Métodos de Previsão e Séries Temporais

Mestrado em Estatística para Ciência de Dados

A. Manuela Gonçalves
mneves@math.uminho.pt

Departamento de Matemática
http://www.math.uminho.pt

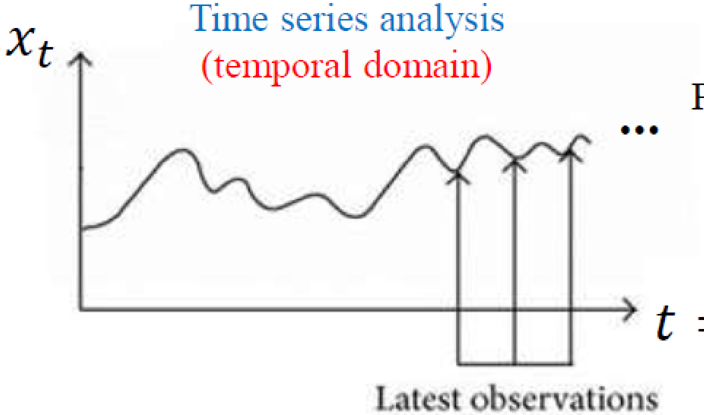
Part I – Concepts

Introduction

Time Series and Spectral Analysis

What is a time series? A **time series** is a set of observations obtained by measuring a single variable regularly over a period of time.

Time series analysis
(temporal domain)

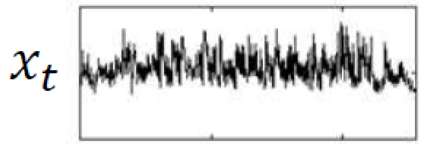


$t = 1,2,\dots,n$

Prediction of future values
of the time series
(Model forecasting)

Knowledge improvement of the underlying
mechanism that generates the time series
(Data decomposition or Model fitting)

Time series analysis
(temporal domain)

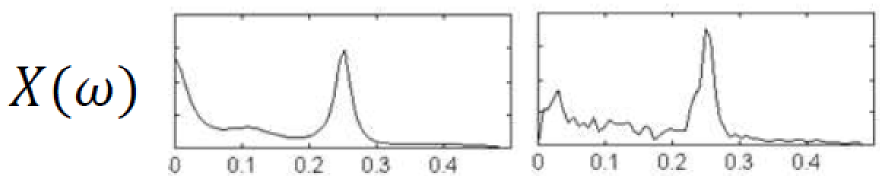


Time scale t (e.g. sec)

Spectral analysis
(frequency domain)

Parametric
(model-driven)

Non-parametric
(data-driven)



Frequency scale ω (e.g. Hz $= \frac{cycles}{sec}$)

Introduction

Time Series and Spectral Analysis

Ideas (Shumway and Stoffer, 2011):

The first step in any time series investigation always involves careful examination of the recorded data plotted over time. This scrutiny often suggests the method of analysis as well as statistics that will be of use in summarizing the information in the data.

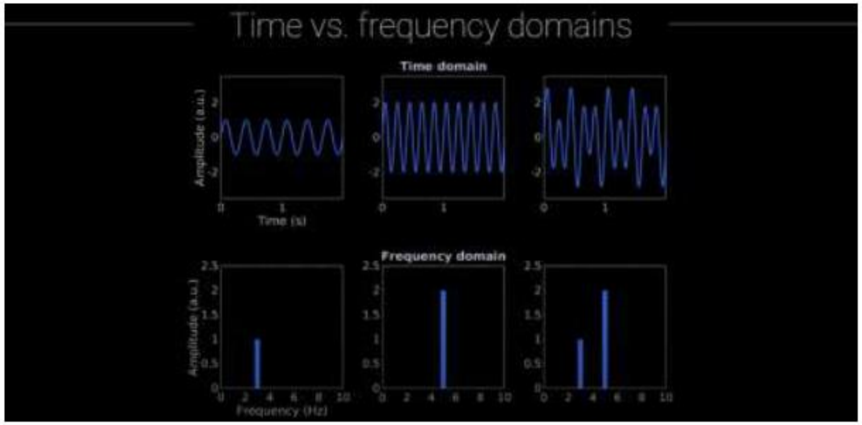
There are two separate approaches (not necessarily mutually exclusive) commonly identified as

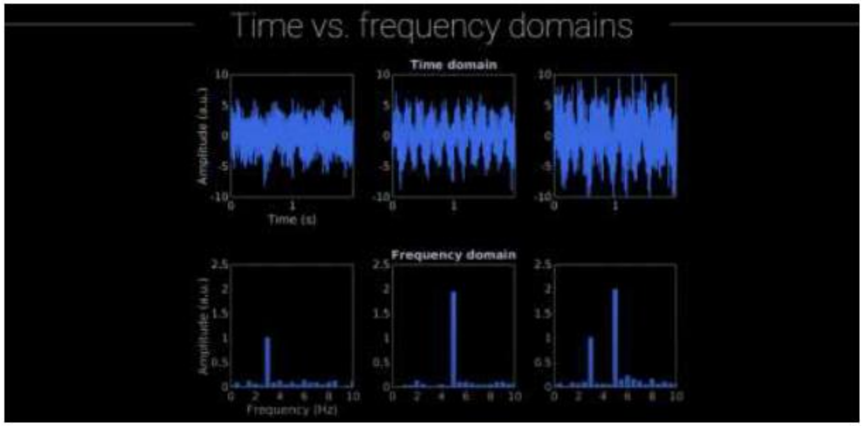
- Time domain approach: correlation explained in terms of the dependency of the presente with respect to the past values of the series → **temporal analysis**
- Frequency domain approach: characterization of the time series as a function of its periodic (sinusoidal) variation → **spectral analysis**

Introduction

Time Series and Spectral Analysis

The time domain displays the changes in a signal (information) over a span of time whereas the frequency domain displays how much of the signal exists within a given frequency band concerning a range of frequencies.





Time and frequency domains (10 min video) by Mike X Cohen

https://www.youtube.com/watch?app=desktop&v=fYtVHhk3xJ0

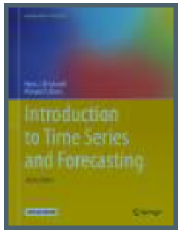
Introduction

Recommended bibliography

- Shumway, R.H. and Stoffer, D.S. (2011), Time Series Analysis and its Applications with R examples (4th ed), Springer texts in Statistics. (https://www.stat.pitt.edu/stoffer/tsa4/tsa4.pdf)

- Brockwell, P.J. and Davies, R.A. (2016), Introduction to Time Series and Forecasting (3rd ed), Springer Texts in Statistics. (https://link.springer.com/content/pdf/10.1007%2F978-3-319-29854-2.pdf)

- Hyndman, R.J. and Athanasopoulos, G. (2021), Forecasting: Principles and Practice (3rd ed), OTexts: Melbourne, Australia. (https://otexts.com/fpp3/)



Other (more specific) references along the text of the course…

Introduction

Packages in R for time series analysis

https://cran.r-project.org/web/views/TimeSeries.html

astsa = applied statistical time series analysis

- https://cran.r-project.org/web/packages/astsa/astsa.pdf
- Shumway, R.H. and Stoffer, D.S. (2011), Time Series Analysis and its Applications with R examples (4th ed), Springer texts in Statistics. (https://www.stat.pitt.edu/stoffer/tsa4/tsa4.pdf)

forecast

- https://cran.r-project.org/web/packages/forecast/forecast.pdf
- Hyndman, R.J. and Athanasopoulos, G. (2021), Forecasting: Principles and Practice (3rd ed), OTexts: Melbourne, Australia. (https://otexts.com/fpp3/)

Packages in R for visualization and graphics

https://cran.r-project.org/web/views/Graphics.html

https://www.r-graph-gallery.com/index.html

ggplot2 (https://cran.r-project.org/web/packages/ggplot2/index.html)

gganimate (https://gganimate.com/)

