A short introduction to Demand Forecasting

Mines Saint-Étienne – Master Génie industriel

Nicolas Durrande (durrande@emse.fr)

Objectives of the course:

- Being able to recognize common patterns in time series
- Explain and apply various forecasting methods
- Being able to asses the quality of a forecast

This course is inspired from
www.slideshare.net/knksmart/forecasting-slides

What is forecasting ?

Forecasting is the process of making predictions for an event that has not been observed. A typical example is to estimate the value of a variable at a future date.

Why do we need forecasting ?

Most managerial decisions will be applied in a near or distant future. In many cases, it is thus important to have an idea of the future environment in which these decisions will be applied.

Is forecasting difficult ?

it really depends on the problem!

Example

Forecasting in operation management:

- Predicting demands of new or existing products
- Predicting cost of materials
- ...

Decisions requiring forecasting:

- Choosing new facility location
- Identifying labour requirement
- Projecting material requirement
- Creating maintenance schedules

Succesfull forecasting is a combination of science and art:

- Science since it relies on rigorous mathematical methods
- Art because the decision maker uses his experience, logic and intuition to supplement the forecasting quantitative analysis.

There are two kinds of forecast types:

qualitative where the forecast is based on opinions. We can cite expert opinion, sale force survey, consumer survey, Delphi method, ...

quantitative methods where the forecast is based on data. Common examples are time series analysis and associative methods.

The process of decision making usually uses both.

Forecast cannot predict the actual future value but it tries to **minimize the prediction error**.

Given the quantity to forecast and the time horizon, the main forecasting steps are :

- Selecting the forecasting model(s)
- Validate the model
- Making the forecast

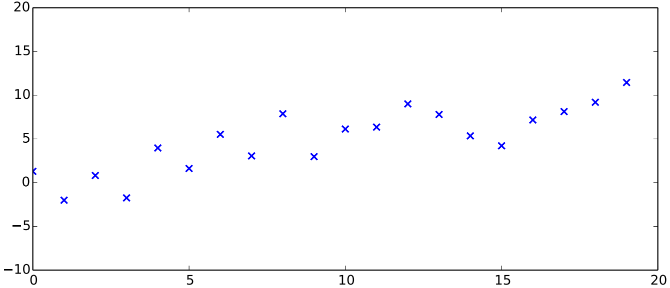
The choice of the forecasting model depends on the characteristics of the time series.

A time series corresponds to the observations of a variable over a set of regularly spaced points.

Example

2 representations of the same time series

t	0	1	2	3	4	5	6	...	19
y	-2.19	-4.7	-1.1	4.2	1.8	4.8	5.7	...	10.7

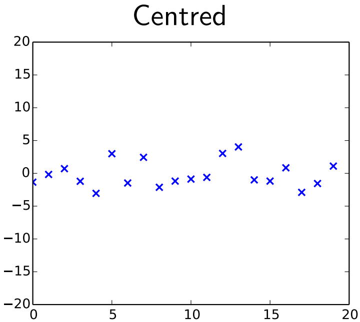


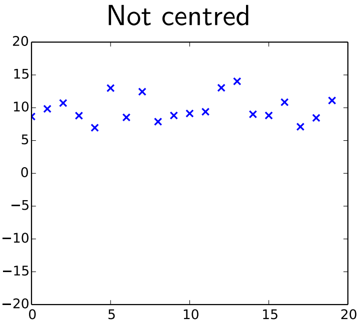
A time series can show various patterns or features:

- centred
- trend
- seasonal
- cyclical
- stationary
- noise
- missing data

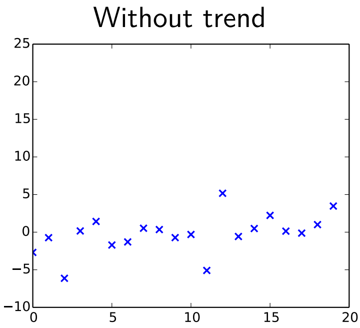
We will now illustrate these features on various examples.

Centred : Is the mean value 0 ?





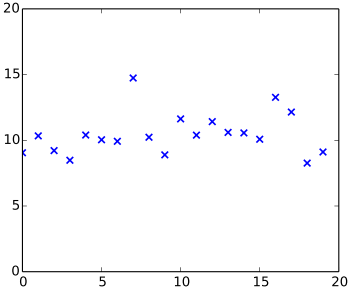
Trend : It is often the most obvious pattern



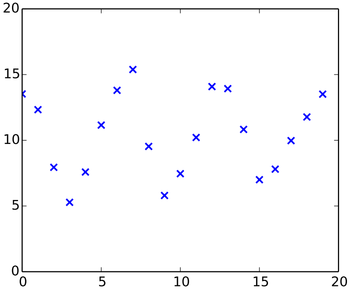
A trend is not necessarily linear !

