Notes on Time Series

Grégoire Clarté

2025-2026

Contents

Stemmatologic thanks: These notes originates from the notes by:

- Ioannis Papasthopoulos (UoE)
- Djalil Chafaï, Célive Lévy-Leduc, Angelina Roche (Dauphine)
- V. Monbet (ENSAI)

Further results come from the following books:

- The Analysis of Time Series - an Introduction, Chatfield (2004, Chapman and Hall).
- Time series: Theory and Methods, Brockwell and Davis (2009, Springer).
- Time Series Analysis and Its Applications (with R examples) - Shumway and Stoffer, Springer.

Part I
Models for time series

1 Time series basics

1.1 Introduction

In basic data analysis, the observations are usually independent. Observations $X_1, \dots, X_n$ are assumed to have joint density of the form $f_1 \otimes \cdots \otimes f_n$, that is independent. Often, we even assume $f_1 = \cdots = f_n$, that is iid observations.

In this course, we will deal with a common form of dependency of the observations, dependency on time:

Definition 1 (Time Series)**.** A time series is a *finite* sequence of observations indexed by time.

For example $(X_t)_{t\in\{t_1,\dots,t_n\}}$, with $X_t \in \mathbb{R}^k$.

Example 2. *Daily rain observations, stock values, physiological measurements, are all typical time series.*

Remark 3. *We should make a difference between Time series as an observation (which is always a finite sequence) and a Time series as a mathematical model that we will call stochastic process, defined below, which are infinite sequences.*

Studying time series can have several aims:

- **describe** the time series;
- **analyse** by building statistical models, for example to take into account the dependence of the observations;
- **monitor/control** a process, for example a mechanical or biological process, to raise an alarm;
- **predict** future observations: meteorology, physiological processes, etc.

You can observe some examples of Time Series in Figure 1.

1.2 Basic notations and quantities

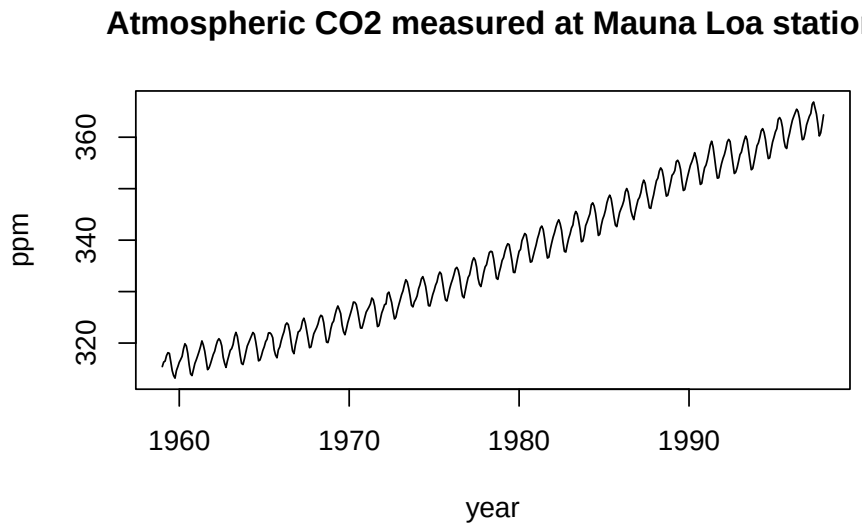
Definition 4.
1. A *stochastic process* $(X_t)_{t\in\mathcal{T}}$ indexed by $\mathcal{T}$ is a set of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.
2. A *realisation* of (X_t) is the outcome $(x_t)_{t\in\mathcal{T}} = (X_t(\omega))_{t\in\mathcal{T}}$ for some ω.

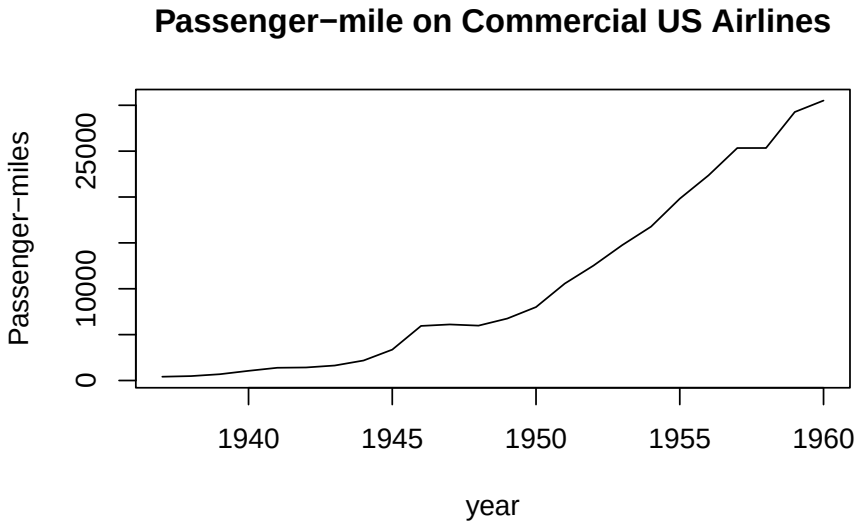
Remark 5. *For the sake of lightness of the writing I may write X for the time series, equivalently with (X_t) or with $(X_t)_{t\in\mathcal{T}}$. X_t designate a single point of the time series. This is not the same object, one is a random variable with value in $\mathbb{R}^{\mathcal{T}}$, the other with value in $\mathbb{R}$ (for example).*

Remark 6. *In general $\mathcal{T} = \mathbb{R}, \mathbb{R}^+$ or $\mathbb{Z}$.*

It is not possible to numerically simulate values indexed by real numbers, but in practice if the points are close enough (the time step is small enough) simulating from a discrete model leads to similar results.

In these notes, we will assume $\mathcal{T} = \mathbb{Z}$, so that we write $(X_t)_{t\in\mathcal{T}} = (\dots, X_{-n}, \dots, X_{-1}, X_0, X_1, \dots)$.





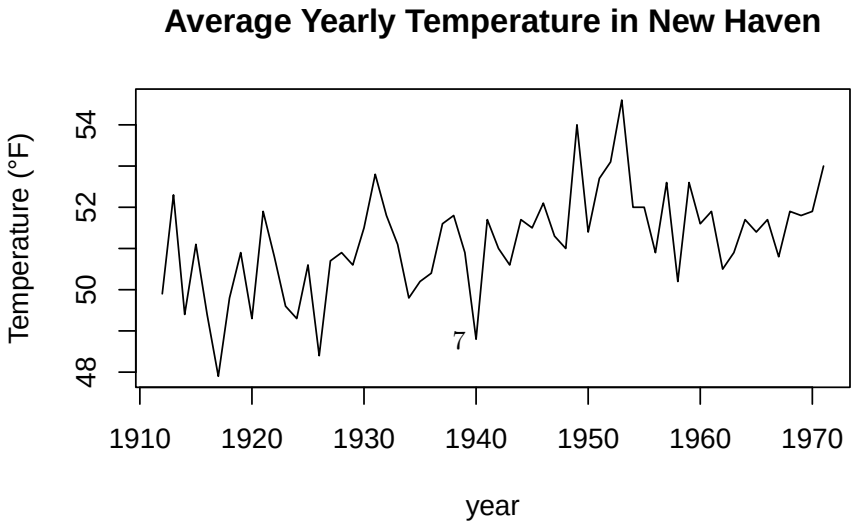


Figure 1: Example of time series

Remark 7. *If the observations come from a random function $X : \mathbb{R} \to \mathbb{R}$, it's possible to define the observations as points of the function: $(X(t_i))_{i\in\mathbb{Z}}$, but also to define it as integrals of the functions $\int_{t_i}^{t_{i+1}} X(t)dt$, for example when studying rainfalls.*

Care must be taken when working on these series as for how they were produced.

1.2.1 Mean, covariances, correlations

Definition 8. A stochastic process (X_t) with values in $\mathbb{R}$ is said to be a *second order stochastic process* (or *of the second order*) if $\forall t,\ E(X_t^2) < \infty$.

Most of the processes we will study are of the second order. This definition basically is the counterpart in stochastic processes of the L^2 space for functions. Not surprisingly you will find in the following results that reminds of Hilbert spaces theory.

Definition 9. Let $(X_t)_{t\in\mathcal{T}}$, be a second order stochastic process. Then,

1. We define the *mean* (or *expected value*) of the process as:

$$\mu_t = E(X_t).$$

   If the mean depends on time, we call it *trend*.

2. We define the *(auto)covariance* function as:

$$\gamma(s,t) = cov(X_s, X_t) = E((X_s - \mu_s)(X_t - \mu_t)),$$

   we can also define the autocorrelation for $s, t \in \mathcal{T}$ by:

$$\rho(s,t) = \frac{\gamma(s,t)}{(\gamma(s,s)\gamma(t,t))^{1/2}}.$$

Remark 10. *$Var(X_t) = cov(X_t, X_t) = \gamma(t,t)$.*

Cauchy-Schwaz gives $\forall s, t \in \mathcal{T}, |\rho(s,t)| \leq 1$, with $\forall t \in \mathcal{T}, \rho(t,t) = 1$.