Introduction

Fundamental concepts in time series

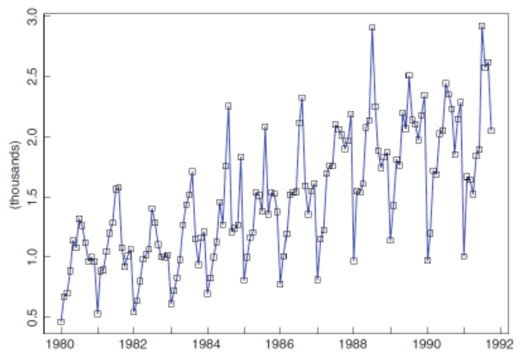


Figure 1-1
The Australian red wine sales, Jan. 1980–Oct. 1991

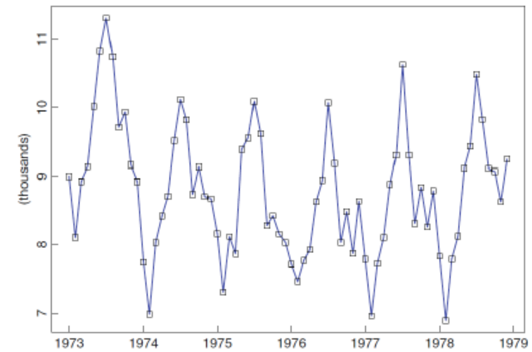


Figure 1-3
The monthly accidental deaths data, 1973–1978

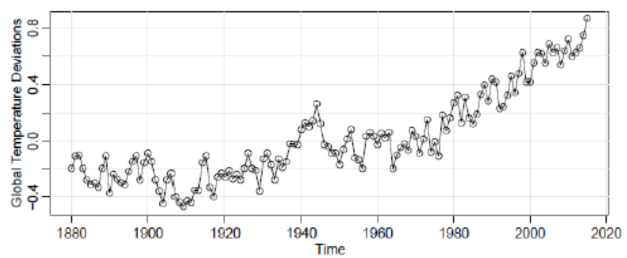


Fig. 1.2. Yearly average global temperature deviations (1880–2015) in degrees centigrade.

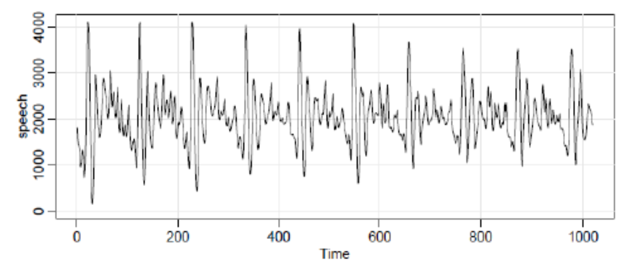


Fig. 1.3. Speech recording of the syllable aaa ⋯ hhh sampled at 10,000 points per second with $n = 1020$ points.

Figures from Brockwell and Davies (2016) and Shumway and Stoffer (2010).

Introduction

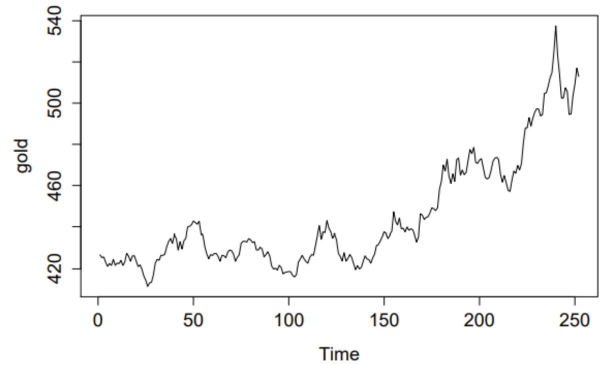


Figure 4: Daily price of gold (in dollars per ounce) for the 252 trading days of 2005

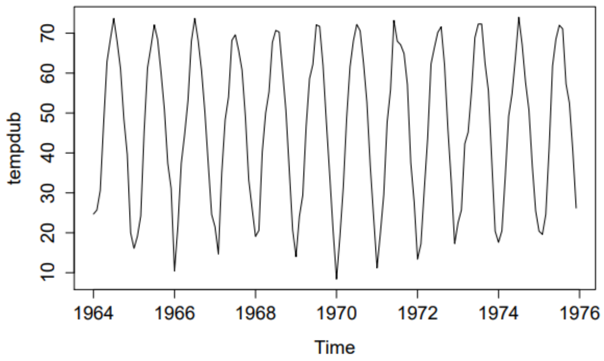


Figure 5: Monthly average temperature (in degrees Fahrenheit) in Dubuque

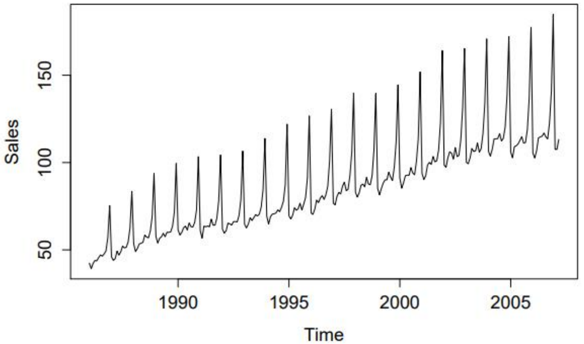


Figure 6: Annual sales of certain large equipment

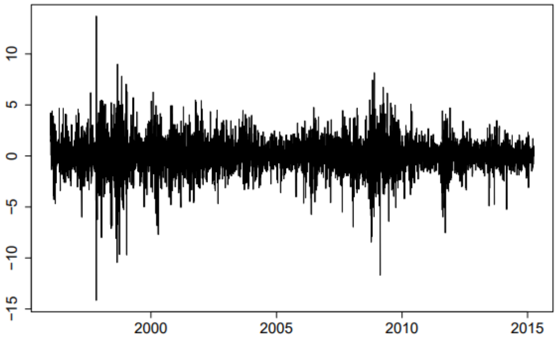
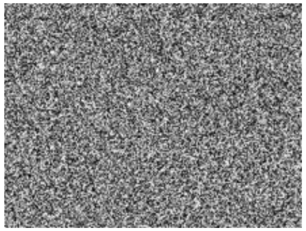


Figure 7: Daily returns of the WIG20 index from January 2, 1996 until March 31, 2015

Simple Models

White Noise*

*The designation *white* originates from the analogy with white light where all possible periodic oscillations are present with equal strength



White noise: sound and video

https://www.youtube.com/watch?v=t0I4mTEdAf8

- **white noise**: $e_t \sim \mathrm{WN}(0, \sigma_e^2)$

Set of uncorrelated random variables with $\mathrm{E}(e_t) = 0$ and $\mathrm{Var}(e_t) = \sigma_e^2 < \infty$.

- **white independent noise:** $e_t \sim \mathrm{iid}(0, \sigma_e^2)$

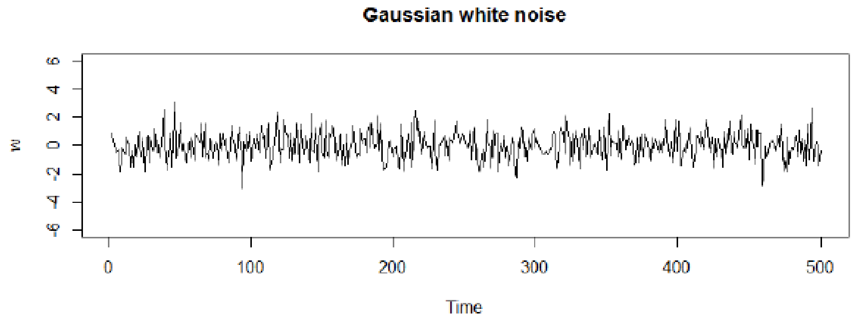
Set of independent and identically distributed (iid) random variables with $\mathrm{E}(e_t) = 0$ and $\mathrm{Var}(e_t) = \sigma_e^2 < \infty$.

- **gaussian white noise:** $e_t \sim \mathrm{iid}\ \mathrm{N}(0, \sigma_e^2)$

Set of iid random variables with normal distribution $\mathrm{N}(0, \sigma_e^2)$.