Seasonal : It is also a pattern that is easy to identify

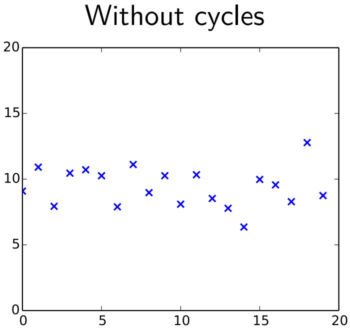
Without seasonality

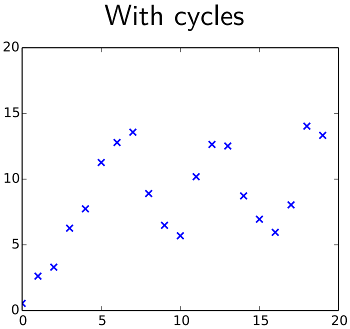


With seasonality



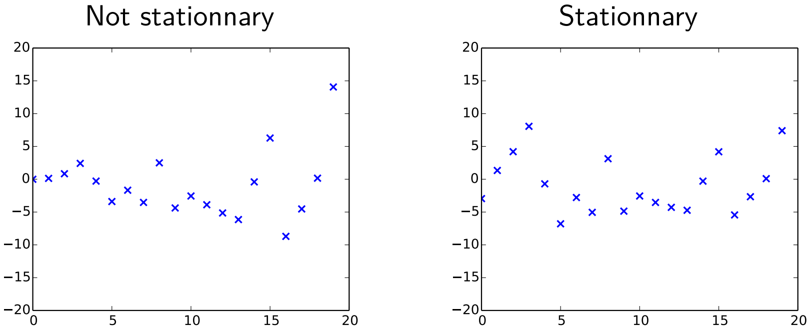
Cyclical : Similar to seasonality but not periodic





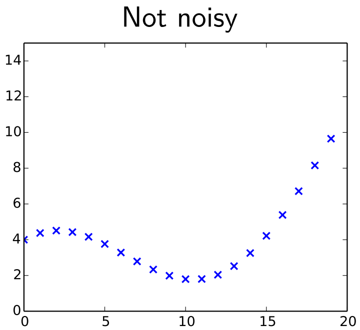
This can make the forecast difficult.

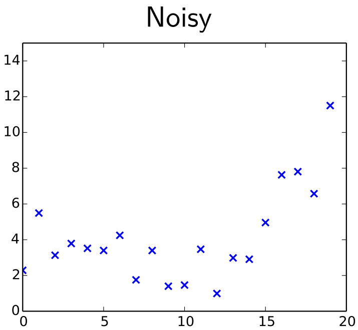
Stationary : A time series is stationary when the distribution of $y(t)$ does not change over time.



This is a key concept in time series analysis.

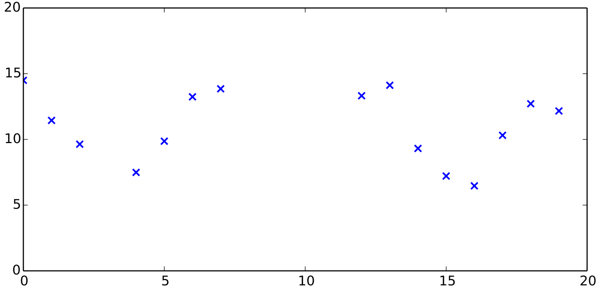
Noise : Noise is the component of the signal that cannot be explained.





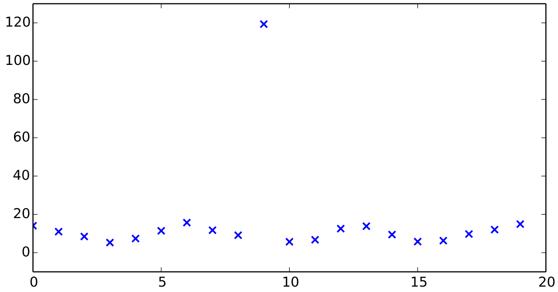
It is often difficult to tell if some variations are due to noise.

Missing data : In practice, full data is not always available.



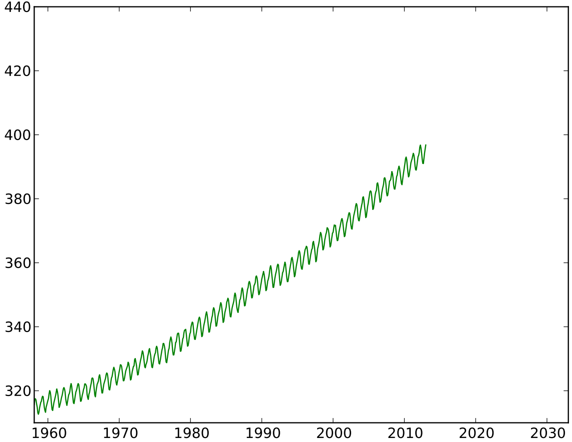
Some methods can cope very well with missing data, others cannot.

Outliers : Some of the data may not be reliable.



Outliers should be removed in a preprocessing step.