The function γ is symmetric positive semi-definite: $\sum a_i a_j \gamma(t_i, t_j) \geq 0$, for any $a_1, \ldots, a_k \in \mathbb{R}$ and any $t_1, \ldots, t_k \in \mathcal{T}$.

Exercise 11. *Let $t_1, \ldots, t_n \in \mathbb{Z}$, show that $\Gamma = (\gamma(t_i, t_j))_{i,j}$ is symmetric positive semi-definite.*

1.2.2 Backshift operator

A concept that is useful to study time series is the *Backshift operator* B:

Definition 12 (Backshift operator)**.** The Backshift operator is an operator on the spaces of stochastic processes that shifts the process by one step towards the past:

$$\forall t \in \mathbb{Z}, (BX)_t = X_{t-1}.$$

Remark 13. *This is an operator on the process. In particular $B(X_t)$ does not have any sense. BX_t must be read as $(BX)_t$.*

1.3 Stationarity

If we can decompose a stochastic process (X_t) into

$$X_t = d_t + w_t,$$

where d_t is a deterministic function and w_t is a random noise, a good property to study the process would be that the w_t have same distribution. In this notation d corresponds to the trend, we will discuss this part in a later section. This leads to the following definition:

Definition 14 (Strict stationarity)**.** A process $(X_t)_{t\in\mathbb{Z}}$ is strictly stationary when for all $h \in \mathbb{Z}$, and all finite sequence $t_1, \dots, t_n \in \mathbb{Z}$, $n \geq 1$, the vectors $(X_{t_1}, \dots, X_{t_n})$ and $(X_{t_1+h}, \dots, X_{t_n+h})$ have the same distribution.

Example 15. *If $\forall t, X_t \sim \mathcal{F}$, independently, then (X_t) is strictly stationary.*

This definition is quite stringent, and often a weaker concept of stationarity is required. More often, we call simply stationarity this *weak* stationarity.

Definition 16 ((Weak) stationarity)**.** We say that a second order process (X_t) is (weakly) stationary when for all $h \in \mathbb{Z}$, and all finite sequence $t_1, \dots, t_n \in \mathbb{Z}$, $n \geq 1$, the random vectors

$$(X_{t_1}, \dots, X_{t_n}) \text{ and } (X_{t_{1+h}}, \dots, X_{t_n+h}),$$

have same *mean* and *covariance matrix*.
In other words, this means that μ_t is constant and that γ is translation invariant.

Remark 17. *This concept is often called* second-order stationarity.

Exercise 18. *Show that if (X_t) is stationary, then $\gamma(s,t) = \gamma(0, t-s)$ for all $s, t \in \mathbb{Z}$.*

This means that the covariance function γ is actually a function of a single value (the difference in time between the points). We call this argument the *lag*:

$$\gamma(s, s+h) = \gamma(h).$$

Sometimes, we write γ_h instead of $\gamma(h)$.

Exercise 19. *Show that if (X_t) is stationary, then $\sigma_t = Var(X_t)$ is also constant.*

The terms *strict* and *weak* are well chosen:

Theorem 20. *Let (X_t) be a second order process. Then if (X_t) is strictly stationary it is stationary. The reciprocal is false.*

Remark 21. *In practice, it is not possible to check strict stationarity. In addition most result require only second order stationarity.*

We could define order-k stationarity by requiring all moments up to order k to be constant.

In practice, we usually preprocess the data to remove seasonality and trend and model the remaining series with a stationary model. See later sections on this matter.

Exercise 22. *Let $X_t = X_{t-1} + w_t$, with $X_0 = 0$ and w_t iid $\mathcal{N}(0,1)$. Show that this process is not stationary*

This exercise actually contains a small trap. This process cannot be stationary because X_0 is constant (mean and variance are null), while X_1 follows a normal distribution with non null variance.

It is otherwise possible to compute the autocovariance of the process:

$$X_t = w_1 + \cdots + w_t, \text{ and } X_{t+h} - X_t = w_{t+1} + \cdots + w_{t+h}$$

are independent and centered, thus $Cov(X_t, X_{t+h}) = 0$, hence by bilinearity:

$$\gamma(t, t+h) = Cov(X_t, X_{t+h} - X_t + X_t) = Cov(X_t, X_{t+h} - X_t) + Var(X_t) = Var(X_t) = t\sigma^2.$$

Note that it is not possible to define this process without a starting point (like AR models). Note as well that this previous result is linked with what you will remember from your Markov chain course. If you don't remember this course check your past notes.

Stationary processes have nice properties for the time average that will be used in statistical inference:

Property 23 (Variance of a time average)**.** *Let (X_t) be a stationary second order stochastic process, and let $\bar{X} = n^{-1}\sum_{i=1}^{n} X_{t_i}$, for some $n \geq 1$, $t_1, \ldots, t_n$. We have that:*

$$Var(\bar{X}) = \frac{\gamma(0)}{n}\left(1 + \frac{2}{n}\sum_{h=1}^{n-1}(n-h)\rho_h\right).$$

Property 24. *Let $\gamma : \mathbb{Z} \to \mathbb{R}$ the autocovariance function of a stationary process. Then it verifies: $|\gamma(h)| \leq \gamma(0)$.*

1.3.1 Estimating mean and covariance of a stochastic process

We observed $x_1, \ldots, x_n$ from a Time Series. We assume it comes from a stochastic process $(X_t)_t$ at times $t = 1, \ldots, n$. We assume X to be stationary (at least weakly), so that $E(X_t) = \mu$ and γ_X exists.

We propose the following method:

1. We estimate μ with the sample mean: $\hat{\mu} = \bar{x} = n^{-1}\sum_{i=1}^{n} x_i$;
2. $\gamma_X(k)$ is estimated as:

$$c_k = \frac{1}{n-k-1}\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x});$$

3. the autocovariance is then estimated as $r_k = \frac{c_k}{c_0}$.

Remark 25. *Sometimes the denominator of c_k is $n-k$, it's a question of bias.*

Remark 26. *The formulas for c_k and r_k should be used for k small compared to n, e.g. $k \leq n/3$.*

Definition 27. A graph of r_k as a function of k is called a correlogram.

1.4 White noise, gaussian process

Definition 28 (Gaussian Process)**.** A stochastic process (X_t) is a *Gaussian Process* when $(X_1, \ldots, X_t)$ is a Gaussian vector for any $t_1, \ldots, t_n \in Z$, $n \geq 1$.

Property 29. *A Gaussian Process is always of the second order.*

Theorem 30. *If (X_t) is a stationary Gaussian Process, it is strictly stationary.*

Definition 31 (White noise)**.** A stochastic process (X_t) is called a *white noise* if:

$$\forall t, s \in \mathbb{Z},\ E(X_t) = 0, \gamma(t, s) = \mathbf{1}_{s=t}\sigma$$

For the sake of simplicity, we often write $X \sim WN(0, \sigma^2)$ or even $\sim (0, \sigma^2)$ to designate a white noise.

The term "white" comes from the analogy with white light, where all frequencies are equally represented. This will be discussed later in this course.

Often we will write $(0, \sigma^2)$ the characteristics of a white noise.

1.5 Testing White-noise-ness

If $X_t = w_t + \mu$, where $w_t \sim (0, \sigma^2)$, then the estimator of the autocorrelation r_k approximately follows the distribution $\mathcal{N}(0, 1/n)$.

With this property we can design a test for White-noise-ness:

- $H_0 : X$ is a white noise with constant mean μ and variance σ^2.
- under H_0, when n is large:

  $$r_k \sim \mathcal{N}(0, 1/n), \quad \forall k \in \{1, \ldots, m\}$$

  where m is usually nor more than $n/3$. Note that under H_0 the $(r_k)_k$ are independent.
- we count the number B of k for which r_k falls outside of an α confidence interval. We know that $B \sim \mathcal{B}(m, \alpha)$, and we can reject the assumption if B is too large.

Example 32 (Lake Huron)**.** *We use the* `LakeHuron` *dataset available on* `R`*.*

We produce the two graphs of Figure 2. The correlogram already includes the significance lines. We can safely claim that the time series observed is not a realisation of a white noise (9 of the values lie outside of the boundaries while we would expect $\sim \mathcal{B}(20, 0.05)$).