Simple Models

White Noise




```
set.seed(1234) # reproduce the results, if needed
w = rnorm(500, mean = 0, sd = 1) # 500 random numbers generated from N(0,sd^2) distribution
plot.ts(w, ylim=c(-6,6), main = "Gaussian white noise")
boxplot(w) # boxplot display
hist(w,probability = TRUE) # histograma display
lines(density(w)) # adds the kernel density (smooth estimate of the probability function)
qqnorm(w) # Quantile-Quantile plot = QQplot
qqline(w) # adds reference line to the Qqplot
cor(x = w[1:length(w)-1], y = w[2:length(w)], method = "pearson")
cor.test(x = w[1:length(w)-1], y = w[2:length(w)], method = "pearson")
```

If the stochastic behavior of a time series could be explained in terms of the white noise model, classical statistical methods would suffice. If not, how to introduce more smoothness into a time series models or even serial correlation?

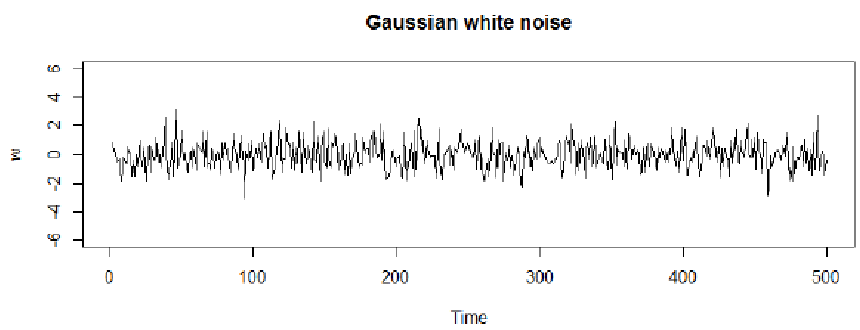
Simple Models

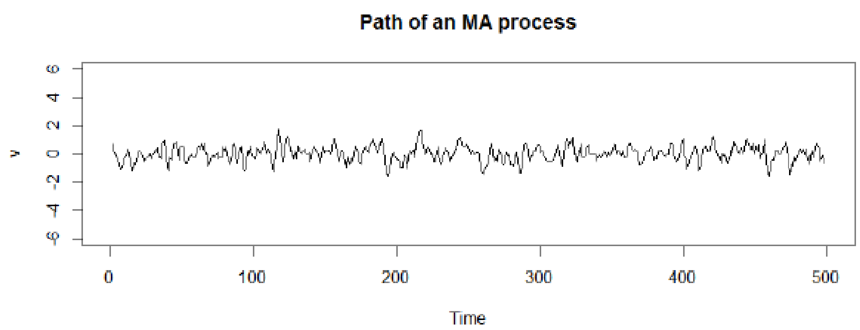
Moving Averages and Filtering

Smoothness can be introduced into time series models with a *moving average* (MA) process. E.g.,

$$X_t = \frac{1}{3}(e_{t-1} + e_t + e_{t+1}) \qquad t = 1,2,\dots,n$$

considers an average of its current value and its immediate neighbors (past and future) and constitutes a three-point moving average of the gaussian white noise.





The resulting series is a smoother version of the first series, reflecting the fact that the slower oscillations are more apparent and some of the faster oscillations are taken out.

Simple Models

Autoregressions

Serial correlation can be introduced in a time series model with an *autoregression* (AR).

Considering a white noise as input, the second-order equation

$$X_t = X_{t-1} - 0.9X_{t-2} + e_t \qquad t = 1,2,\ldots$$

represents a regression of X_t, the current value of a time series, as a function of X_{t-1} and X_{t-2}, the past two values of the series (autoregression).

There is an issue with the AR startup because, for $t = 1$, the above equation becomes

$$X_1 = X_0 - 0.9X_{-1} + e_1.$$

Thus, X_0 and X_{-1} must be set initially to generate the succeeding values of the time series. One way to deal with this shortcoming is to set $X_0 = X_{-1} = 0$, i.e.

$$X_1 = e_1, X_2 = X_1 + e_2, X_3 = X_2 - 0.9X_1 + e_3, \ldots$$

run the equation for longer than needed and remove the initial AR values.

Simple Models

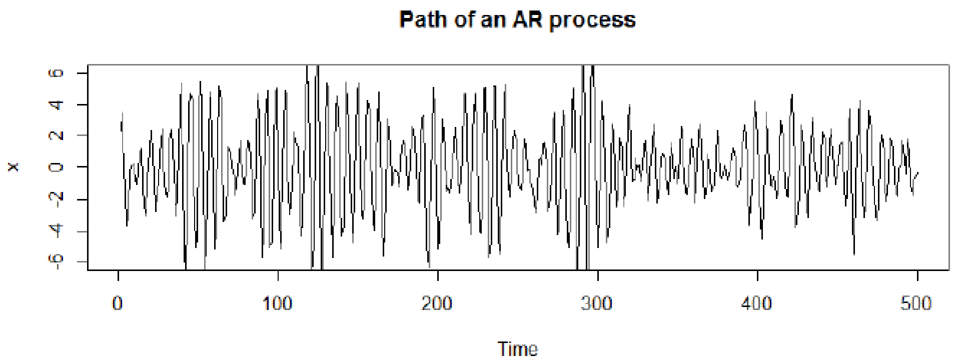
Autoregressions

Serial correlation can be introduced in a time series model with an *autoregression* (AR).

Considering a white noise as input, the second-order equation

$$X_t = X_{t-1} - 0.9X_{t-2} + e_t \qquad t = 1,2,\dots$$

represents a regression of X_t, the current value of a time series, as a function of X_{t-1} and X_{t-2}, the past two values of the series (autoregression).



The resulting series shows a periodic behavior, and it seems to be more predictable than the gaussian white noise (serial correlation).

Simple Models

Simulation of MA and AR paths by filtering

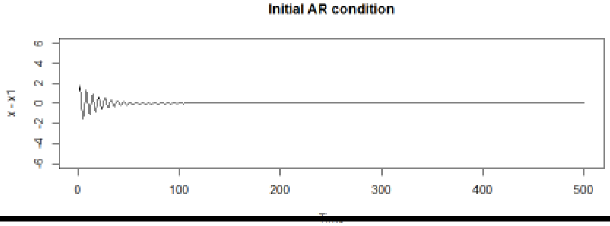
A linear combination of values in a time series is referred to, generically, as a filtered series; hence the command *filter* is used to generate MA and AR paths.

```
# White noise
set.seed(1234) # reproduce the results, if needed
sd = 1
w = rnorm(500, mean = 0, sd = sd) # 500 random numbers generated independently from N(0,sd^2) distribution
plot.ts(w, ylim=c(-6,6), main = "Gaussian white noise")

# Path of an MA process
v = filter(w, sides=2, filter=rep(1/3,3))
plot.ts(v, ylim=c(-6,6), main="Path of an MA process")

# Path of an AR process
w1 = c(rnorm(50, mean = 0, sd = sd), w) # generate extra 50 to deal with the startup problems
x = filter(w1, filter=c(1,-.9), method="recursive") [-(1:50)] # filter and remove the first 50
plot.ts(x, ylim=c(-6,6), main="Path of an AR process")

# Evaluate the effect of the initial AR condition
x1 = filter(w, filter=c(1,-.9), method="recursive")
plot.ts(x-x1 , ylim=c(-6,6), main="Initial AR condition")
```



Simple Models

Random Walk

The random walk model is given by

$$X_t = e_1 + e_2 + \cdots + e_{t-1} + e_t = X_{t-1} + e_t$$

for $t = 1,2,\dots$ with initial condition $X_0 = 0$ and where e_t is white noise. The term random walk comes from the fact that the value of X_t is the value of X_{t-1} plus a completely random movement determined by e_t.