Most time series show a combinations of the above features:



CO_2 Concentration in the atmosphere in ppm at the Mauna Loa observatory (Hawaii).

The quantitative forecasting methods can be divided in two sets:

time series models

- Moving average
- Exponential smoothing
- Autoregressive models

associative models

- Linear regression

Moving average

Moving average is a basic method that can be used for:

- noise filtering
- missing data prediction
- forecasting

The prediction $\hat{y}_t$ at point t is given by the average over the neighbours of t:

$$\hat{y}_t = \frac{1}{|V|} \sum_{v \in V} y_v$$

The main question here is how to define a neighbourhood.

Example (Denoising)

We consider the following time series that gives the wind velocity (in km/h) every two hours:

time	0	2	4	6	8	10	12	14	16	18
wind	10	3	6	0	2	7	14	21	17	28

- What is the smoothed times series if we consider a window $(t-1, t, t+1)$ at time t?
- Should the smoothing window be large or narrow?

Example (Missing data)

Same data as before with missing observation in 8:

time	0	2	4	6	8	10	12	14	16	18
wind	10	3	6	0		7	14	21	17	28

- What smotting window can be considered to predict the missing data?

Example (prediction)

Same settings:

time	0	2	4	6	8	10	12	14	16	18	20	22
wind	10	3	6	0	2	7	14	21	17	28		

- What smoothing window can be considered to predict the wind in 20?
- What about 22 ?

Weighted Moving average

Moving average is not very efficient when there is trend in data (see lab session).

An alternative is weighted moving average:

$$\hat{y}_t = \frac{1}{\sum w} \sum_{v \in V} w_{v-t} y_v$$

The weights w are often such that we give less importance to distant observations.

Exponential smoothing

One very popular weighted average method is **exponential smoothing**. In this case, the weights decline exponentially.

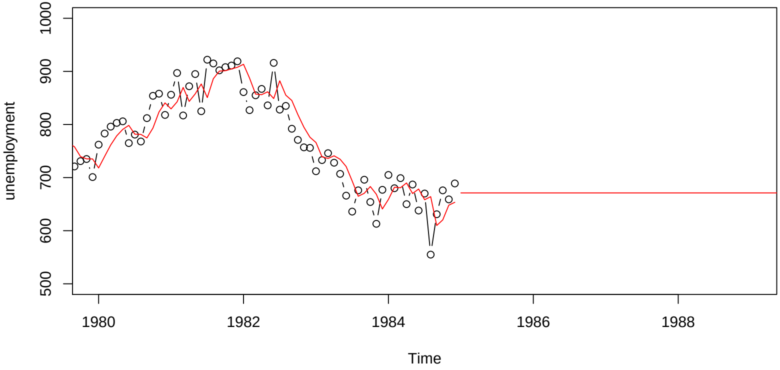
The two following definitions are equivalent:

$$\hat{y}_t = \alpha y_{t-1} + \alpha(1-\alpha) y_{t-2} + \cdots + \alpha(1-\alpha)^{t-1} y_0$$
$$\hat{y}_t = \alpha y_{t-1} + (1-\alpha)\hat{y}_{t-1}$$

The parameter α has to be chosen by the user.

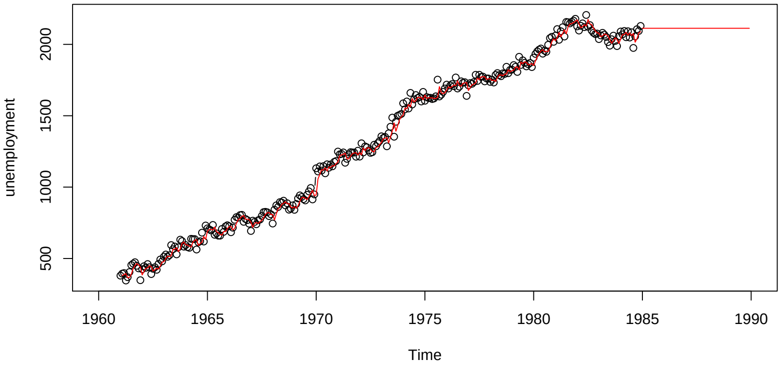
Exponential smoothing

Prediction at a two-step horizon is fine when there is no trend in the data :



Exponential smoothing

But not very satisfying otherwise:



Holt-Winters

Let $\hat{y}(t, h)$ denote the prediction at time t for horizon h.