1.5.1 Standard portmanteau test

This test is due to Box & Pierce (1970, JASA), and Ljung & Box (1978, Bka). It is based on the fact that for a white noise sequence with $m \ll n$ and n large

$$Q_m = n(n+2)\sum_{h=1}^{m}(n-h)^{-1}r_h^2 \sim \chi^2_m.$$

The sensitivity of the test depends on m:

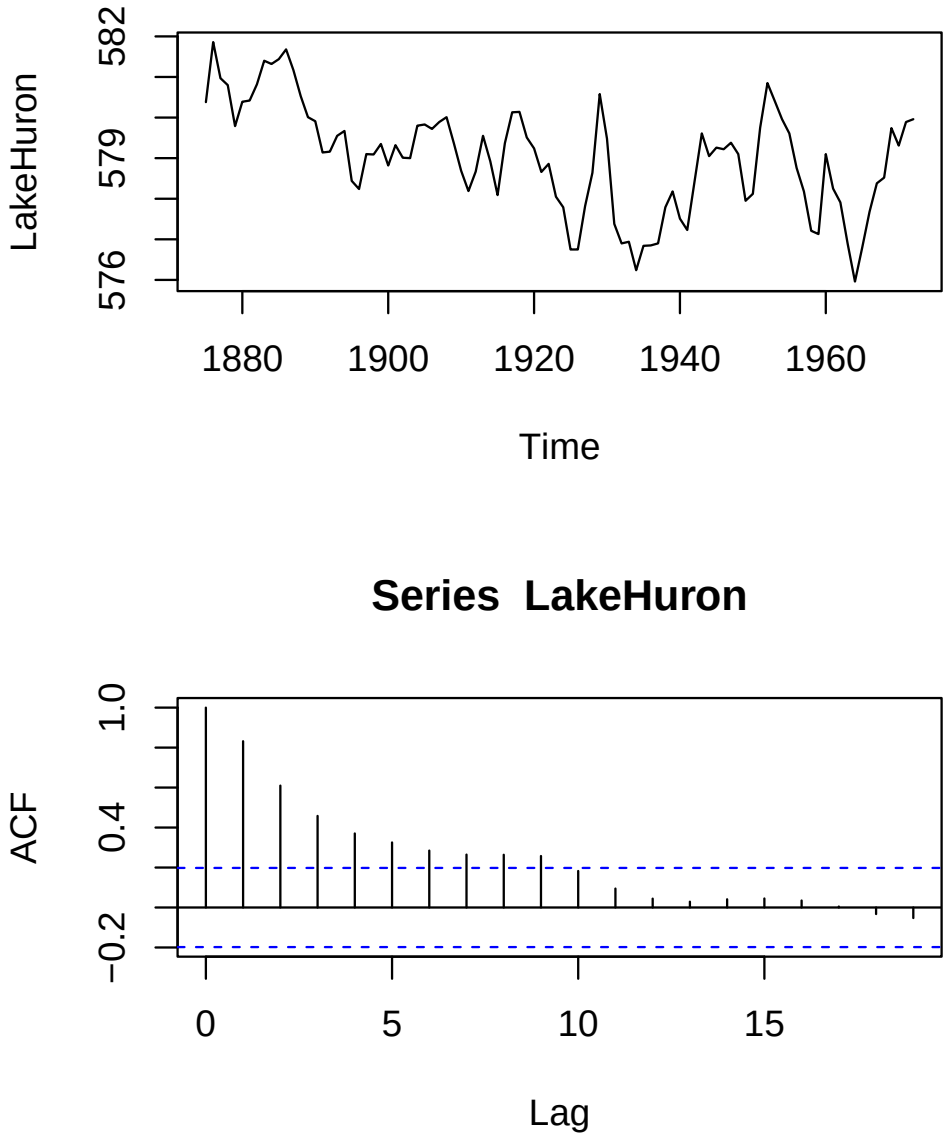


Figure 2: Lake Huron Dataset and ACF associated

– if m is too large, the sensitivity is greatly reduced because some of the r_h will not contribute to inform the lack of fit.

– m too small will reduce the sensitivity because some of the contributing r_h are not included.

Typically, we should plot the significance levels of Q_m and see if it is too small for a range of values of m.

Example 33 (Ljung-Box test)**.** *On the lake Huron data we have:*

k	1	2	3	4	5
r_k	0.84	0.63	0.48	0.39	0.35

The test statistic can then be computed as:

$$Q_5 = 164.34$$

while $\chi^2_5(0.95) = 11.07$*. Which means we strongly reject* H_0 *for this data.*

1.6 Testing for stationarity

We introduce here a test for stationarity based on the model:

$$X_t = \xi t + \eta_t + \epsilon_t, \quad \eta_t = \eta_{t-1} + w_t, \quad w \sim (0, \sigma_w^2).$$

The test is only valid if the previous model is sensible for the data.

Note that in the previous model:

– if $\sigma_w = 0$ and $\xi = 0$ then $(X_t)_t$ is stationary;

– if $\sigma_w = 0$ then $(X_t - \xi t)$ is stationary;

In econometrics, the first case is called level stationarity, and the second trend stationarity.

We can then propose a test based on these properties. The KPSS test (Kwiatkowski *et al.*, Journal of Econometrics, 1992) implemented in `R` as `kpss.test` is to build a score test for the hypothesis $\sigma_w = 0$, leading to the test statistic:

$$C(l) = \hat{\sigma}(l)^{-2} \sum_{t=1}^{n} S_t^2, \qquad \text{where } S_t = \sum_{j=1}^{t} e_j, \quad t = 1, \dots, n;$$

and where $e_1, \dots, e_n$ are the residuals from the linear regression $X_t = \alpha + \beta t + \epsilon_j$ and $\hat{\sigma}(l)$ is based on residuals truncated at lag l. Under conditions on the dependence of (ϵ_t), we can show that $C(l)$ has a tractable (and complex) distribution (integral of the squared brownian bridge).

1.7 An example: Autoregressive model

Let (w_t) be a gaussian white noise, and define

$$X_t = \alpha X_{t-1} + w_t, \; X_0 \sim \mathcal{N}(0, \sigma_0)$$

We will prove the existence of this process later for some α.

This model is called Autoregressive of order 1 (AR(1)), as it relies on the previous observations to build the next.

Let's find the parameters for which this process is stationary. A necessary condition is the constance of the mean:

$$E(X_t) = \alpha E(X_{t-1}) = \alpha E(X_0) = 0.$$

Another necessary condition is the constance of the variance:

$$Var(X_t) = \alpha^2 Var(X_{t-1}) + 1,$$

thus,

$$\sigma_0 = \frac{1}{1-\alpha^2}.$$

Reciprocally, if the process verifies these two assumptions, it is strictly stationary (normal with constant mean and variance).

For $\alpha = 1$, note that the model is not stationary.

The variance of the averages of the AR(1) process are thus, following Prop. 23 for large n:

$$Var(\bar{X}) \simeq \frac{\gamma(0)}{n}\left(\frac{1+\alpha}{1-\alpha}\right) = \frac{\gamma(0)}{n(1-\alpha)^2}$$

1.8 An example: Moving Average process

Definition 34 (Moving Average of rank 1 Process)**.** A process X is said to be a moving average process (MA(1)) if:

$$\forall t \in \mathbb{Z}, \qquad X_t = w_t + \theta w_{t-1}$$

where w is a white noise with variance σ^2 and mean 0.

Property 35. *We have:* $\mu(t) = E(w_t) + \theta E(w_{t-1}) = 0$.

$$\gamma(s,t) = E(X_s X_t) = \begin{cases} 0 & \text{if } |s-t| > 1; \\ \theta\sigma^2 & \text{if } |s-t| = 1; \\ (1+\theta^2)\sigma^2 & \text{if } s = t. \end{cases}$$

The process is thus stationary.

Remark 36. *More generally, we can define higher order MA processes as:*

$$X_t = \theta_0 w_t + \cdots + \theta_q w_{t-q}.$$

Property 37. *Let X be a MA(q) process:*

$$\forall t \in \mathbb{Z}, \quad X_t = \theta_0 w_t + \cdots + \theta_q w_{t-q};$$

it is a stationary process and its covariance verifies $\gamma_X(h) = 0, \forall h \geq q+1$.

1.9 Differenciation

As we saw in some of the examples, sometimes the time series exhibit a clear trend that prevents us from using stationarity. A simple way to remove this trend is to study the *first difference of the time series*:

Definition 38. Difference operator We define Δ the difference operator on a time series (X_t) as:

$$\Delta X = (I - B)X.$$

That is $(\Delta X)_t = X_t - X_{t-1}$.
We can define by recurrence Δ^k by $\Delta^k X = \Delta(\Delta^{k-1}X)$.

Property 39. *If X is stationary, ΔX is stationary.*

Property 40. *If $X_t = p(t) + w_t$ where p is a kth degree polynomial, and w_t is a stationary stochastic process, then $\Delta^k X$ is stationary.*

Remark 41. *If X_t is a random walk, then ΔX_t is stationary.*

In practice, we can use this differentiation process until the resulting series seems stationary enough; often $k = 1, 2$ is enough.

Remark 42. *In some cases, longer lag difference might be interesting to study. For example if a series has a period d, then $(I - B^d)X$ should be more stationary.*

These last comment are quite hand-wavy, they are practical recommendations.

Remark 43. *Differentiating usually complicates the dependency structure.*

1.10 Remindals: L^p spaces for random variables

We can define the space of random varibles with pth order moment, associated with the norm:

$$\|X\| = \sqrt[p]{E(|X|^p)}.$$

The space of all real random variables that have finite second order moment, written L^2 is a Hilbert space for the following dot product (scalar product):

$$\langle X, Y \rangle = E[XY].$$

In this space we can do more: orthogonal projections, etc.
Convergence in L^p and a.s. are not equivalent.