The random walk with drift model extends the previous as

$$X_t = \delta + X_{t-1} + e_t$$

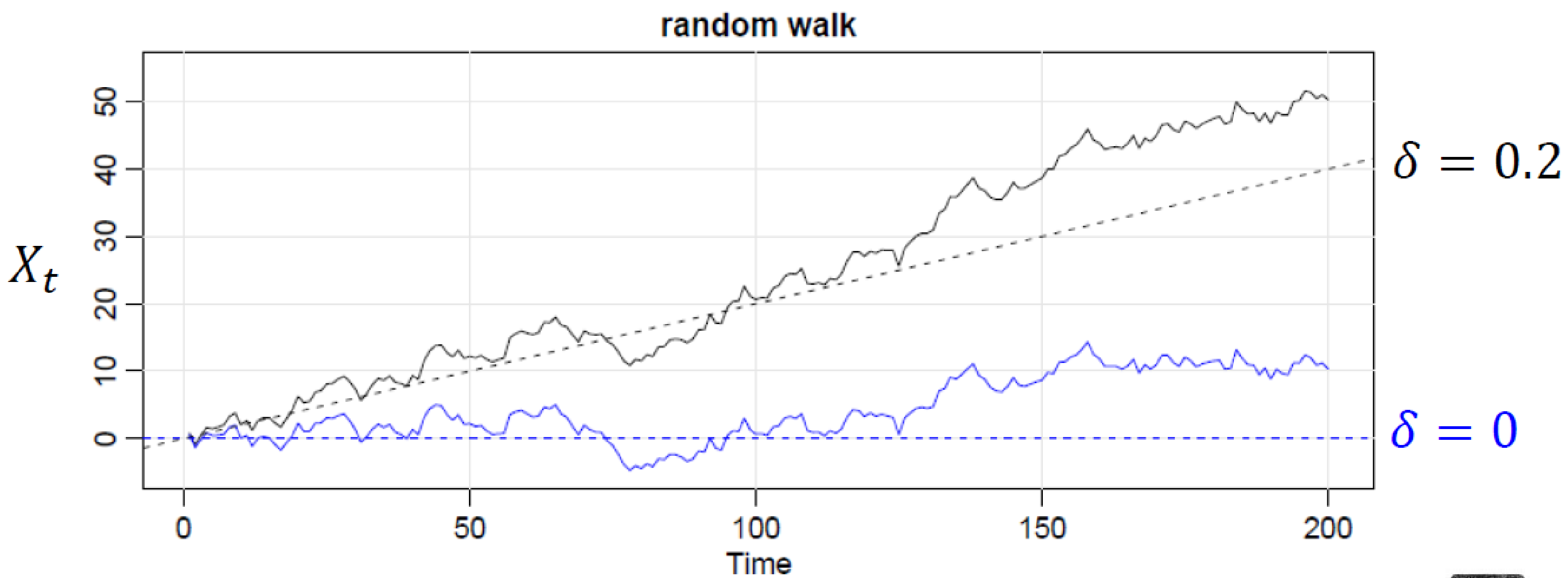
where the constant δ is called *drift* (when $\delta = 0$ there is no drift). This model can be rewrite as a cumulative sum of white noise variates

$$X_t = \delta t + \sum_{i=1}^{t} e_i$$

which highlights the existence of a linear trend component in X_t.

Simple Models

Random Walk




```
set.seed(154) # so results are reproducible
w = rnorm(200); x = cumsum(w) # two commands in one line
wd = w +.2; xd = cumsum(wd)
plot.ts(xd, ylim=c(-5,55), main="random walk", ylab='')
lines(x, col=4); abline(h=0, col=4, lty=2); abline(a=0, b=.2, lty=2)
```

Simple Models

Signal in Noise

Many realistic models for generating time series assume an underlying signal with some consistent periodic variation contaminated by adding a random noise. E.g.,

$$X_t = 2\cos\left(2\pi\frac{t+15}{50}\right) + e_t \qquad t = 1,2,\ldots$$

where the first term is regarded as the *signal* and e_t the *noise*.

In general, a sinusoidal waveform can be written as

$$A\cos(2\pi\omega t + \phi)$$

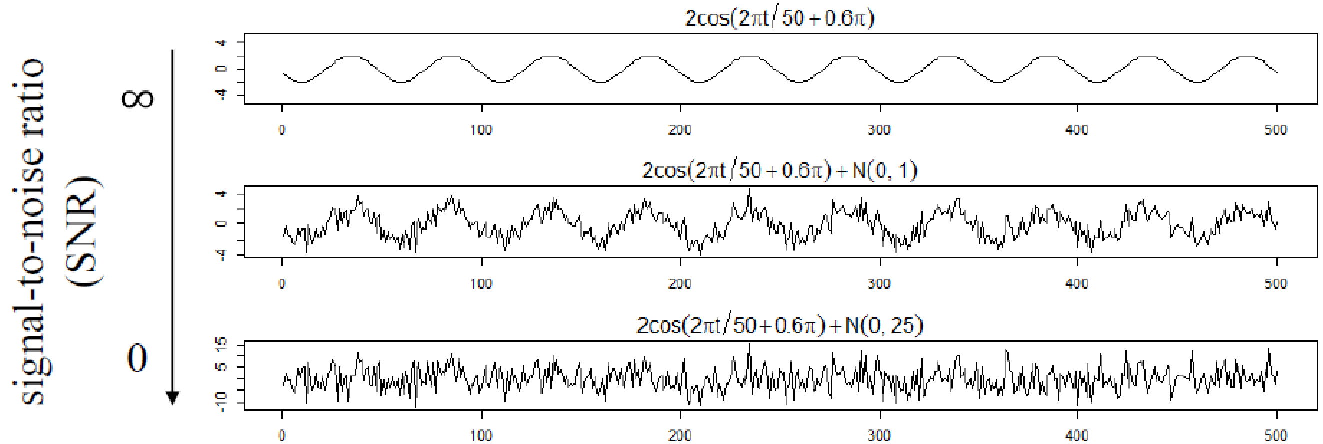
where A is the amplitude, ω is the frequency of oscillation, and ϕ is a phase shift. In the example,

$$2\cos\left(2\pi\frac{t+15}{50}\right) = 2\cos\left(2\pi\frac{1}{50}t + 2\pi\frac{15}{50}\right)$$

so that $A = 2$, $\omega = 1/50$ (one cycle every 50 time points) and $\phi = 0.6\pi$.

Simple Models

Signal in Noise



In this example, the noise was taken as white noise with $\sigma_e^2 = 1$ and $\sigma_e^2 = 25$.

```
cs = 2*cos(2*pi*1:500/50 + .6*pi); w = rnorm(500, mean = 0, sd = 1)
par(mfrow=c(3,1), mar=c(3,2,2,1), cex.main=1.5)
plot.ts(cs, ylim=c(-5,5), main=expression(2*cos(2*pi*t/50+.6*pi)))
plot.ts(cs+w, main=expression(2*cos(2*pi*t/50+.6*pi) + N(0,1)))
plot.ts(cs+5*w, main=expression(2*cos(2*pi*t/50+.6*pi) + N(0,25)))
```

The addition of the noise obscures the signal (cosine waveform), depending on the amplitude of the signal and the variability of e_t.

Dependence

Measures of dependence

A complete description of a time series, observed as a collection of n random variables

$$X_1, X_2, \ldots, X_n$$

at arbitrary time points $-\infty < t_1 < t_2 < \ldots < t_n < +\infty$, with $n \in \mathbb{N}$, is provided by the joint distribution function, evaluated as the probability that the values of the series are jointly less or equal than the n constants, $x_1, x_2, \ldots, x_n$ i.e.,

$$F_{X_1,\, X_2,\, \ldots,\, X_n}(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n).$$

Unfortunately, these multidimensional distribution functions cannot usually be written easily unless the random variables are

- purely random (iid) or
- jointly normal (gaussian processes).

Although the joint distribution function describes the data completely, it is an unwieldy tool for displaying and analyzing time series data (function of n variables).

Dependence

Measures of dependence

Alternatively, the marginal distribution functions

$$F_t(x) = P(X_t \leq x)$$

or the corresponding marginal density functions

$$f_t(x) = \frac{\partial F_t(x)}{\partial x}$$

when existing, are often informative for examining the marginal behavior of a series.

Moreover, one should consider the first and the second order moments of the joint distributions, namely

- expected value,
- autocovariance function (ACVF),
- autocorrelation (normalized autocovariance) function (ACF).

Dependence

Expected value (or mean)

The mean function is defined as

$$\mu_t = \mathrm{E}(X_t) = \int_{-\infty}^{+\infty} x.\, f_t(x)\, dx$$

provided it exists, where $f_t(x)$ is the marginal density function of X at a given time t and E denotes the usual expected value operator.