We have just seen that for exponential smoothing

$$\hat{y}(t, h) = \hat{a}_t$$

The idea of Holt-Winter algorithm is to have a forecast of the form

$$\hat{y}(t, h) = \hat{a}_t + \hat{b}_t h$$

where $\hat{a}$ is called the *level* and $\hat{b}$ the *slope*. These parameters are updated in a similar fashion:

$$\hat{a}_{t+1} = \alpha y_{t+1} + (1 - \alpha)(\hat{a}_t + \hat{b}_t)$$
$$\hat{b}_{t+1} = \beta(\hat{a}_{t+1} - \hat{a}_t) + (1 - \beta)\hat{b}_t$$

Holt-Winters

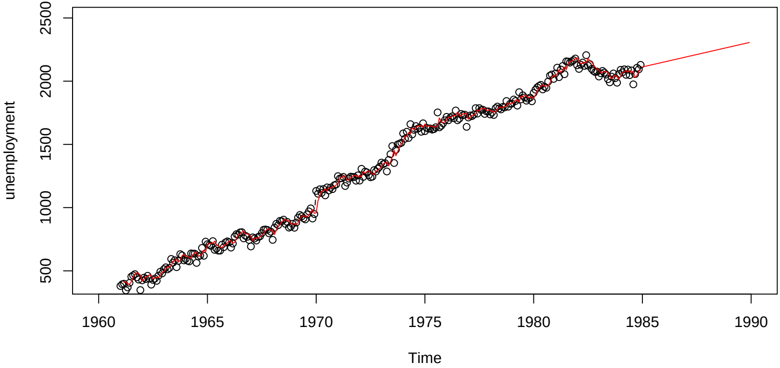
Holt-Winters can be initialized with

- $\hat{a}_2 = y_2$
- $\hat{b}_2 = y_2 - y_1$

If we choose $\beta = 0$ and $b_2 = 0$ we recover exponential smoothing.

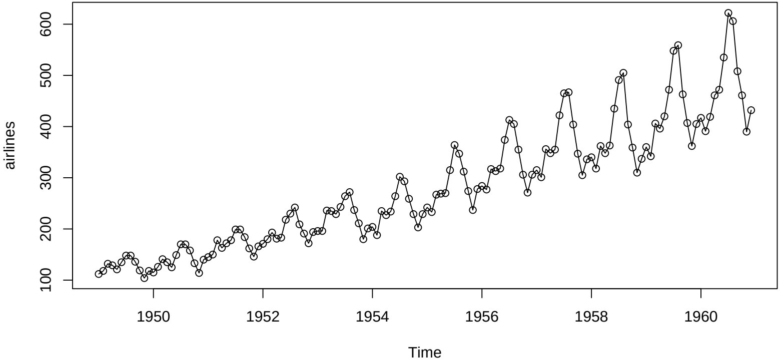
Holt-Winters

The method is much more adequate in the presence of a trend:



Holt-Winters

We have just seen that a trend can be added to exponential smoothing. In a similar fashion it is possible to add a seasonality in order to cope with datasets such as



Holt-Winters

Holt-Winters seasonal algorithm is based on the smoothing of three components:

- The level without seasonality $\hat{a}_n$
- The trend slope without seasonality $\hat{b}_n$
- The seasonality $\hat{c}_n$

We distinguish two kinds of models

- Holt-Winters with additive seasonality
- Holt-Winters with multiplicative seasonality

The period s is assumed to be known.

Holt-Winters

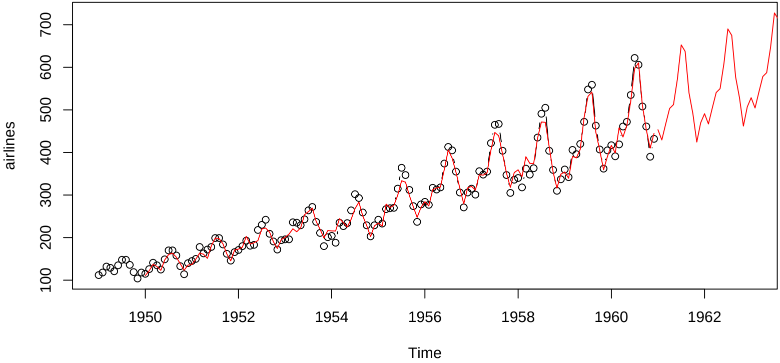
$$\hat{y}(t,h) = \hat{a}_t + \hat{b}_t h + \hat{c}_{n-ks+h}$$

The updates of $\hat{a}_n$, $\hat{b}_n$ and $\hat{c}_n$ are:

$$\begin{aligned}
\hat{a}_{t+1} &= \alpha(y_{t+1} - \hat{c}_{t+1-s}) + (1-\alpha)(\hat{a}_t + \hat{b}_t) \\
\hat{b}_{t+1} &= \beta(\hat{a}_{t+1} - \hat{a}_t) + (1-\beta)\hat{b}_t \\
\hat{c}_{t+1} &= \gamma(y_{t+1} - \hat{a}_{t+1}) + (1-\gamma)\hat{c}_{t+1-d}
\end{aligned}$$

