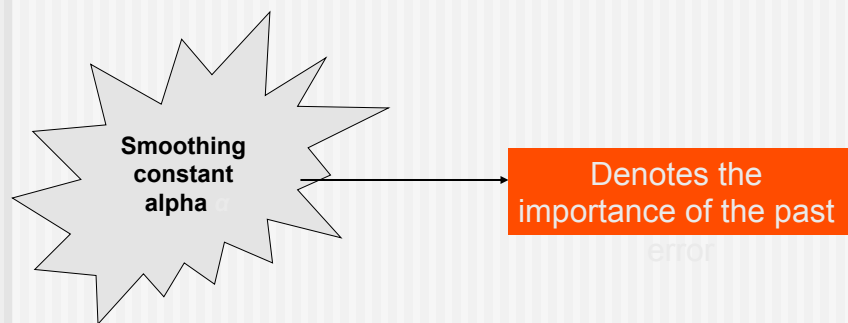
1. Depending on the importance that we feel past data has
2. Depending on known seasonality (weights of past data can also be zero).

**WMA is better than SMA
because of the ability to
vary the weights**

Time Series: Exponential Smoothing (ES)

Main idea: The prediction of the future depends mostly on the most recent observation, and on the error for the latest

forecast



Why use exponential smoothing?

1. Uses less storage space for data
2. Extremely accurate
3. Easy to understand
4. Little calculation complexity
5. There are simple accuracy tests

Exponential smoothing: the method

Assume that we are currently in period t . We calculated the forecast for the last period (F_{t-1}) and we know the actual demand last period (A_{t-1}) ...

$$F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1})$$

The smoothing constant α expresses how much our forecast will react to observed differences...

If α is low: there is little reaction to differences.

If α is high: there is a lot of reaction to differences.

Example: bottled water at Kroger

Month	Actual	Forecasted
<i>Jan</i>	1,325	1,370
<i>Feb</i>	1,353	1,361
<i>Mar</i>	1,305	1,359
<i>Apr</i>	1,275	1,349
<i>May</i>	1,210	1,334
<i>Jun</i>	?	1,309

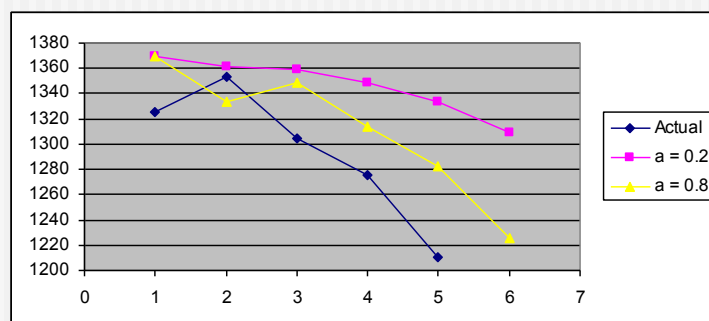
$\alpha = 0.2$

Example: bottled water at Kroger

Month	Actual	Forecasted
<i>Jan</i>	1,325	1,370
<i>Feb</i>	1,353	1,334
<i>Mar</i>	1,305	1,349
<i>Apr</i>	1,275	1,314
<i>May</i>	1,210	1,283
<i>Jun</i>	?	1,225

$\alpha = 0.8$

Impact of the smoothing constant



Linear regression in forecasting

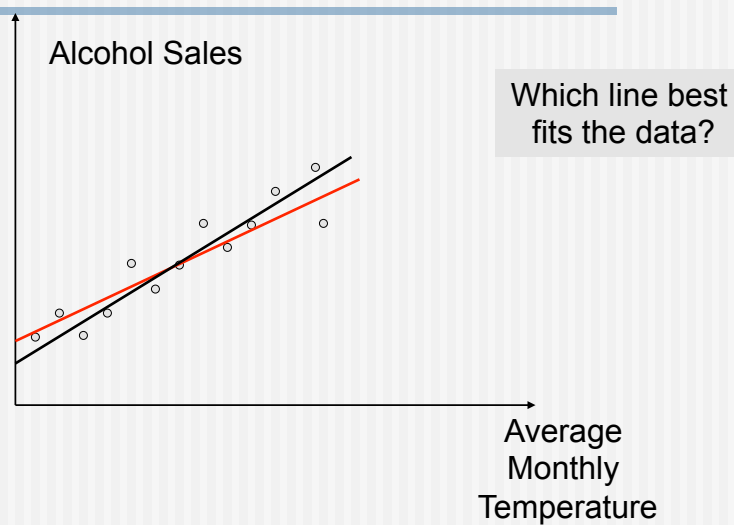
Linear regression is based on

1. Fitting a straight line to data
2. Explaining the change in one variable through changes in other variables.

$$\text{dependent variable} = a + b \times (\text{independent variable})$$

By using linear regression, we are trying to explore which independent variables affect the dependent variable

Example: do people drink more when it's cold?