Properties: Being X and Y two random variables and $a, b \in \mathbb{R}$, then

- $\mathrm{E}(a) = a$;
- $\mathrm{E}(aX) = a\mathrm{E}(X)$;
- $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$;
- $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$, if X and Y are independent;
- $\mathrm{E}(g(X)) = \int_{-\infty}^{+\infty} g(x).\, f(x)\, dx$ represents the expected value of any function of X

Dependence

Expected value (or mean)

- If e_t denotes a white noise series, then $\mathrm{E}(e_t) = 0$ for all t.
- For the smoothed series $X_t = \frac{1}{3}(e_{t-1} + e_t + e_{t+1})$

$$\mathrm{E}(X_t) = \mathrm{E}\left(\frac{1}{3}(e_{t-1} + e_t + e_{t+1})\right) = \frac{1}{3}\mathrm{E}(e_{t-1}) + \frac{1}{3}\mathrm{E}(e_t) + \frac{1}{3}\mathrm{E}(e_{t+1}) = 0.$$

- For the random walk with drift $X_t = \delta t + \sum_{i=1}^{t} e_i$,

$$\mathrm{E}(X_t) = \mathrm{E}\left(\delta t + \sum_{i=1}^{t} e_i\right) = \delta t + \sum_{i=1}^{t} \mathrm{E}(e_i) = \delta t$$

which is a straight line with slope δt.

- For the noisy cosine model $X_t = 2\cos\left(2\pi \frac{t+15}{50}\right) + e_t$,

$$\mathrm{E}(X_t) = \mathrm{E}\left(2\cos\left(2\pi\frac{t+15}{50}\right) + e_t\right) = 2\cos\left(2\pi\frac{t+15}{50}\right)$$

which corresponds to the deterministic part of the model.

Dependence

Autocovariance function

The autocovariance function is defined as the second order product

$$\gamma_X(t,s) = \mathrm{Cov}(X_t, X_s) = \mathrm{E}\big((X_t - \mu_t)(X_s - \mu_s)\big)$$

for all time points $t, s \in \{1,2,\dots\}$.

Interpretation: This function measures the linear dependence between two points on the same series observed at different times:
Very smooth series have autocovariance functions that stay large even when the t and s are far apart, whereas choppy series tend to have autocovariance functions that are nearly zero for large separations.

Properties (from definition):

- $\gamma_X(t,s) = \gamma_X(s,t)$ for all t and s.
- for $t = s$, the autocovariance reduces to the (assumed finite) variance, as
  $\gamma_X(t,t) = \mathrm{Cov}(X_t, X_t) = \mathrm{E}\big((X_t - \mu_t)^2\big) = \sigma_t^2 = \mathrm{Var}(X_t)$

Dependence

Autocovariance function

In classical statistics, $\text{Cov}(X, X) = \text{Var}(X)$ where Var denotes the usual variance operator. Being X and Y two random variables and $a, b \in \mathbb{R}$, then

- $\text{Var}(a) = 0$;
- $\text{Var}(aX + b) = a^2 \,\text{Var}(X)$;
- $\text{Var}(aX \pm bY + c) = a^2 \,\text{Var}(X) + b^2 \,\text{Var}(X)$, if X and Y are independent;
- In general, $\text{Var}(aX \pm bY + c) = a^2 \,\text{Var}(X) + b^2 \,\text{Var}(X) \pm 2ab \,\text{Cov}(X, Y)$;

Note: If X and Y are independent then $\text{Cov}(X, Y) = 0$.

- $\text{Var}(X) = \text{E}(X^2) - \text{E}(X)^2$.

Challenge: Consider a random sample* $(X_1, X_2, \ldots, X_n)$ of a population X with $\text{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$. Show that the sample average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has $\text{E}(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/n$.

* A random sample is a set of iid random variables.

Dependence

Autocovariance function

Moving to the time series context, recall

- if $\gamma_X(t,s) = 0$ then X_t and X_s are not linearly related but there can be some dependence structure between them;
- if X_t and X_s are bivariate normal, $\gamma_X(t,s) = 0$ implies that X_t and X_s are independent.

Challenge: Show that the white noise e_t with $\mathrm{Var}(e_t) = \sigma_e^2$ has

- expected value $\mathrm{E}(e_t) = 0$ and
- autocovariance function given by

$$\gamma_e(t,s) = \begin{cases} \sigma_e^2 \,, t = s \\ 0 \,, t \neq s \end{cases}$$

Dependence

Autocovariance function

There is often the need to calculate the autocovariance in filtered series. A useful result is given in the following proposition.

Proposition: Covariance of linear combinations

If the random variables U and V are linear combinations of (finite variance) random variables X_j and Y_k, respectively, i.e.

$$U = \sum_{j=1}^{m} a_j \, X_j \quad \text{and} \quad V = \sum_{k=1}^{r} b_k \, Y_k$$

then

$$\text{Cov}(U, V) = \sum_{j=1}^{m} \sum_{k=1}^{r} a_j \, b_k \, \text{Cov}(X_j, Y_k) \, .$$

Furthermore, $\text{Cov}(U, U) = \text{Var}(U)$.

Dependence

Autocovariance function of an MA process

The autocovariance of the MA process $X_t = \frac{1}{3}(e_{t-1} + e_t + e_{t+1})$ is

$$\gamma_X(t,s) = \mathrm{Cov}(X_t, X_s) = \frac{1}{9}\mathrm{Cov}(e_{t-1} + e_t + e_{t+1}, e_{s-1} + e_s + e_{s+1})$$

and further result will depend on how t relates with s.

E.g., when $t = s$,

$$\gamma_X(t,t) = \frac{1}{9}\big(\mathrm{Cov}(e_{t-1}, e_{t-1}) + \mathrm{Cov}(e_{t-1}, e_t) + \cdots + \mathrm{Cov}(e_{t+1}, e_{t+1})\big) =$$

$$= \frac{1}{9}\big(\mathrm{Cov}(e_{t-1}, e_{t-1}) + \mathrm{Cov}(e_t, e_t) + \mathrm{Cov}(e_{t+1}, e_{t+1})\big) = \frac{3}{9}\sigma_e^2$$

When $t = s + 1 \Leftrightarrow t - 1 = s \Leftrightarrow t - s = 1$ then

$$\gamma_X(t, t-1) = \frac{1}{9}\big(\mathrm{Cov}(e_{t-1} + e_t + e_{t+1}, e_{t-2} + e_{t-1} + e_t)\big) = \frac{2}{9}\sigma_e^2$$

and the same result is obtained for $t = s - 1 \Leftrightarrow t - s = -1$.

Dependence

Autocovariance function of an MA process

and so on that it can be written as

$$\gamma_X(t,s) = \begin{cases} \frac{3}{9}\sigma_e^2, \ t = s \\ \frac{2}{9}\sigma_e^2, \ |t-s| = 1 \\ \frac{1}{9}\sigma_e^2, \ |t-s| = 2 \\ 0, \ |t-s| \geq 3 \end{cases}$$

This result clearly shows that the smoothing operation introduces

- a covariance function that decreases as $h = |t-s|$ (the separation between the two time points) increases and
- disappears completely when $h \geq 3$.

This is interesting because the ACVF depends on the time separation (or lag) h and not on the absolute location of t and s. This is related with the concept of stationarity.

Dependence

Autocovariance function of a random walk

For the random walk model $X_t = \sum_{i=1}^{t} e_i$,

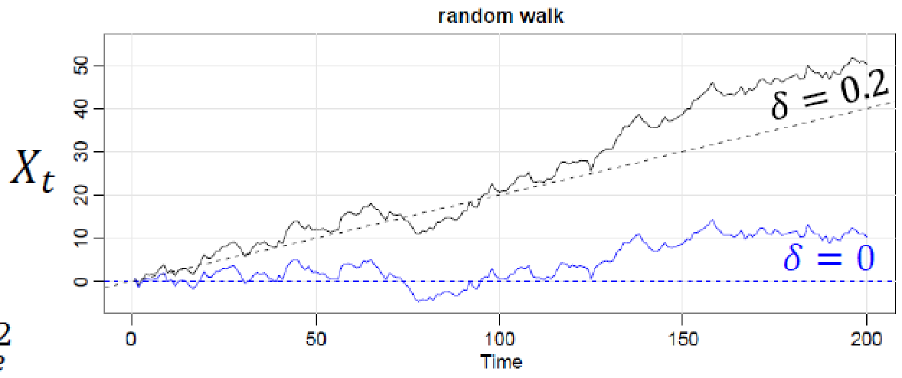
$$\gamma_X(t,s) = \text{Cov}(X_t, X_s) = \text{Cov}\left(\sum_{j=1}^{t} e_j\,, \sum_{k=1}^{s} e_k\right) = \min(t,s)\,\sigma_e^2$$

because e_t are uncorrelated random variables. This autocovariance depends on the particular time values t and s, and not on the time separation or lag.