2 Linear filtering, causality

In the previous part, we mentioned the AR(1) process as an example of stationary process under some assumptions, leaving aside its existence. Considering the equation that defines AR(1):

$$\forall t \in \mathbb{Z}, X_t = w_t + \alpha X_t,$$

with w a white noise and α the parameter of the process. To solve this equation in X, we can iterate n times the definition:

$$X_t = w_t + \alpha(w_{t-1} + \alpha(\dots))) = \sum_{k=0}^{n} \alpha^k w_{t-k} + \alpha^{n+1} X_{t-(n+1)}.$$

Now, if $|\alpha| < 1$ and if a stationary solution of the equation existed, then:

$$\|\alpha^{n+1} X_{t-(n+1)}\|_2^2 = E[(\alpha^{n+1} X_{t-(n+1)})^2] = |\alpha|^{2(n+1)}(\gamma(0) + \mu_X^2) \longrightarrow_{n\to\infty} 0,$$

that is this series converge in L^2 to the stationary solution of the AR(1) equation.

Studying when and how these stochastic series converge is the goal of this part. This will allow us to study more precisely the ARMA processes.

If X is a process and if $(\alpha_k)_{k\in\mathbb{Z}}$ is a determinist sequence of real numbers with finite support (there is a finite number of non null α_k), then the random variable:

$$Y_t = \sum_{k\in\mathbb{Z}} \alpha_k X_{t-k}$$

is defined (as the sum is finite). This moving weighted average constitutes a *convolution smoothing* called *linear filtering*, equivalent to the integrals $f * g(t) = \int f(t-h)g(h)dh$.

If α doesn't have finite support we can rely on the following theorem to make sense of the infinite sum:

Definition 44 (Summability)**.** We say that $(x_k)_{k\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$ is *summable* if and only if $\sup_{n,m} \sum_{k=-m}^{n} |x_k| < \infty$. In this case, the series $\sum_{k\geq 0} x_k$ and $\sum_{k\leq 0} x_k$ are absolutely convergent, which allows to define $\sum_{k\in\mathbb{Z}} x_k$.

We write $\ell^1(\mathbb{Z})$ the space of those series and $x \mapsto \|x\|_1 = \sum_{k\in\mathbb{Z}} |x_k|$ defines a norm on that space, making it a Banach space.

Theorem 45 (Filtering Theorem)**.** *Let $\alpha \in \ell^1(\mathbb{Z})$ and let X be a process such that $\sup_{t\in\mathbb{Z}} E(|X_t|^p) < \infty$. If we define:*

$$\forall t \in \mathbb{Z}, \forall m, n \in \mathbb{N}, \quad Y_{t,m,n} = \sum_{k=-m}^{n} \alpha_k X_{t-k}.$$

Then for all $t \in \mathbb{Z}$, the family $(Y_{t,m,n})_{m,n}$ converges almost surely and in L^p when $m, n \to \infty$ to a random variable $Y_t \in L^p$:

$$Y_{t,m,n} \xrightarrow[m,n\to\infty]{a..s.} Y_t \in L^p \text{ and } \lim_{m,n\to\infty} = E(|Y_{t,m,n} - Y_t|^p) = 0.$$

In addition, the process (Y_t) is well defined a.s. and bounded in L^p.

This theorem allows us to state the following result for stationary processes:

Theorem 46 (Filtering of stationary processes)**.** *Let $\alpha \in \ell^1(\mathbb{Z})$ and X a stationary process with mean μ_X and covariance function γ. Then the process $Y := F_\alpha(X)$ defined by:*

$$\forall t \in \mathbb{Z}, \quad (F_\alpha X)_t = \sum_{k\in\mathbb{Z}} \alpha_k X_{t-k},$$

is a second order process stationary with mean and covariance:

$$\mu_Y = \mu_X \sum_{k\in\mathbb{Z}} \alpha_k \quad \text{and} \quad \gamma_Y(h) = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \alpha_j \alpha_k \gamma(h+j-k).$$

Remark 47. *Note that the parity of the covariance function comes from the parity of γ.*

Remark 48. *We can interpret $F_\alpha X$ using operator polynomials (or series).*
If $P_\alpha = \sum_{k\in\mathbb{Z}} \alpha_{-k} X^k$, then:
$(F_\alpha X)_t = (P_\alpha(B))(X)_t$.

Remark 49. *Note that if the polynomials are defined as finite support sequence of numbers we have yet another link appearing.*

Example 50 (Filtering of a white noise)**.** *Let w be a white noise with variance σ^2, $\mu \in \mathbb{R}$ and $\alpha \in \ell^1(\mathbb{Z})$, then Th. 46 states that $X = \mu + F_\alpha w$ is a stationary process with mean μ and autocovariance*

$$\gamma_X(h) = \sigma^2 \sum_{j\in\mathbb{Z}} \alpha_j \alpha_{j+h}.$$

Remark 51. *The Backshift operator previously introduced can be used to write a linear filtering.*
Let $\alpha \in \ell^1(\mathbb{Z})$, then we can write $F_\alpha = P_\alpha(B)$:

$$(P_\alpha(B)X)_t = \sum_{k\in\mathbb{Z}} \alpha_k (B^k X)_t = \sum_{k\in\mathbb{Z}} \alpha_k X_{t-k}.$$

2.1 Causality and invertibility

The previous definitions and results lead us to introduce these two concepts tightly linked with the definition of ARMA processes.

Definition 52. Let Z be a stationary process. We say that the filter $X = \mu + F_\alpha Z$ of Z is a

- *causal process* if $\alpha_k = 0$ for all $k < 0$ (that is the process only depends on the past).
  An example of such a process is the MA(q) processes that verify:

$$\forall t \in \mathbb{Z}, \quad X_t = w_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}.$$

- *invertible process* if Z is causal process of X, that is there exists $\beta \in \ell^1(\mathbb{Z})$ such that $Z = F_\beta(X)$, with $\beta_k = 0$ for all $k < 0$.
  An example of such a process is the AR(p) process that verify $Z = F(X)$ as:

$$\forall t \in \mathbb{Z}, X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = w_t.$$

In other words, invertibility means that there exists a linear filter that allows to recover the underlying process.

Remark 53. *The invertibility should be understood in the algebraic sense in the Algebra defined by the composition of linear filters. We leave these theoretical results to the side but they are available in several books.*

The definition of invertibility can be tricky to check. To circonvene this issue the following theorem can be used in practice:

Theorem 54 (Invertibility of stochastic models)**.** *Let $X = \mu + F_\alpha w$, a stochastic process with α* ***with finite support****.*
Then X is invertible iif $P_\alpha = \sum \alpha_k x^k$ as a polynomial only has (possible complex) roots ***outside of the unit disc****.*

3 AR and MA models

3.1 Moving Average process (MA)

Definition 55 (MA process)**.** A process $(X_t)_{t\in\mathbb{Z}}$ is a MA(q) process for an integer $q \geq 0$ if it is a second order process and solution of the recurrence equation:

$$X_t = w_t + \sum_{k=1}^{q} \theta_k Z_{t-k} = (F_\theta w)_t = (\Theta(B)w)_t$$

where $(w_t)_{t\in\mathbb{Z}}$ is a white noise with mean 0 and variance σ_2, and where $\theta \in \mathbb{R}^q$ is a fixed vector. We say that q is the order of the process and that (θ, σ^2) are its parameters.

Remark 56. *If $(X_t)_{t\in\mathbb{Z}}$ is MA(q) then it is MA(r) for all $r \geq q$.*

Property 57. *Let $(X_t)_{t\in\mathbb{Z}}$ be a MA(q) process. Then, defining $\theta_0 = 1$,*

- $E[X_t] = E[w_t] + \sum_{k=1}^{q} \theta_k E[w_{t-k}] = 0$.
- $\gamma(k) = cov(X_t, X_{t+k}) = E[X_t X_{t+k}] - 0 = \sum_{r=0}^{q}\sum_{s=0}^{q} \theta_r\theta_s E[w_{t-r}w_{t+k-s}]$
  $= \begin{cases} 0 & \text{if } |k| > q \\ \sigma^2 \sum_{r=0}^{q-|k|} \theta_r\theta_{r+|k|}; & \text{if } |k| \leq q. \end{cases}$

To prove the weak stationarity we can use the Filtering Theorem 46. Computing everything by hand is completely doable.

The null covariance after some steps is an important property of the MA processes.

The MA process is a weighted sum of a fixed number of white noise events. It has several applications for example:

- the number of sick people in a population, as sick people need time to recover;
- the effect on production of random events, as the productivity will require time to recover.

Property 58. *For a given autocorrelation ρ there exists a single invertible MA(q) process that has this autocorrelation.*

Using the backshift operator, we can write:

$$X_t = (\Theta(B)Z)_t$$

in other words, we are looking for $\Theta(B)^{-1}$, so that we can rewrite the process as:

$$(\Theta(B)^{-1}X)_t = w_t$$

the existance of such an inverse is ensured by the following result:

Property 59. *A MA(q) process is invertible if the (potentially complex) roots of Θ all lie* outside of the unit circle.