Holt-Winters

We obtain on the previous example



Holt-Winters

Holt-Winters with multiplicative seasonality:

$$\hat{y}(t,h) = (\hat{a}_t + \hat{b}_t h)\hat{c}_{n-ks+h}.$$

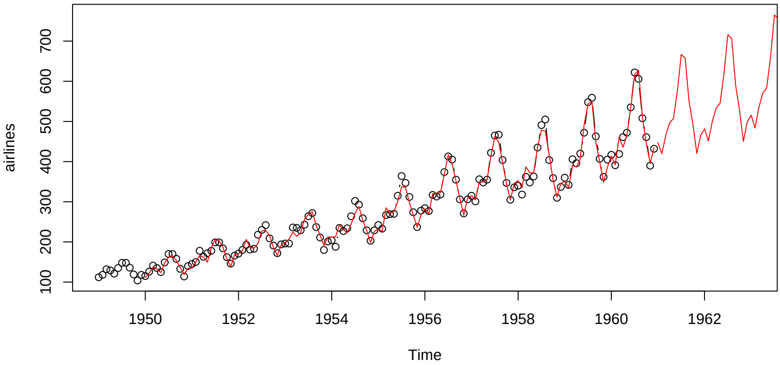
The updates of $\hat{a}_n$, $\hat{b}_n$ and $\hat{c}_n$ are:

$$\hat{a}_{t+1} = \alpha \frac{y_{t+1}}{\hat{c}_{t+1-s}} + (1-\alpha)(\hat{a}_t + \hat{b}_t)$$
$$\hat{b}_{t+1} = \beta(\hat{a}_{t+1} - \hat{a}_t) + (1-\beta)\hat{b}_t$$
$$\hat{c}_{t+1} = \gamma \frac{y_{t+1}}{\hat{a}_{t+1}} + (1-\gamma)\hat{c}_{t+1-d}$$

there are $2+s$ initial values that need to be specified.

Holt-Winters

We obtain on the previous dataset



Linear regression

Linear regression approximates some observations by a linear combination of basis functions.

Let $B(t) = (b_1(t), \dots, b_n(t))$ be the set of basis functions. We look for the function of the form $\hat{y}(t) = B(t)\beta$ that minimises the squared error $\sum(\hat{y}(t_i) - y(t_i))^2$ where the t_i correspond the locations of the observations.

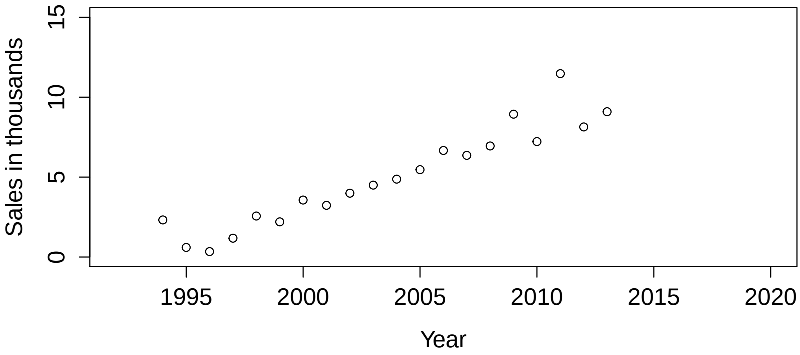
If we write X the matrix of general term $X_{i,j} = b_i(t_j)$ and Y the vector of observations, we have $\beta = (X^tX)^{-1}X^tY$ and thus:

$$\hat{y}(t) = B(t)(X^tX)^{-1}X^tY$$

Linear regression

Example

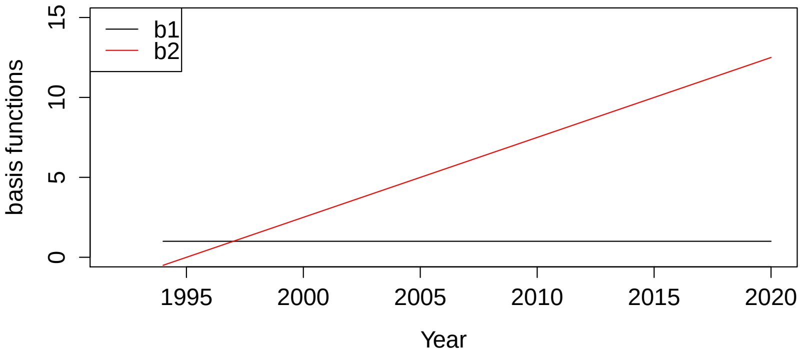
We consider a basic example :



Linear regression

Example

We use two basis functions :

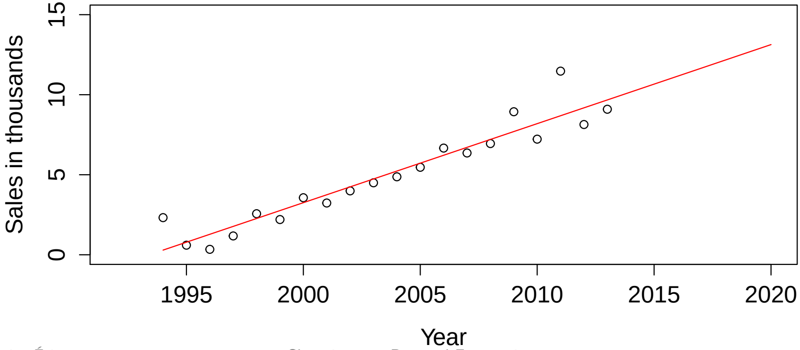


Linear regression

Linear regression

Example

We we obtain the following prediction :



Up to now, we assumed we had some observations of a variable but we had no **probabilistic model**.