Then, the variance is $\text{Var}(X_t) = \gamma_X(t,t) = t\,\sigma_e^2$, which increases without bound as time t increases.

Therefore, for this random walk

- $\text{E}(X_t) = 0$,
- $\text{Var}(X_t) = \gamma_X(t,t) = t\,\sigma_e^2$,
- $\gamma_X(t,s) = \text{Cov}(X_t, X_s) = \min(t,s)\,\sigma_e^2$



Dependence

Autocovariance function

As in classical statistics, it is more convenient to deal with a measure of association that varies between -1 and 1. This leads to the autocorrelation function (ACF)

$$\rho_X(t,s) = \frac{\gamma_X(t,s)}{\sqrt{\gamma_X(t,t)\,\gamma_X(s,s)}}$$

The ACF measures the linear predictability of X_t using X_s. It is easily shown that

$$-1 \leq \rho_X(t,s) \leq 1$$

e.g. using the Cauchy–Schwarz inequality, which implies that $|\gamma_X(t,s)|^2 \leq \gamma_X(t,t)\,\gamma_X(s,s)$.

If X_t can be predicted *perfectly* from X_s through a linear relationship

$$X_t = \beta_0 + \beta_1 X_s$$

then the correlation will be $+1$ when $\beta_1 > 0$ and -1 when $\beta_1 < 0$. Hence, $\rho_X(t,s)$ is a rough measure of the ability to forecast the series at time t from the value at time s.

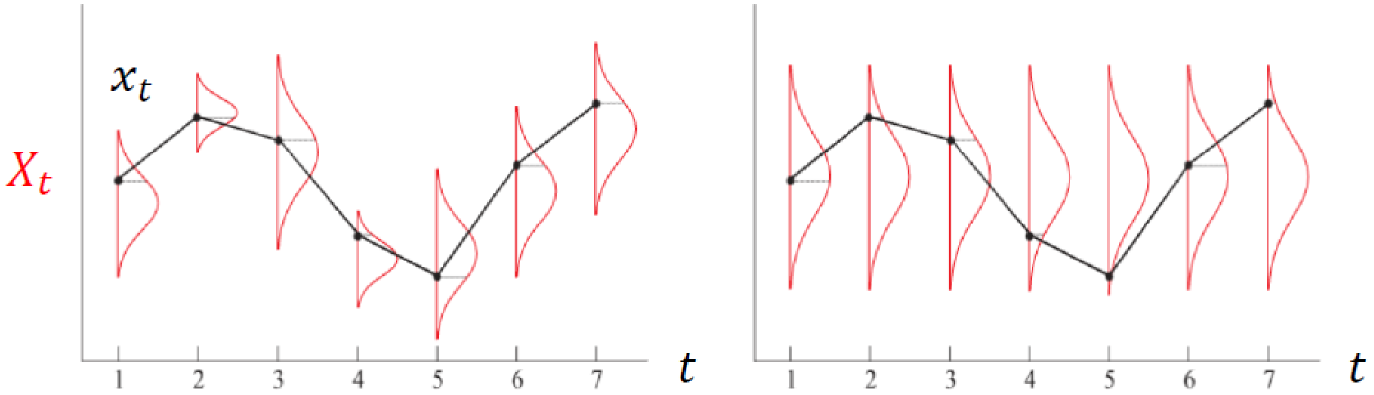
Stationary models

Stationarity

The preceding definitions of mean and autocovariance are completely general and can vary as a function of t. The notion of *regularity* is here introduced using a concept called stationarity.

Stochastic Process X_t versus **Time Series x_t**

A stochastic process X_t is a collection of random variables ordered by an index set t (usually time). A **time series x_t** is a sample path (a realization) of the stochastic process.



Stationary models

Stationarity

Stationarity is a rather intuitive concept, meaning that the statistical properties of the process do not change over time.

Loosely speaking, the process X_t, $t = 1,2, \ldots$ is said to be stationary if its statistical properties are similar to those of the "time-shifted" process X_{t+h}, $t = 1,2, \ldots$ for each integer h.

STRONG

There are two important definitions of stationarity: WEAK

- strong (or strict) stationarity;

Strong stationarity concerns the shift-invariance (in time) of its finite-dimensional distributions.

- weak (or wide-sense) stationarity.

Weak stationarity concerns the shift-invariance (in time) of the first and second moments of a process (mean and autocovariance). Thus, weak stationarity is defined by restricting attention to those properties that depend only on the first- and second-order moments of X_t.

Stationary models

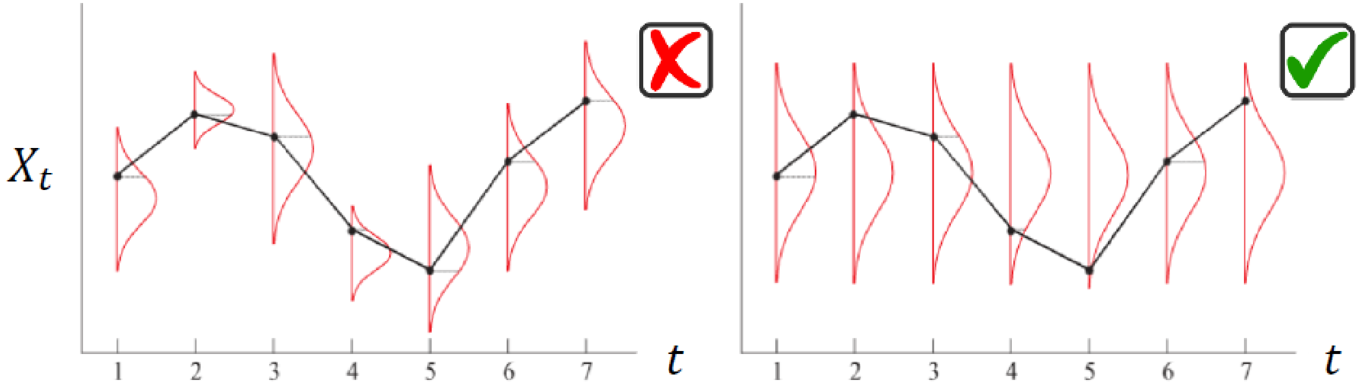
Strong stationarity

The process $\{X_t, t \in \mathbb{N}\}$ is strongly stationary if

$$F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = F_{X_{1+h}, X_{2+h}, \dots, X_{n+h}}(x_1, x_2, \dots, x_n)$$

for all $h \in \mathbb{N}$ with any finite set of indices $\mathrm{t} \in \{1, 2, \dots, n\} \subset \mathbb{N}$ and $n \in \mathbb{N}^+$. Here, F denotes the joint distribution of n random variables.

Note: If $\{X_t, t \in \mathbb{N}\}$ is strongly stationary then
(i) $X_1, X_2, \dots$ have the same probability distribution function ($n = 1$) and
(ii) the joint distribution of $(X_1, \dots, X_n)$ is invariant under translation ($n > 1$).



Stationary models

Weak stationarity

The concept of weak stationarity can be defined by restricting attention to those properties that depend only on the first- and second-order moments of X_t.

The process $\{X_t,\, t \in \mathbb{N}\}$ is weakly stationary if

(1) $\mathrm{E}(X_t) = \mu_t = \mu$, i.e. μ_t is independent of t;

(2) $\mathrm{Cov}(X_t, X_{t+h}) = \gamma_X(t, t+h) = \gamma_X(h)$, i.e. γ_X is independent of t for each h;

(3) the second moment of X_t is finite for all t, i.e. $\mathrm{E}\left({X_t}^2\right) < \infty$ for all t.