Autoregressive model

- AR(p) refers to the autoregressive model of order p

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

- Parameters $\varphi_1, \dots, \varphi_p$
- c is a constant, and ε_t white noise

Moving-average model

- MA(q) refers to the moving average model of order q

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i}$$

where the $\theta_1, \dots, \theta_q$ are the parameters of the model, μ is the expectation of X_t (often assumed to equal 0), and the $\varepsilon_t, \varepsilon_{t-1}, \dots$ are again, white noise error terms.

ARMA

$$X_t = c + \varepsilon_t + \sum_{i=1}^p \varphi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i}.$$

- After choosing p and q the ARMA models can be fitted by least squares regression to find the values of the parameters which minimize the error term
- Good practice: find the smallest values of p and q which provide an acceptable fit to the data

Lag model

- Lag model is a model for time series data in which a regression equation is used to predict current values of a dependent variable
- It is based on both the current values of an explanatory variable and the lagged (past period) values of this explanatory variable

Autoregressive Integrated Moving Average (ARIMA)

- Generalization of ARMA
 - Are an adaptation of discrete-time filtering methods developed in 1930's-1940's by electrical engineers (Norbert Wiener et al.)
- *A series which needs to be differenced to be **made stationary** is an “integrated” (I) series (new)*
- Lags of the stationarized series are called “autoregressive” (AR) terms
- Lags of the forecast errors are called “moving average” (MA) terms

ARIMA terminology

- A non-seasonal ARIMA model can be (almost) completely summarized by three numbers:
 - p = the number of autoregressive terms
 - d = the number of nonseasonal differences
 - q = the number of moving-average terms
 - This is called an “ARIMA(p,d,q)” model
 - The model may also include a constant term (or not)

ARIMA forecasting equation

- Let Y denote the original series
- Let y denote the differenced (stationarized) series
 - No difference ($d=0$): $y_t = Y_t$
 - First difference ($d=1$): $y_t = Y_t - Y_{t-1}$
 - Second difference ($d=2$):

$$y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$$

Forecasting equation for y

$$\hat{y}_t = \underbrace{\mu}_{\text{constant}} + \underbrace{\phi_1 y_{t-1} + \dots + \phi_p y_{t-p}}_{\text{AR terms (lagged values of } y)}$$

By convention, the
AR terms are + and
the MA terms are -

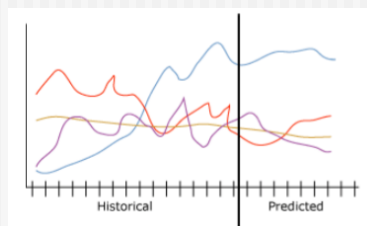
$$\underbrace{-\theta_1 e_{t-1} \dots - \theta_q e_{t-q}}_{\text{MA terms (lagged errors)}}$$

Not as bad as it looks! Usually $p+q \leq 2$ and
either $p=0$ or $q=0$ (pure AR or pure MA model)

Undifferencing the forecast

- The differencing (if any) must be reversed to obtain a forecast for the original series:
 - If (d=0): $Y_t = Y_t$
 - If (d=1): $Y_t = y_t + Y_{t-1}$
 - If (d=2): $Y_t = y_t + 2Y_{t-1} - Y_{t-2}$

Microsoft Time Series



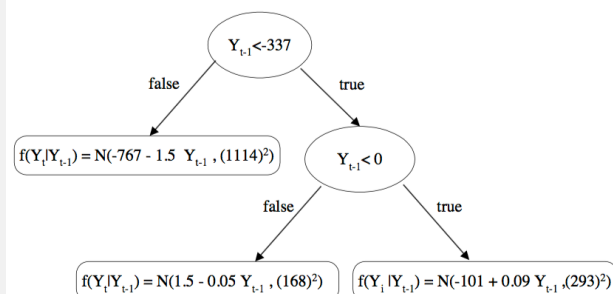
- The ARTXP algorithm is optimized for predicting the next likely value in a series
 - short-term prediction
- The ARIMA algorithm improves accuracy for long-term prediction
 - long-term prediction