In this section, we will assume that the **data we observed is drawn from a random variable**. We will first deal with the case were the distribution of this random variable is known and we will then discuss about its estimation.

We will consider three kinds of probabilistic models

- Autoregressive process (AR)
- Moving average process (MA)
- Autoregressive Moving average process (ARMA)

There are of course many others:

- ARIMA, SARIMA
- ARCH, GARCH, TAR
- ...

We first need to introduce a couple of notions:

random processes $\{Z_t\},\ t \in \mathbb{N}$ are a generalization of random variables. A sample of a random processes is a series (instead of a value). They are characterised by the joint distribution of the $\{Z_t\}$.

The mean function μ and the covariance function γ are defined as

- $\mu(t) = E[Z_t]$
- $\gamma(s,t) = \mathrm{cov}(Z_s, Z_t)$

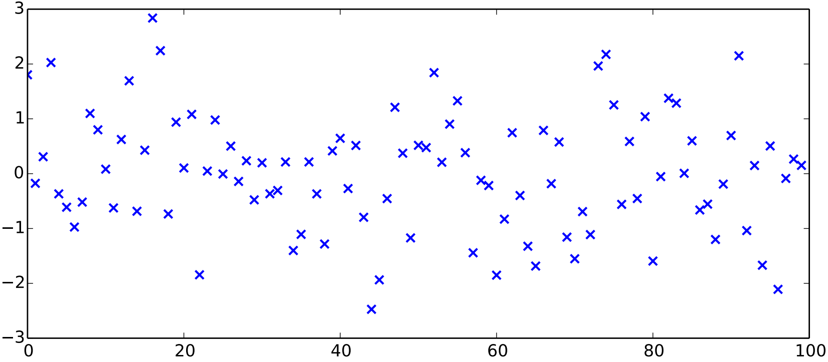
A process is (weakly) **stationary** if

(i) μ is a constant function

(ii) $\gamma(t, t+h) = \gamma(0, h)$

In this case, we often write γ as a function of one variable $\gamma(s-t)$.

From now we will denote by Z_t the white noise process. It is a process where the Z_t are centred and iid.

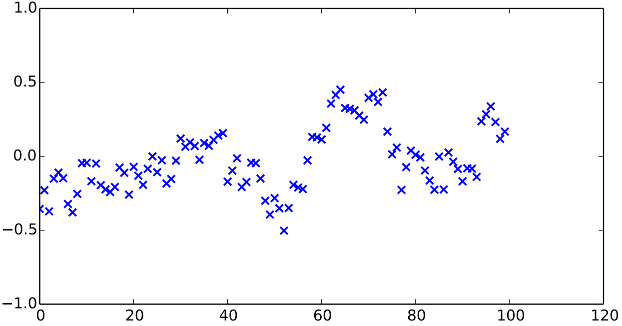


Definition

We call **moving average process of order q** (MA(q)) any process that can be written as :

$$X_t = \alpha_0 Z_t + \alpha_1 Z_{t-1} + \cdots + \alpha_q Z_{t-q}$$

They correspond to a moving average filter applied to white noise.



For this figure we have $w = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$.

Exercise

Given a **moving average process**

$$X_t = \alpha_0 Z_t + \alpha_1 Z_{t-1} + \cdots + \alpha_q Z_{t-q}$$

- Is X a stationary process?
- What is the covariance? $\gamma(h) = \mathrm{cov}(X_t, X_{t+h})$?

Exercise

Given a **moving average process**

$$X_t = \alpha_0 Z_t + \alpha_1 Z_{t-1} + \cdots + \alpha_q Z_{t-q}$$

- Is X a stationary process?
- What is the covariance? $\gamma(h) = \mathrm{cov}(X_t, X_{t+h})$?

The answers are

- Yes.
- $\gamma(h) = \mathrm{cov}(X_t, X_{t+h}) = \sigma^2 \sum_{i=0}^{q-h} \alpha_i \alpha_{i+h}$

In particular, the covariance is null for $h > q$.

Definition

We call **autoregressive process of order p** any process that can be written as :

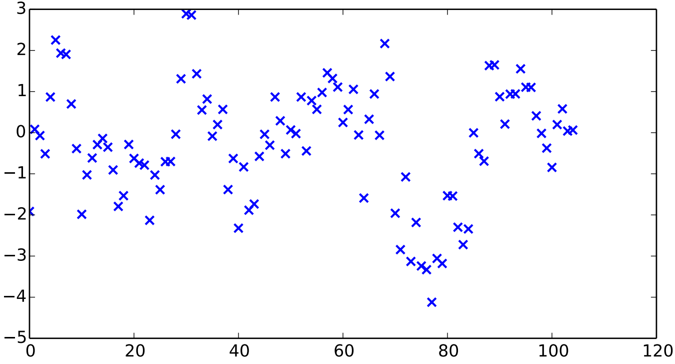
$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p} = \alpha_0 Z_t$$

This process can be seen as the solution of a discretized differential operator.