Note: With respect to condition (2), the covariance at any two time points, t and s,

$$\mathrm{Cov}(X_t, X_s) = \mathrm{Cov}(X_t, X_{t+h})$$

i.e. depends only on $h = s - t \in \mathbb{Z}$, the difference between the two time points and not the on the location of the points along the time axis.

Stationary models

Relation between strong and weak stationarity

- Finite second moments are not assumed in the definition of strong stationarity; therefore, strong stationarity does not necessarily imply weak stationarity*.

*an iid process with standard Cauchy distribution is strictly stationary but not weak stationary because the second moment of the process is not finite.

- If $\{X_t, t \in \mathbb{N}\}$ is strongly stationary with $\mathrm{E}\left({X_t}^2\right) < \infty$ then it is weakly stationary.
- Of course that, in general, a weakly stationary process is not necessarily strongly stationary

$$\text{weak stationarity} \nRightarrow \text{strong stationarity}$$

- There is one important case in which weak stationarity implies strong stationarity.

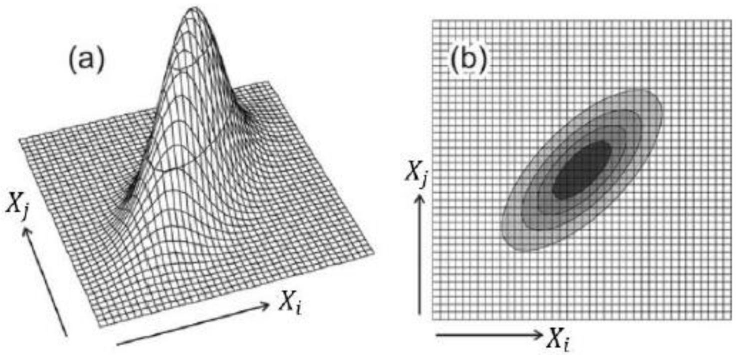
If $\{X_t, t \in \mathbb{N}\}$ is a weakly stationary Gaussian process then it is strongly stationary.

Let's see why…

Stationary models

Relation between strong and weak stationarity

Def. $\{X_t, t \in \mathbb{N}\}$ is a Gaussian process if all of its joint distributions are multivariate normal, i.e. $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ has a multivariate normal distribution for all n.



Bivariate normal distribution ($n = 2$) with correlation 0.6:
(a) Probability density function and
(b) Ellipse representation.

The normal multivariate density of $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ evaluated at $\boldsymbol{x} = (x_1, x_2, \dots, x_n)^T$ is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)$$

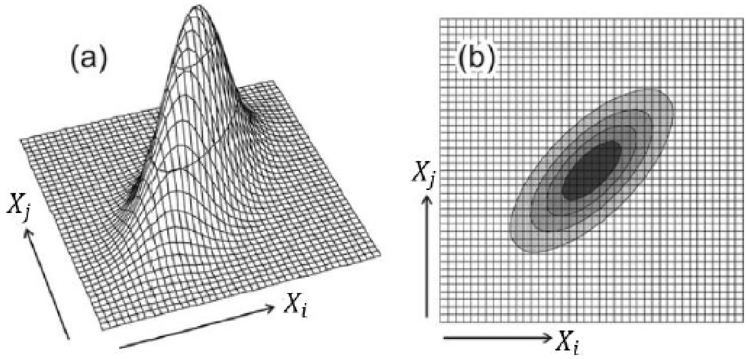
where $\boldsymbol{\mu} = \left(\mathrm{E}(X_1), \mathrm{E}(X_2), \dots, \mathrm{E}(X_n)\right)^T$ is the mean vector, Σ is the covariance matrix with $\Sigma_{ij} = \mathrm{Cov}\left(X_i, X_j\right)$ and $|.|$ represents the determinant.

The multivariate Gaussian distribution is fully characterized by its first two moments.

Stationary models

Relation between strong and weak Stationarity

If $\{X_t,\, t \in \mathbb{N}\}$ is a weakly stationary Gaussian process then it is strongly stationary.



Bivariate normal distribution ($n = 2$) with correlation 0.6:
(a) Probability density function and
(b) Ellipse representation.

Why?

If $\{X_t, t \in \mathbb{N}\}$ is a Gaussian time series, then all of its joint distributions are completely determined by the mean function $\mathrm{E}(X_t) = \mu_t$ and the autocovariance function $\gamma_X(t,s) = \mathrm{Cov}(X_t, X_s)$. If the process is weakly stationary, then $\mu_t = \mu$, for all t and $\gamma_X(t,s) = \gamma_X(h)$, for all t and $h = |s-t|$. In this case, the joint distribution of $(X_1, X_2, \dots, X_n)$ is the same as that of $(X_{1+h}, X_{2+h}, \dots, X_{n+h})$ for all integers h and $n > 0$. Hence for a Gaussian time series strict stationarity is equivalent to weak stationarity.

Stationary models

Mean and ACF under stationarity

Let $\{X_t, t \in \mathbb{N}\}$ be a stationary process. Then,

- the expected value of X_t is constant through time i.e. $\mathrm{E}(X_t) = \mu_t = \mu, \forall t$
- the autocovariance function (ACVF) of X_t depends on the lag h is is defined by
$$\mathrm{Cov}(X_t, X_{t+h}) = \gamma_X(h)$$
- the autocorrelation function (ACF) of X_t at lag h is the normalized ACVF
$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)}$$

Both ACVF and ACF are symmetric i.e.

$$\gamma_X(h) = \gamma_X(-h) \text{ and } \rho_X(h) = \rho_X(-h)$$

and are usually represented for $h \geq 0$.

If the process is not stationary, $|\rho_X(h)|$ slowly decays to zero.

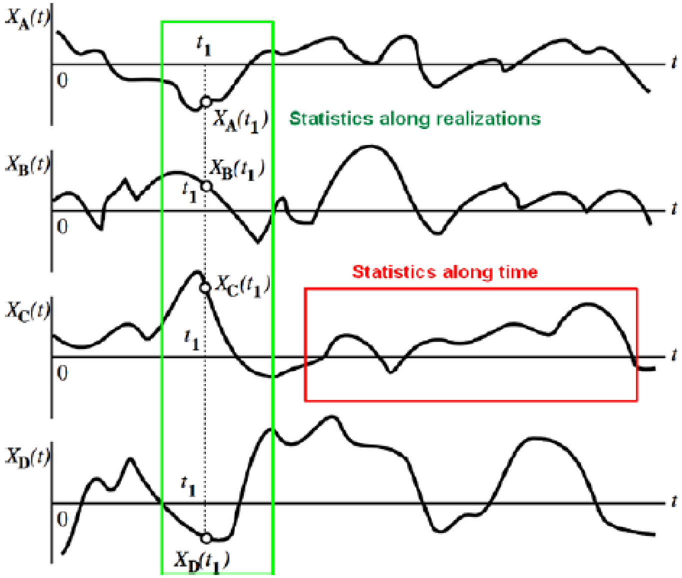
Stationary models

Properties of the sample mean and sample ACF

Although the theoretical mean and ACF are useful to describe the properties of processes, most of the analyses are performed using sampled data.

The sampled information $(X_1, X_2, \dots, X_n)$ is that available for estimating the mean, ACVF and ACF. From the point of view of classical statistics, this poses a problem because there no iid copies of X_t available for the estimation.

Usually, there is only one realization of X_t and the assumption of stationarity becomes critical. Therefore, averages over time in the realization have to be used to estimate the process/population mean (μ), ACVF ($\gamma_X(h)$) and ACF ($\rho_X(h)$).



How to estimate μ, $\gamma_X(h)$ and $\rho_X(h)$ from one realization of the stochastic process?