Example 60. *The MA(1) model writes $X_t = w_t + \theta w_{t-1}$, with $\theta \neq 0$, it can be written with the backshift operator as:*

$$X_t = (1 + \theta B)w_t.$$

To prove invertibility we merely need to check that the root $-1/\theta$ is outside of the unit circle, which is true iif $|\theta| < 1$.

Example 61. *Lets consider a MA(2):*

$$X_t = (1 - B + 0,6B)w_t$$

the roots are: $\frac{1\pm i\sqrt{1,4}}{1,2}$*. We need to compute their modulus:* $\frac{\sqrt{2,4}}{1,2} > 1$*, so this process is invertible.*

3.2 Autoregressive process in more details

Definition 62 (Autoregressive Process)**.** A process $(X_t)_{t\in\mathbb{Z}}$ is AR(p) for some $p \geq 0$ if it is a second order process and solution of the recurrence equation:

$$X_t = w_t + \sum_{k=1}^{p} \phi_k X_{t-k}.$$

In other words, $w_t = (F_\phi X)_t = (\Phi(B)X)_t$
where $(w_t)_t$ is a white noise $(0, \sigma^2)$ and $\phi \in \mathbb{R}^p$ is a fixed vector. We say that p is the order of the AR process.

Remark 63. *If a process is AR of order p it is also AR of order* $r \geq p$*.*

The existence of such a process is not trivial, there is no time origin to the process. Note that it would be possible to define such a process with a time origin, but in this case it wouldn't be stationary. We rely on the following theorem to prove its existence for $p = 1$:

Theorem 64 (Existence of AR(1))**.**

- *If* $|\phi_1| = 1$ *then the AR*(1) *equation has no stationary solution.*
- *if* $|\phi_1| < 1$ *then the solution is given by the causal linear process written*

$$X_t = \sum_{k=0}^{\infty} \phi_1^k w_{t-k}$$

and has mean 0 *and autocovariance* $\gamma_X(h) = \sigma^2\phi_1^{|h|}/(1 - \phi_1^2)$ *for all* $h \in \mathbb{Z}$*. It is actually an MA*(∞) *process;*

- *if* $|\phi_1| > 1$ *then the solution is given by the linear non causal process*

$$X_t = -\sum_{k=1}^{\infty} \phi_1^{-k} w_{t+k},$$

with mean 0 *and autocovariance* $\gamma_X(h) = \sigma^2\phi_1^{-|h|}/(-1 + \phi_1^2)$*. This is once again the only stationary solution.*

Remark 65. *If we are looking for non stationary solution, we can actually find an infinite quantity of such solutions.*

This result is only for AR(1) processes. For more general AR(p) we must rely on other types of result:

Definition 66 (Causality for the AR(p) process)**.** An AR(p) process X is causal if there exists Ψ such that:

– $\Psi(B) = 1 + \psi_1 B + \dots$

– $\sum_{i=0}^{\infty} |\psi_i| < \infty$

– $X_t = \Psi(B) w_t$.

Following a similar proof of as for invertibility of the MA(q) process, we can show that:

Theorem 67 (Causality of the AR processes)**.** *An AR process X is causal iif the (possibly complex) roots of ϕ lie outside of the unit circle.*

Remark 68. *If X is causal, we can write:*

$$X_t = \Psi(B) w_t$$

which means that the process can be written as an MA(∞) process.

Example 69. *The AR(1) model $X_t = \alpha X_{t-1} + w_t$ can be written:*

$$\Phi(B) X_t = w_t,$$

with $\Phi(B) = (1 - \alpha B)$.
The process is causal if $|\alpha| < 1$, and $\Psi^{-1}(B) = 1 + \alpha B + \alpha^2 B^2 + \dots$.

3.3 Estimating the mean and autocovariance of an AR or MA process

After we built these process, the natural question in practice is to infer the parameters of the process given Time Series data.

We observed $x_1, \dots, x_n$ from a Time Series. We assume it comes from a stochastic process $(X_t)_t$ at times $t = 1, \dots, n$. We assume X to be stationary (at least weakly), so that $E(X_t) = \mu$ and γ_X exists.

We propose the following method:

1. We estimate μ with the sample mean: $\hat{\mu} = \bar{x} = n^{-1} \sum_{i=1}^{n} x_i$;
2. $\gamma_X(k)$ is estimated as:

$$c_k = \frac{1}{n-k-1} \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x});$$

3. the autocovariance is then estimated as $r_k = \frac{c_k}{c_0}$.

Remark 70. *Sometimes the denominator of c_k is $n - k$, it's a question of bias.*

Remark 71. *The formulas for c_k and r_k should be used for k small compared to n, e.g. $k \leq n/3$.*

Definition 72. A graph of r_k as a function of k is called a correlogram.

A correlogram allows to understand better the coefficient. In Figure 3 are represented the correlograms of an AR(1) and a MA(1) processes.

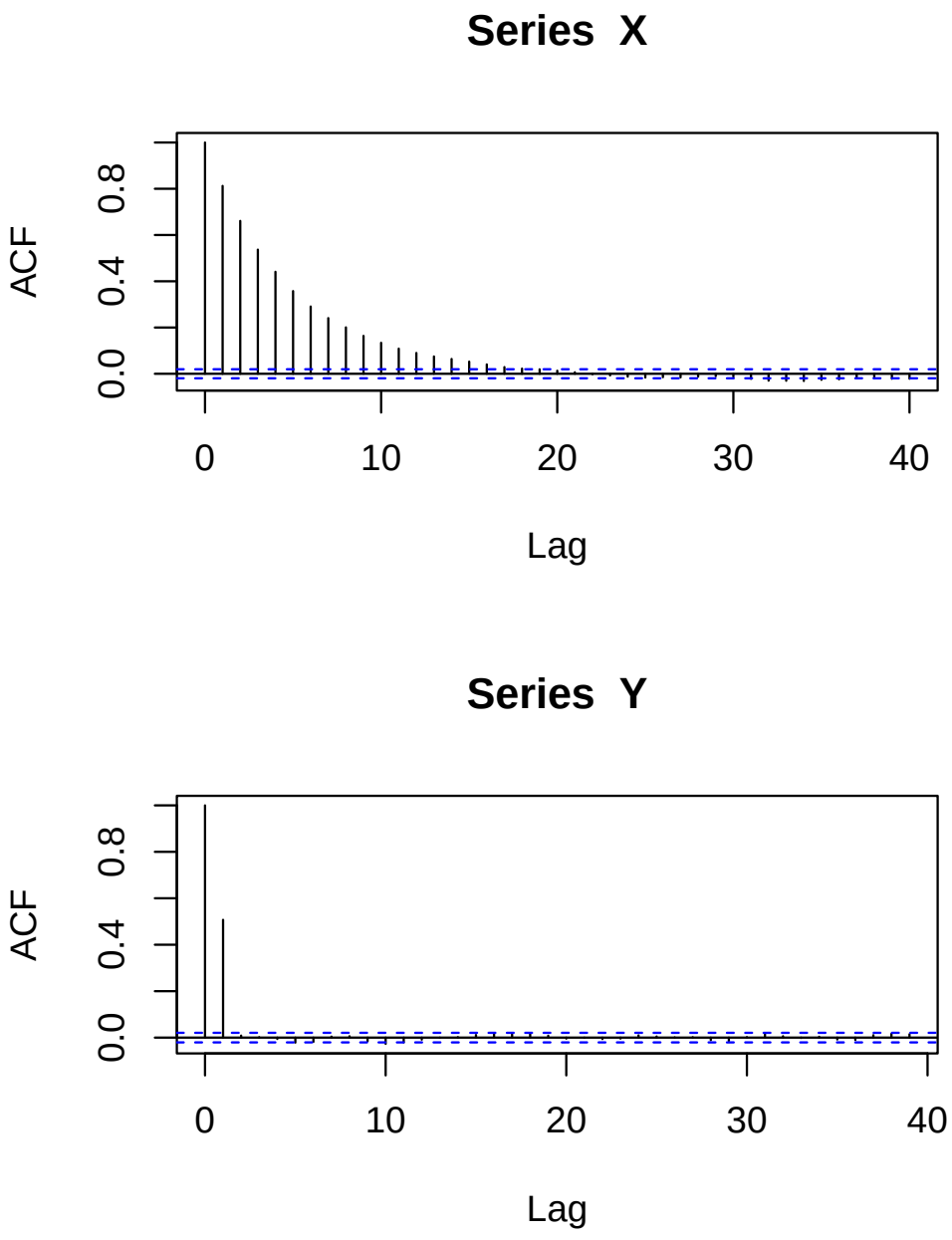


Figure 3: Correlograms of AR(1) and MA(1) processes.

3.4 Detection of MA(q) processes

As stated above, for a MA(q) process, $\rho(k) = 0, \forall |k| \geq q+1$. Hence, on a dataset following an MA model, we shoudl expect:

- $r_1, \ldots, r_q$ to be close to $\rho(1), \ldots, \rho(q)$;
- $r_{q+1}, \ldots$ to be close to 0 randomly around it.

In particular, if $|r_1|$ is large, but $r_2, \ldots$ are close to 0 it is most likely an MA(1) process.