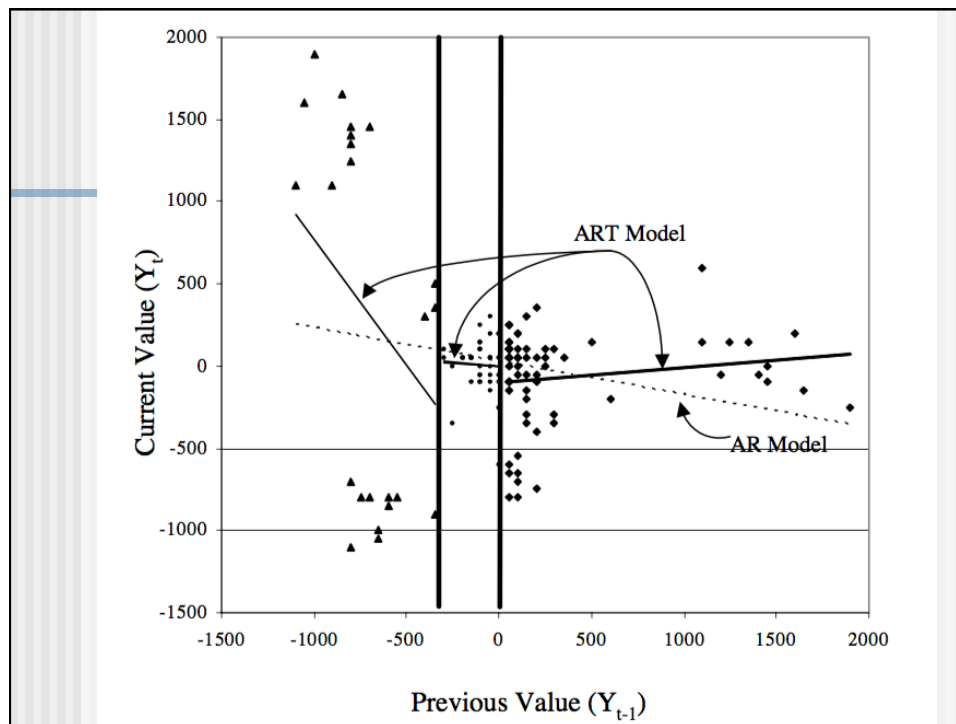
- ARTXP algorithm is based on Autoregressive Tree Models
- Generalization of standard-autoregressive models $AR(p)$
- The decision tree has linear regressions at its leaves

- Standard “windowing” transformations of time-series are used to generate sub-time-series
- Sub-series are used for for predicting values by regression analysis $AR(p)$
- Each sub-series corresponds to $AR(p)$

- The decision tree has linear regressions at its leaves
- Decision tree with one single leaf is equal to AR(p) model
- Several leaves can model non-linear relationships in time series data (combination of several AR(p) models)

- Temporal sequence $Y=(Y_1, Y_2, \dots, Y_T)$
 - Each sub-series corresponds to $AR_i(p)$
 - Binary search in the tree for the probable $AR_i(p)$





Causal Models and Neural Networks

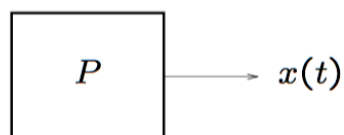
- Let Y be the quantity to be forecasted
- (X_1, X_2, \dots, X_n) are n variables that have predictive power for Y .
- A causal model is $Y = f(X_1, X_2, \dots, X_n)$.
- Non Linear Model $f(X_1, X_2, \dots, X_n)$ represented by NN

Neural Networks for Time Series Prediction

- Based on earlier slides by Dave Touretzky and Kornel Laskowski

What is a Time Series for NN?

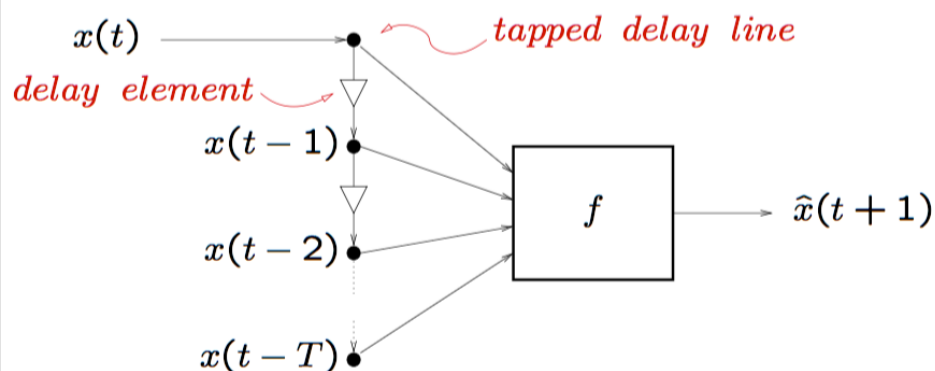
- sequence of vectors (or scalars) which depend on time t . In this lecture we will deal exclusively with scalars:
- $\{x(t_0), x(t_1), \dots, x(t_{i-1}), x(t_i), x(t_{i+1}), \dots\}$ It's the output of some process P that we are interested in:



Possible Types of Processing

- Predict future values of $x[t]$
- Classify a series into one of a few classes
 - “price will go up”
 - “price will go down” — sell now “no change”
- Describe a series using a few parameter values of some model
- Transform one time series into another
 - oil prices \rightarrow interest rates

Using the Past to Predict the Future



The Problem of Predicting the Future

- Extending backward from time t , we have time series $\{x[t], x[t-1], \dots\}$. From this, we now want to estimate x at some future time
- $x[t+s] = f(x[t], x[t-1], \dots)$
- s is called the horizon of prediction.