Stationary models

Properties of the sample mean and sample ACF

If $\{X_t,\ t \in \mathbb{N}\}$ is stationary, then $\mu_t = \mu, \forall t$ which can be estimated by the sample mean

$$\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$$

Note that $\bar{X}$ is an unbiased estimator for μ as $\mathrm{E}(\bar{X}) = \mu$. The variance of $\bar{X}$ is given by

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) = \frac{1}{n^2}\mathrm{Cov}\left(\sum_{t=1}^{n} X_t, \sum_{s=1}^{n} X_s\right) =$$

$$= \frac{1}{n^2}\big(n\,\gamma_X(0) + (n-1)\gamma_X(1) + (n-2)\gamma_X(2) + \cdots + \gamma_X(n-1) + (n-1)\gamma_X(-1)$$

$$+ (n-2)\gamma_X(-2) + \cdots + \gamma_X(-(n-1))\big) = \frac{1}{n}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma_X(h)$$

As $n \to \infty$, $\mathrm{Var}(\bar{X}) = \mathrm{E}\big((\bar{X} - \mu)^2\big) \to 0$ if $\gamma_X(n) \to 0$ which states the conditions for which $\bar{X}$ converges in mean square to μ.

Stationary models

Properties of the sample mean and sample ACF

Remarks:

- If X_t is an uncorrelated process, $\mathrm{Var}(\bar{X})$ reduces to σ_X^2/n by recalling that $\gamma_X(0) = \sigma_X^2$ and $\gamma_X(h) = 0$ for $h \neq 0$ for an uncorrelated process.
- In the case of dependence, $\mathrm{Var}(\bar{X})$ may be smaller or larger than the white noise case (iid) depending on the nature of the correlation structure.

To make inferences about μ using $\bar{X}$, it is necessary to know the distribution of $\bar{X}$ (or an approximation). If X_t is a gaussian process

$$\frac{1}{\sqrt{n}}(\bar{X} - \mu) \sim N\left(0, \sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma_X(h)\right)$$

and straightforwardly follow the exact confidence bounds for μ if γ_X is known or approximate bounds if γ_X has to be estimated from the sample.

Stationary models

Properties of the sample mean and sample ACF

The sample ACVF is defined as

$$\hat{\gamma}_X(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_{t+h} - \bar{X})(X_t - \bar{X}),$$

for $h = 0,1,\ldots,n-1$ and with $\hat{\gamma}_X(-h) = \hat{\gamma}_X(h)$.

Remarks:

- The sum runs for a restricted range of t, because X_{t+h} is not available for $t + h > n$.
- $\hat{\gamma}_X(h)$ is approximately equal to the sample covariance of the $n-h$ pairs (X_1, X_{1+h}), $(X_2, X_{2+h}), \ldots, (X_{n-h}, X_n)$. The difference comes from the divisor n in $\hat{\gamma}_X(h)$ and $\bar{X}$ being computed from n parcels (instead of $n-h$).
- The divisor n ensures that $\hat{\gamma}_X(h)$ is non-negative definite (see Shumway and Stoffer, 2016, p.27) guaranteeing that the variance of a linear combination of X_t will never be negative.

Stationary models

Properties of the sample mean and ACF

The sample ACF function is defined, analogously, as

$$\hat{\rho}_X(h) = \frac{\hat{\gamma}_X(h)}{\hat{\gamma}_X(0)}$$

Both $\hat{\gamma}_X(h)$ and $\hat{\rho}_X(h)$ are asymptotically unbiased estimators.

Remarks:

- Without further information beyond $(X_1, X_2, \ldots, X_n)$ it is impossible to give reasonable estimates of $\gamma_X(h)$ and $\rho_X(h)$ for $h \geq n$.
- Even for h slightly smaller than n, the estimates are unreliable, since there are just a few pairs (X_t, X_{t+h}) available $(h = n - 1)$.
- A useful guide is provided by Jenkins (1976), p. 33 who suggest that $n > 50$ and $h \leq n/4$.

Stationary models

Properties of the sample mean and ACF

The ACF sampling distribution allows to assess whether the data comes from an uncorrelated process or whether correlations are *statistically significant* at some lags h.

Result: Under general conditions*, for large n, the sample ACF of an iid sequence $(X_1, X_2, \dots, X_n)$ is distributed as follows

$$\hat{\rho}_X(h) \underset{\text{approx.}}{\sim} N(0, 1/n)$$

This conveys a rough method for assessing the statistical significance of peaks in sample ACF values. If $(x_1, x_2, \dots, x_n)$ is a realization of the iid process, roughly about 95% of the sample ACF values should fall within $\pm 1.96/\sqrt{n}$.

* The general conditions are that X_t is iid with finite fourth moment (see Shumway and Stoffer (2016) Appendix A). A sufficient condition for this to hold is that X_t is a Gaussian white noise (see Brockwell and Davis (1991) p. 222).

Stationary models

Properties of the sample mean and ACF

Many modeling procedures depend on reducing a time series to a white noise using several transformations. After such a procedure is applied, the plotted ACFs of the residuals should resemble that of a white noise process. As,

$$\hat{\rho}_X(h) \underset{\text{approx.}}{\sim} N(0, 1/n)$$

then approx. 95% of the ACF values must lie within $\pm 1.96/\sqrt{n}$. If not, the series is probably not white noise. Thus, ACF is provided with the $\pm 1.96/\sqrt{n}$ critical values.

Rule of thumb: When computing the sample ACF up to lag 40 and find that more than two or three values (95%) fall outside the bounds, or that one value falls far outside the bounds, the iid hypothesis should be rejected (5% significance).

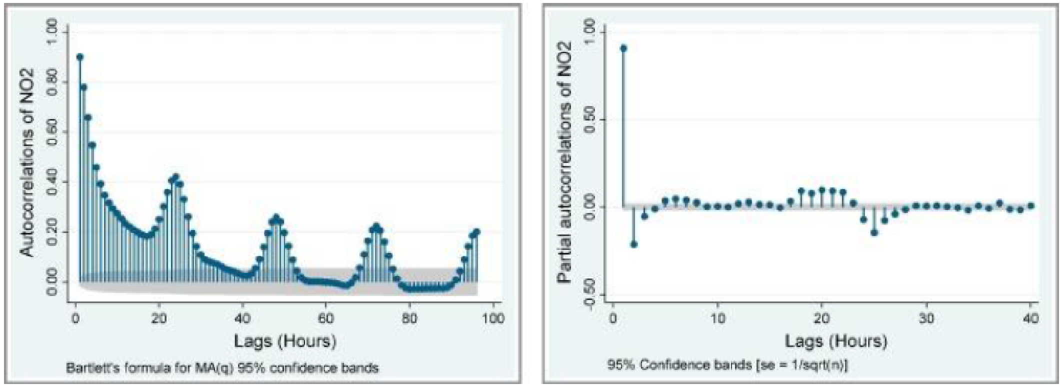
```
w = rnorm(500, mean = 0, sd = 1) # 500 random numbers generated from N(0,sd^2) distribution
stats::acf(w, lag.max = 40, main = "acf") # sample ACF (stats)
stats::pacf(w, lag.max = 40, main = "pacf") # sample PACF (stats)
astsa::acf1(w, main = "acf1") # sample ACF
astsa::acf2(w, main = "acf2") # sample ACF
forecast::Acf(w)
forecast::Pacf(w)
```

Stationary models

Partial autocorrelation of a time series X_t, $t = 1,2, \ldots n$ gives the correlation of X_t on its own past and future values, whilst controlling for the values of X_t at all shorter lags.

$$\text{PACF}(h) = \frac{\text{Cov}(X_t, X_{t+h} | X_{t+1}, X_{t+2}, \ldots, X_{t+h-1})}{\text{Var}(X_t | X_{t+1}, X_{t+2}, \ldots, X_{t+h-1})}$$

Sample ACF and PACF for a time series X_t, $t = 1,2, \ldots n$ (example)



Statistical inference on ACF or PACF

The grey shadows represent the critical regions defined by

$$\pm 1.96/\sqrt{n}$$

for which the null hypothesis in

H_0: ACF$(h) = 0$ vs H_1: ACF$(h) \neq 0$

cannot be rejected at a 5% level. The same region is defined for PACF(h).

Figures from M Catalano and F Galatioto, "Enhanced transport-related air pollution prediction through a novel metamodel approach", Transportation Research 2017, 55, 262-276, doi: 10.1016/j.trd.2017.07.009