Example 73. *In Figure 4 are two ACF for two different MA models. What order can you estimate them to be? Can you propose an estimation of the parameters?*

3.5 Detextion of AR(1) process

We saw that the autocorrelation function for the AR(1) process defined by $X_t = \alpha X_{t-1} + w_t$, with $w \sim (0, \sigma^2)$ is given by:

$$\rho(k) = \gamma(k)/\gamma(0) = \alpha^{|k|}.$$

Therefore we expect an exponential decrease with rate α. Note that contrary to MA, we do not expect the coefficients to be close to 0, but only to decrease to 0.

As r_1 is expected to be close to α, it can be used as rough estimator.

3.6 Detection of AR(p) process

It can be much more difficult to detect the order of an AR(p) process as there is no simple observation to make.

Nevertheless, for a model of the form $X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + w_t$ with $w \sim (0, \sigma^2)$ estimates of the $\alpha_1, \ldots, \alpha_p$ cab be obtained by minimising:

$$n^{-1} \sum_{t=p+1}^{n} \left(x_t - \sum_{i=1}^{p} \alpha_i x_{t_i} \right)^2$$

that is performing least squares estimation. We note $\hat{\alpha}_i$ one of these estimates.

Remark 74.
- *The estimate $\hat{\alpha}_p$ is called the* sample partial autocorrelation coefficient at lag p.
- *To obtain the sample partial autocorrelation coefficient at lag e.g. k, we need to fit an AR(k) to the data.*
- *The sample partial autocorrelation coefficient at lag p measures the autocorrelation at lag p that is not accounted for by the autocorrelation at lags $1, \ldots, p-1$.*
- *We can plot the sample partial autocorrelation coefficients as a function of their lag, similarly to the correlograms. This gives the* partial autocorrelation function.
- *For an AR(p) process the partial autocorrelation coefficients $\hat{\alpha}_{p+1}, \ldots$ should drop to around 0, as there shouldn't be anything left to explain in the dependency. This is the easiest way to estimate the order of an AR process.*

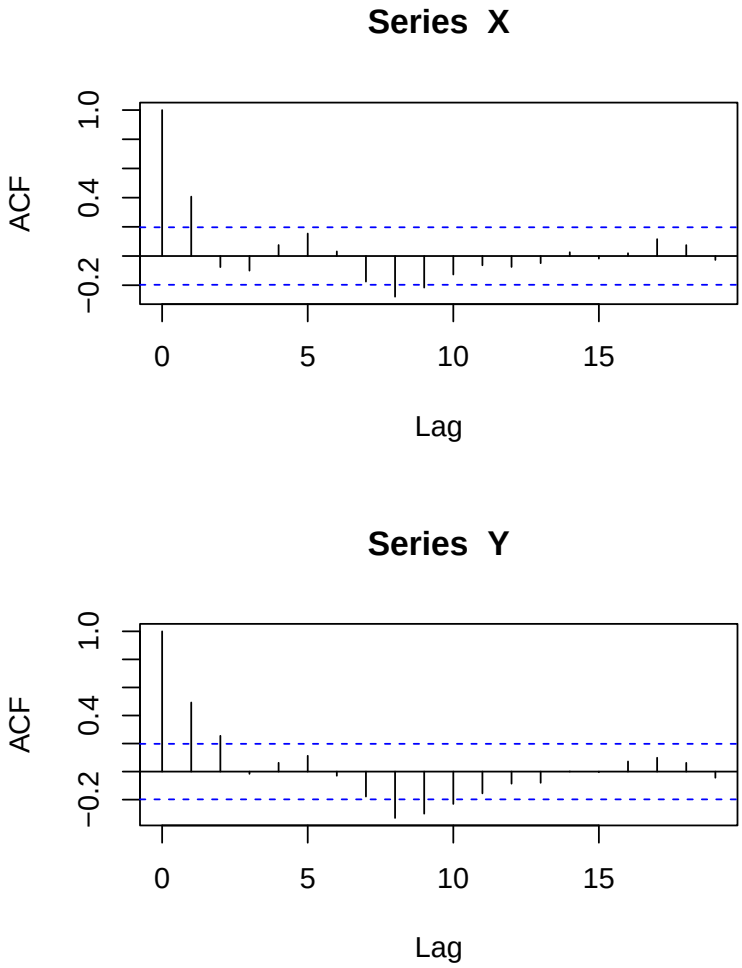


Figure 4: Example of detection of MA parameters

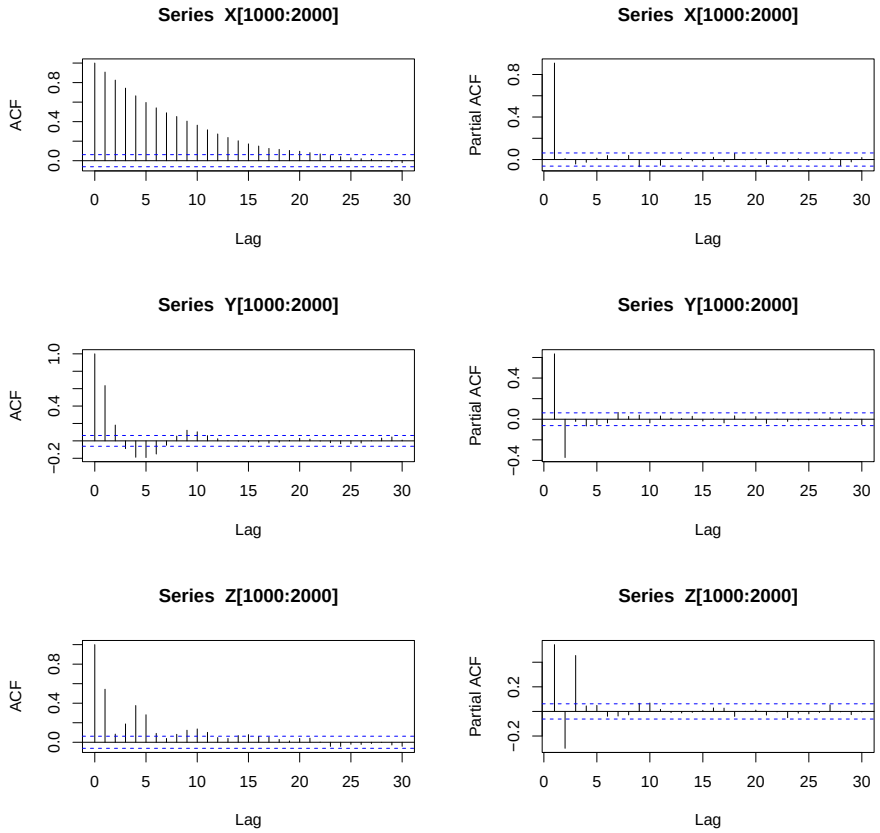


Figure 5: ACF and PACF for AR(1), AR(2) and AR(3)

3.6.1 Approximate distribution of $\hat{\alpha}_k$

The partial autocorrelation coefficients at lag k ($\hat{\alpha}_k$) is significantly different from 0, at 5% confidence it should be outside of $(-2/\sqrt{n}, 2/\sqrt{n})$.

This gives a slightly handwaivy test for the order. Note that in `R`, the blue lines represent those confidence intervals.

3.7 ACF and PACF

In Figure 5 we represent the ACF and PACF for several orders of AR models. Note that the PACF is relatively well informative of the order.

3.8 Time series residuals

The residuals are defined as $\hat{w}_t = observations - fitted\ value$. For example for an AR(1) model with observations $(x_t)_t$ and estimator $\hat{\alpha}$ for the (unique) parameter:

$$\hat{w}_t = x_t - \hat{\alpha} x_{t-1}.$$

The fitted valye at time t is the forecast depending on the previous obsevations.

As for linear regression, studying those residuals allow to run basic model fitness checks. For a well fitting model, the $(\hat{w}_t)$ should be close to a white noise, giving the following methods:

1. Plot the residuals as a function of time. Residuals should be uncorrelated and independently distributed around 0 following the same distribution. There shouldn't by any pattern, and this can also allow to detect outliers.
2. Using the Ljung-Box test.
3. Plot a correlogram of the residuals, if any autocorrelation coefficient has values outside of $\pm 2/\sqrt{n}$ it is significantly different from 0 at 5% significance level.

Note that the residuals are always approximately following a white noise, and care must be taken on those methods.

4 ARMA and ARIMA processes

These processes are one step further in complexity for stochastic processes. By combining AR and MA process, we can propose a process defined by the recurrent equation:

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q} \tag{1}$$

where $w \sim (0, \sigma^2)$.

We can rewrite this equation as:

$$\Phi(B)X = \Theta(B)w$$

where $\Phi(B) = 1 - \alpha_1 B - \cdots - \alpha_p B^p$, and $\Theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$

Definition 75 (ARMA process)**.** The process defined by equation 1 is called an ARMA process. We write it ARMA(p, q).

Property 76. *MA and AR processes are special cases as MA(q) corresponds to ARMA$(0, q)$ and AR(p) to ARMA$(p, 0)$.*

A white noise is then merely ARMA$(0, 0)$.