Example 1/3: AR(1)

We consider

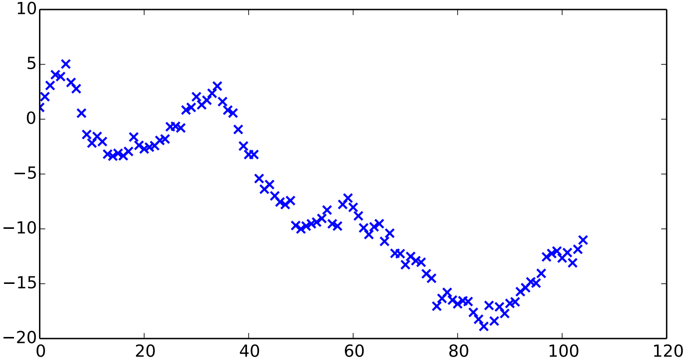
$$X_t - 0.8X_{t-1} = Z_t$$



Example 2/3: AR(1)

Special case : A random walk

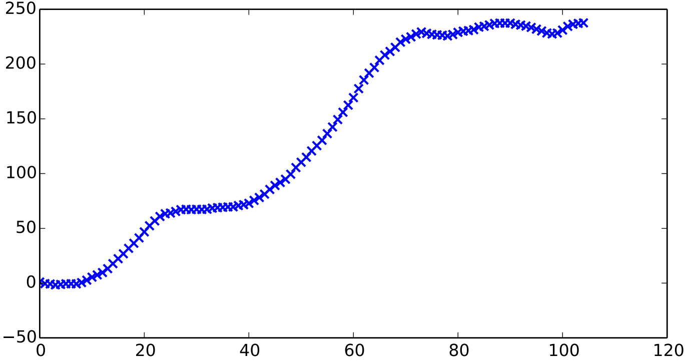
$$X_t - X_{t-1} = Z_t$$



Example 3/3: AR(2)

We consider

$$X_t - 2X_{t-1} + X_{t-2} = Z_t$$



Let X_t be an AR(1) process:

$$X_t - \alpha X_{t-1} = Z_t,$$

then X_t is stationary if $\alpha \neq 1$.

Furthermore, if $\alpha < 1$, X_t can also be written as a $MA(\infty)$ process:

$$X_t = \sum_{j=0}^{\infty} \alpha^j Z_{t-j},$$

The AR and MA processes can be generalized as ARMA processes

Definition

We call **autoregressive moving average process of order p, q** any process X that can be written as :

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p} = \beta_0 Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}$$

This class of processes is very famous in time series analysis. The forecasting methods we have just seen can directly be applied to ARMA processes.

The optimal 1 step ahead forecast is:

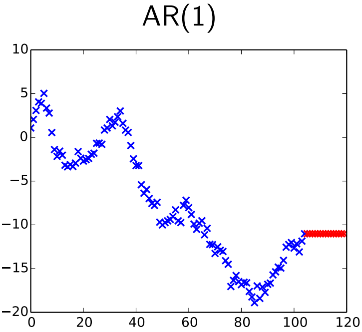
$$\hat{X}_t = E[X_t | X_{t-1}, \dots, X_0]$$

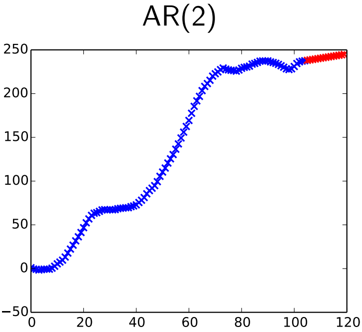
In practice, we focus on the best linear predictor.

Exercise

- Show that, for an AR process, $\hat{X}_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p}$
- What about the two step ahead forecast?

We obtain for the previous examples:





For such process, the optimal forecast is not as straightforward as before. We are looking for the values $a_0, \dots, a_n$ that minimises the error:

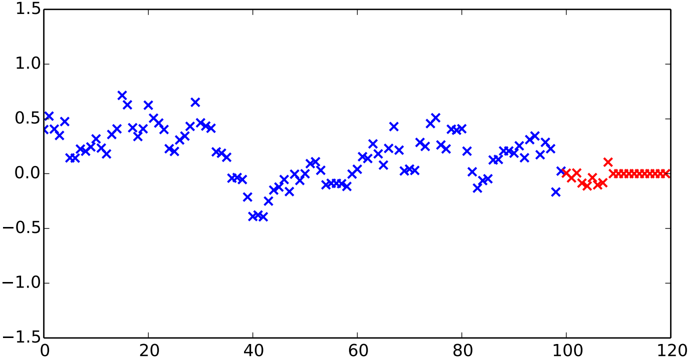
$$E[(X_{n+h} - a_0 - a_1 X_n - \dots - a_n X_1)^2] = E[(X_{n+h} - \mathbf{a}^t \mathbf{X_p})^2]$$