Horizon of Prediction

- So far covered many neural net architectures which could be used for predicting the next sample in a time series. What if we need a longer forecast, ie. not $x[t+1]$ but $x[t+s]$, with the horizon of prediction $s > 1$?

- Train on $\{x[t], x[t-1], x[t-2], \dots\}$ to predict $x[t+s]$
- Train to predict all $x[t+i]$, $1 \geq i \geq s$ (good for small s).
- Train to predict $x[t+1]$ only, but iterate to get $x[t+s]$ for any s .

Predicting Sunspot Activity

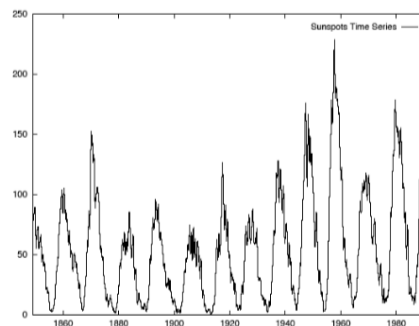
- Sunspots affect ionospheric propagation of radio waves.
- Telecom companies want to predict sunspot activity six months in advance.
- Sunspots follow an 11 year cycle, varying from 9-14 years. Monthly data goes back to 1849.

Predicting Sunspot Activity

- ***Fessant, Bengio and Collobert.***
- Authors focus on predicting $IR5$, a smoothed index of monthly solar activity.

$$IR5[t] = \frac{1}{5} (R[t-3] + R[t-2] + R[t-1] + R[t] + R[t+1])$$

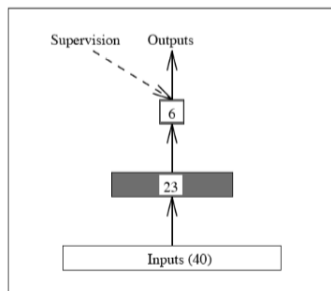
where $R[t]$ is the mean sunspot number for month t and $IR5[t]$ is the desired index.



Simple Feedforward NN

Output: $\{\hat{x}[t], \dots, \hat{x}[t+5]\}$

(1087 weights)

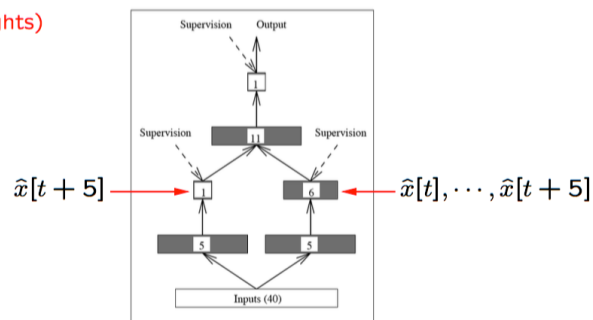


Input: $\{x[t-40], \dots, x[t-1]\}$

Modular Feedforward NN

Output: $\hat{x}[t+5]$

(552 weights)

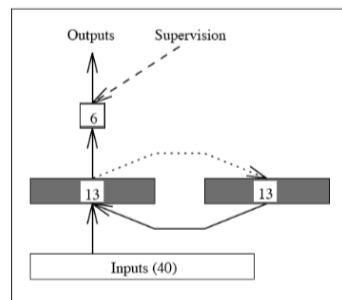


Input: $\{x[t-40], \dots, x[t-1]\}$

Elman NN

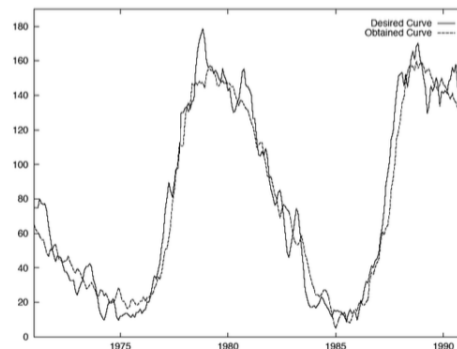
(786 weights)

Output: $\{\hat{x}[t], \dots, \hat{x}[t+5]\}$



Input: $\{x[t-40], \dots, x[t-1]\}$

Fessant et al: Results



Train on first 1428 samples Test on last 238 samples	CNET heuristic	Simple Net	Modular Net	Elman Net
Average Relative Variance	0.1130	0.0884	0.0748	0.0737
# Strong Errors	12	12	4	4