Similarly to the previous parts, and following similar a similar proof we have that:

Theorem 77 (Invertibility and causality of ARMA processes)**.** *An ARMA(p, q) process associated with Φ and Θ is*

- causal *if the roots of Φ lie outside the unit circle*
- invertible *if the roots of Θ lie outside the unit circle.*

If an ARMA process is both invertible and causal it can be expressed as both an infinite order MA and AR process.

Remark 78. *The sufficient conditions of theorem 77 are not necessary. As the shape of the ARMA trajectory only depends on the rational fraction Θ/Φ, if there exists an irreducible form $\tilde{\Theta}/\tilde{\Phi}$, the condition can be used on $\tilde{\Theta}$ and $\tilde{\Phi}$.*

Definition 79. If an ARMA process

1. is invertible
2. is causal
3. is such that Θ and Φ have no common roots (i.e. Θ/Φ is irreducible)

then it is said *regular*.

Property 80 (Wold Decomposition)**.** *If the ARMA equation writes:*

$$\Phi(B)X_t = \Theta(B)w_t$$

provided the process is invertible, we can write

$$X_t = \frac{\Theta(B)}{\Phi(B)}X_t = \Psi(B)X_t = \sum_{i=0}^{\infty}\psi_i w_{t-i}.$$

Note that we can also use a similar decomposition to write Z as a function of past X's.

This decomposition allows to compute the covariance of the process.

Theorem 81 (autocovariance of the ARMA process)**.** *Let X be a regular ARMA(p,q) process, then we can write for $k \in \mathbb{N}$:*

$$\rho_X(h) = \frac{\sum_{i=O}^{\infty}\psi_i\psi_{i+k}}{\sum_{i=0}^{\infty}\psi_i^2}$$

where ψ are the coefficients of the Wold decomposition.

Example 82 (Full study of an ARMA$(1,1)$ process)**.** *Let X be a process defined by the following equation:*

$$X_t = \alpha X_{t-1} + w_t + \beta w_{t-1}$$

where $\alpha, \beta \neq 0$ and Z is a Gaussian white noise with mean 0 and variance σ^2. This equation rewrites:

$$(1-\alpha B)X_t = (1+\beta B)w_t$$

the roots of the polynom are $1/\alpha$ and $-1/\beta$ respectively. For the process to be regular, we need that:

$$|\alpha|, |\beta| < 1, \quad \alpha \neq \beta.$$

We can thus compute a Wold decomposition of the process:

$$\begin{aligned} X_t &= \frac{1+\beta B}{1-\alpha B}w_t \\ &= (1+\alpha B+\alpha^2 B^2+\cdots+)(1+\beta B)w_t \\ &= (1+(\alpha+\beta)B+(\alpha^2+\alpha\beta)B^2+\dots)w_t \\ &= \sum_{i=0}^{\infty}\psi_i w_{t_i}. \end{aligned}$$

where $\psi_i = (\alpha+\beta)\alpha^{i-1}$ *for* $i \in \mathbb{N}^*$ *and* $\psi_0 = 1$.
From there we can compute the variance, autocorrelation, etc.:

$$var(X_t) = \sum_{i=0}^{\infty} \psi_i^2 var(Z_{t-i}) = \left(1 + \frac{(\alpha+\beta)^2}{1-\alpha^2}\right)\sigma^2$$

$$\rho_X(0) = 1 + \frac{(\alpha+\beta)^2}{1-\alpha^2} = \frac{1+\beta^2+2\alpha\beta}{1-\alpha^2}.$$

Also,

$$\psi_0\psi_1 + \psi_1\psi_2 + \psi_2\psi_3 + \cdots = (\alpha+\beta) + \left(\frac{(\alpha+\beta)^2\alpha}{1-\alpha^2}\right).$$

Therefore,

$$\rho_X(1) = \frac{(\alpha+\beta)(1+\alpha\beta)}{1+\beta^2+2\alpha\beta}.$$

Remark 83. *The ARMA process can be generalised by adding a constant term:*

$$\Phi(B)X_t = c + \Theta(B)w_t$$

This model is called ARMA(p,q) model with constant mean.
By letting $\mu = \frac{c}{1-\alpha_1-\cdots-\alpha_p} = E[X_t]$*, the problem may be reframed without the constant mean by considering* $Y_- = X_t - \mu$.

To see that, it suffices to see that $\Phi(B)Y_t = \Phi(B)(X_t - \mu) = \Phi(B)X_t - \Phi(B)\mu = c + \Theta(B) - c = \Theta(B)w_t$. As $\Phi(B)\mu = 1 - \alpha_1 - \cdots - \alpha_p$.
After Wold decomposition, we can write:

$$X_t = \mu + \sum_{i=0}^{\infty} \psi_i w_{t-i}.$$

The autocorrelation of X and Y are the same.

4.1 Fitting an ARMA model

An ARMA(p, q) has $p + q + 2$ parameters:

- $\phi_1, \ldots, \phi_p$;
- $\theta_1, \ldots, \theta_q$;
- μ, σ.

If $x_1, \ldots, x_n$ is the dataset we can propose the following procedure:

1. Identifying the possible models, this can involve removing trends in the data to ensure stationarity. We can compare correlograms and plots of the PACF to guide the choice. A quantitative criterion, such as AIC can also be used.
2. Estimating the model parameters. For example using MLE, or AIC.
3. Check of the selected model. For example by checking the residuals which should be white noise. Correlograms of the residuals could be of interest.

5 ARIMA

In practice the time series are generally not stationary. The non stationary parts of the model must be removed in order to use the previously described procedure. The most basic example is when the *trend* $E[X_t]$ is not constant over time.

As stated previously, studying the differences can allow to remove this trend:

$$Y_t = \Delta X_t = (I - B)X_t = X_t - X_{t-1}.$$

Higher order differences can also be of interest $\Delta^2 X_t = X_t - 2X_{t-1} + X_{t-2}$. As we saw previously if the trend is polynomial we retreive a stationary stochastic model.

If a process X is such that $\Delta^d X_t$ follows an ARMA(p, q) model, we say that X is an *Autoregressive integrated moving average with parameters* p, q and d (ARIMA(p, q, d)).

If X has a polynomial trend of order d, then let's define $W_t = \Delta^d X_t$, so that the process is stationary (either with mean 0 or $c \neq 0$, without loss of generality we can assume null mean). Assume a regular process ARMA(p, q) is used to model W_t, that is the roots of Θ and Φ associated with the process are outside of the unit disc and not in common between the polynoms. Let's define:

$$\Psi(B) = \Phi(B)(I - B)^d,$$

this polynom has d roots on the unit circle and p roots outside of it. We may then write:

$$\Psi(B)X_t = \Phi(B)(1 - B)^d X_t = \Phi(B)W_t = \Theta(B)w_t.$$

We can invert the process X (as all the roots of Θ are outside of the unit disc, and write

$$w_t = \frac{\Psi(B)}{\Theta(B)} X_t = \Pi(B)X_t,$$

it is nevertheless false that the process is causal, because of the added roots (and the impossibility to invert Ψ). Therefore the Wold decomposition does not exist.

Theorem 84. *If X follows an ARIMA(p, q, d) process where the associated ARMA process is regular, and if $\Pi(B) = \Psi(B)/\Theta(B) = \sum_{i=0}^{\infty} \pi_i B$, we have*

$$\Pi(1) = 1 + \pi_1 + \pi_2 + \cdots = 0$$

Part II
Forecasting, seasonal variation, detrending

6 Forecasting

Forecasting corresponds to the estimation of future values for an observed time series, that is estimating X_{n+h} for some integer h given observations $x_1, \dots x_n$.

Forecasting is a notoriously difficult task. Many elements are unknown:

- the quality of the fit;
- the validity of the model in the future;

– the inherent randomness of the process.

These different sources of uncertainty must be taken into account for forecasting.

6.1 A simple forecasting method: Linear regression

It is possible to treat a time series as a linear model to do predictions. In this case, we want to predict the value of X_{n+h} based on the data $x_1, \ldots, x_n$. The linear model will write:

$$X_t = \beta_0 + \beta_1(t - \bar{t}) + w_t, \qquad t = 1, \ldots, n, \qquad (w_t) \sim (0, \sigma^2),$$

for which the (uncorrelated) estimators are $\hat{\beta}_0 = \bar{X}$ and $\hat{\beta}_1 = \sum X_t\,(t - \bar{t}) / \sum (t - \bar{t})^2$.

The natural predictor is $X^n_{n+h} = \hat{\beta}_0 + \hat{\beta}_1(n + h - \bar{t})$ and this has variance

$$\mathrm{var}(X^n_{n+h}) = V(h) = \sigma^2 \left\{ \frac{1}{n} + \frac{(n + h - \bar{t})^2}{\sum (t - \bar{t})^2} \right\},$$

which increases quadratically in $n + h - \bar{t}$, but tends to zero for fixed h as $n \to \infty$.

If the model is correct, the future observations will be $X_{n+h} = \beta_0 + \beta_1\,(n + h - \bar{t}) + w_{n+h}$, where w_{n+h} is independent of the previous data, and then

$$\mathrm{var}(X_{n+h} - X^n_{n+h}) = \sigma^2 + V(h) \to \sigma^2 \tag{2}$$

even if the sample size $n \to \infty$.