This is quadratic in **a** so we can differentiate this expression and find where it is null. We finally obtain

$$\begin{aligned} \mathbf{a} &= E[\mathbf{X_p}\mathbf{X_p}^t]^{-1} E[X_{n+h}\mathbf{X_p}] \\ &= \begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \dots & \gamma(n) \\ \gamma(1) & \gamma(0) & \gamma(1) & \dots & \gamma(n-1) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \gamma(n) & \gamma(n-1) & \dots & \dots & \gamma(0) \end{pmatrix}^{-1} \begin{pmatrix} \gamma(h) \\ \gamma(h+1) \\ \vdots \\ \gamma(h+n) \end{pmatrix} \end{aligned}$$

This approach is general and it also stands for AR processes.

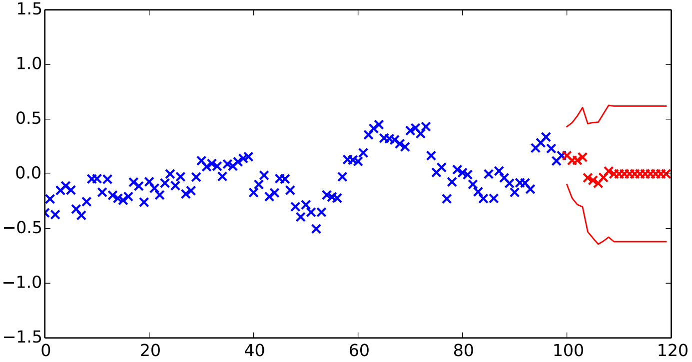
If we consider the previous example we obtain:



A great asset of probabilistic models is to provide not only a "mean predictor" but also a quantification of the uncertainty:

$$v^2(t,h) = Var[X_{t+h}|X_t, \dots, X_0]$$

In the Gaussian case, confidence intervals are $mean \pm 1.96sd$



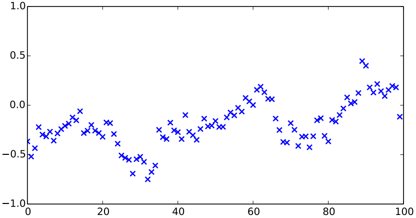
In the previous section we have seen that the predictions can be made for AR, MA and ARMA models when the autocovariance is known.

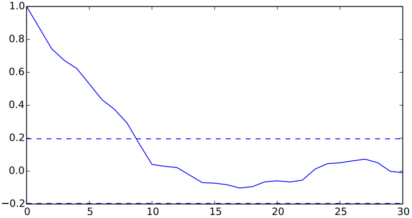
In practice, we just observe the data so the model's parameters:

- the orders p and q
- the parameters $\alpha_1, \ldots, \alpha_p$ and $\beta_0, \ldots, \beta_q$

are unknown and they have to be estimated from the data.

We have seen that for $h > q$, the autocorrelation of a MA is null: $\gamma(h) = 0$. This can be used to estimate q:





in this example we had $q = 10$.

In a similar fashion, for $h > q$, the *partial autocorrelation* of an AR is null $\pi(h) = 0$, where

$$\pi(h) = corr(X_t - E[X_t|X_{t+1}, ..., X_{t+h-1}] \, , \, X_{t+h} - E[X_{t+h}|X_{t+1}, ..., X_{t+h-1}])$$

This can be used to estimate p.

Various methods allow to estimate the model parameters:

- Computation of the *sample autocovariance function* and *sample partial autocovariance function*.
- least square estimation
- maximum likelihood estimation

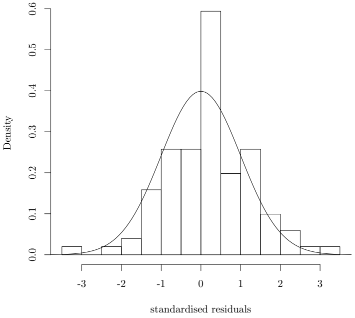
We do not have time do detail these methods which are already implemented in various packages.

A good forecast is a forecast that has a small prediction error. This can be measured with the Mean Square Error :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}(t_i) - y(t_i))^2$$

In order to measure this quantity, we can compare the **forecast with reality on past data**.

Furthermore, if the model provides some uncertainty measure on the forecasts, it is important to validate it as well:



Three important points to remember :

- Forecasts **cannot be perfect** but we want them to be **as accurate as possible**
- Time series analysis is a well suited framework for prediction
- **Validating the model** is of the utmost importance. It can be done on past data.

We will illustrate these notions during next week lab session.

Finally

- The probabilistic framework is of great interest in forecasting
  - it allows to study the optimality of a forecast
  - it can be used to estimate parameters
- Advanced models such as ARMA or Holt-Winters can capture specific patterns such as trend or seasonality
- Methods are already implemented in statistical software