The terms in (2) represent the uncertainty due to intrinsic variability of the system, σ^2, and that due to estimating the system, $V(h)$, to which must be added the variability due to guessing the system (here a linear model) from the data.

6.2 ARMA forecasting

In this section we will describe forecasting methods for ARMA processes, to simplify notations we assume there is no constant term. In this part we will note X^n_{n+h} the forecast made at time n for X_{n+h}. For example, when $h = 1$ we have X^n_{n+1} = forecast for X_{n+1}, made at time n.

6.2.1 Box-Jenkins approach

The general idea is to set $w_t = 0$ (its mean) for any future value of t. Thus, at time n, $w_{n+1}, w_{n+2}, \ldots$ are set to zero. X_t is set to its forecasted value for future values of t. For example, for a forecast made at time n, X_{n+1} is set to X^n_{n+1}. Then, forecasts are made using the form of the model. The approach discussed is known as the *Box-Jenkins* approach.

From the definition of an ARMA(p,q) process, we know that

$$X_{n+1} = (\alpha_1 X_n + \alpha_2 X_{n-1} + \ldots + \alpha_p X_{n+1-p}) + w_{n+1} + \theta_1 w_n + \ldots \theta_q w_{n+1-q}.$$

To obtain X^n_{n+1} the forecast for time $n + 1$, made at time n we set w_{n+1} to zero, so that

$$X^n_{n+1} = (\alpha_1 X_n + \alpha_2 X_{n-1} + \ldots + \alpha_p X_{n+1-p}) + \theta_1 w_n + \ldots \theta_q w_{n+1-q}.$$

So,

$$X^n_{n+1} = X_{n+1} - w_{n+1}. \tag{3}$$

Since (X_t) is regular, we can write w_{n+1} as:

$$\begin{aligned} w_{n+1} =& \pi(B)X_{n+1} \\ =& X_{n+1} + \pi_1 X_n + \pi_2 X_{n-1} + \dots \end{aligned}$$

with $\Pi(B) = \frac{\Phi(B)}{\Theta(B)}$.

Substituting this into (3), we get our forecast, of

$$X^n_{n+1} = -\pi_1 X_n - \pi_2 X_{n-1} - \dots$$

Example 85. *Supppose $\theta(B) = 1$ (so the model is AR(p)). We then have $\pi(B) = \phi(B)$, and hence $\pi_i = -\alpha_i$ for $i = 1, \dots p$ and $\pi_i = 0$ for $i > p$.*

Using the formula obtained for the one step ahead forecast, we get

$$X^n_{n+1} = \alpha_1 X_n + \alpha_2 X_{n-1} + \dots + \alpha_p X_{n+1-p},$$

which also follows from setting $w_{n+1} = 0$ in the formula for the AR(p) model:

$$X_{n+1} = \alpha_1 X_n + \alpha_2 X_{n-1} + \dots + \alpha_p X_{n+1-p} + w_{n+1}.$$

To build higher order forecast, that is X^n_{n+h} for $h > 1$, we will use Wold's decomposition, we have that

$$X_{n+h} = \sum_{i=0}^{\infty} \psi_i w_{n+h-i}. \tag{4}$$

On setting $w_{n+1} = w_{n+2} = \dots = w_{n+h} = 0$, we obtain

$$X^n_{n+h} = \psi_h w_n + \psi_{h+1} w_{n-1} + \dots \tag{5}$$

This is theoretically computable, since we can find $w_n, w_{n-1}, \dots$ from the formula $w_t = \pi(B)X_t$ (the inversion formula).

Denoting the error in the forecast for X_{n+h} by $e_n(h)$, we have that

$$e_n(h) = X_{n+h} - X^n_{n+h},$$

i.e. the error is the difference between the actual value and the forecasted value.

Using (4) and (5), this can be written

$$e_n(h) = w_{n+h} + \psi_1 w_{n+h-1} + \dots + \psi_{h-1} w_{n+1}.$$

6.2.2 Forecasting error

Regarding the error $e_n(h)$ as a random variable, we have

$$\begin{aligned} E[e_n(h)] =& 0 \\ \mathrm{Var}[e_n(h)] =& \left[1 + \psi_1^2 + \dots \psi_{h-1}^2\right] \sigma^2. \end{aligned}$$

Hence, if (w_t) are Gaussian white noise, a 95% confidence interval for X_{n+h} is

$$X^n_{n+h} \pm 1.96 \left[1 + \psi_1^2 + \dots \psi_{h-1}^2\right]^{1/2} \sigma.$$

In particular, when $h = 1$, we have $e_n(1) = w_{n+1} \sim N(0, \sigma^2)$, and a 95% confidence interval for X_{n+1} is

$$X_{n+1}^n \pm 1.96\sigma.$$

In practice, the approach given above is often used to obtain the variance of the estimate. There are other, recursive approaches that can be used to obtain point estimates of the forecasts (see e.g. p83-84 of Chatfield).

Example 86. *Consider the AR(1) process*

$$X_t = 0.5X_{t-1} + w_t$$

where $\{w_t\}$ denotes Gaussian white noise.

We have already seen that an AR(1) process can be written

$$X_t = \psi(B)w_t = \sum_{i=0}^{\infty} \psi_i w_{t-i},$$

where $\psi_i = \alpha^i$ for $i = 1, 2 \ldots$ if the process is causal. The root of $\phi(B) = (1 - 0.5B)$ is $1/0.5$, which lies outside the unit circle. So the process is causal, and we can use the decomposition given above. For this example, we have $\alpha = 0.5$.

The forecast for X_{n+h} is given by

$$X_{n+h}^n = \psi_h w_n + \psi_{h+1} w_{n-1} + \ldots = 0.5^h(w_n + 0.5w_{n-1} + \ldots) = 0.5^h X_n.$$

The forecasting error $e_n(h)$ is given by

$$\begin{aligned} X_{n+h} - X_{n+h}^n &= \sum_{i=0}^{\infty} 0.5^i w_{n+h-i} - 0.5^h(w_n + 0.5w_{n-1} + \ldots) \\ &= w_{n+h} + 0.5w_{n+h-1} + \ldots 0.5^{h-1} w_{n+1}, \end{aligned}$$

so $e_n(h) \sim N(0, (1 + 0.5^2 + 0.5^4 + \ldots 0.5^{2h-2})\sigma^2)$, using that $\{w_t\}$ are uncorrelated and normally distributed, with mean zero and variance σ^2.

A 95% prediction interval for X_{n+h} is given by

$$\begin{aligned} &X_{n+h}^n \pm 1.96\,\sigma \left(\sum_{i=0}^{h-1} 0.5^{2i} \right)^{1/2} \\ =&0.5^h X_n \pm 1.96\,\sigma \left(\frac{1 - 0.5^{2h}}{1 - 0.5^2} \right)^{1/2} \end{aligned}$$

where this final simplification uses the sum of a geometric series.

6.3 Notes on the forecasting methods

The best linear predictor for a process with mean μ has the form

$$X_{n+h}^n = \mu + \sum_{j=h}^{\infty} \psi_j w_{n+h-j}$$

but because the ψ_j converge to zero exponentially fast for any causal ARMA model, we have $X^n_{n+h} \to \mu$ as $h \to \infty$.

Likewise for large h, the prediction error

$$\text{var}(e(h)) = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2 \to \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 \qquad \text{exponentially fast as } h \to \infty$$

6.3.1 Prediction uncertainty

We've already seen that there are three components to prediction uncertainty:

- model uncertainty
- estimation uncertainty
- innovation variability

Previous discussion pertains only to the innovation uncertainty, assuming that the model structure and its parameters are known we can use *bootstrap simulation* (Davison, A. C. and Hinkley (2013). *Bootstrap Methods and their Application*, CUP) to include the other two. The procedure is the following:

- we simulate R new sets of data $X_1^*, \ldots, X_n^*$ from some plausible model;
- we fit the model class to each of these datasets, choosing the 'best' model by AIC or similar, and estimating the parameters
- we use the 'best' fitted model to obtain X^{n*}_{n+h} for each dataset
- finally we use the R replicates of X^{n*}_{n+h} to assess the uncertainty of the original prediction X_{n+h}.

6.4 Forecasting with ARIMA(p,d,q) models

We remind that for ARIMA models, $\Phi(B)$ is replaced with $\Phi(B) = \Phi(B)(I - B)^d$. Under this setting, most of the theory discussed for forecasting with ARMA models can still be used. However, the Wold decomposition is no longer used. Instead, forecasts may be obtained recursively, by setting future values of w to zero as before (we do not go into detail here, although specific examples are given later).

We will discuss two ad-hoc forecasting methods, which may be used on both ARMA and ARIMA models: exponential smoothing and Holt's method. It turns out that these methods are equivalent to the method discussed above for some specific ARIMA models.

6.4.1 Exponential smoothing

The method of *Exponential smoothing* forecasts X^n_{n+1} using a weighted sum of past observations (where (X_t) is a time series with no trend):

$$X^n_{n+1} = (1-\theta)X_n + (1-\theta)\theta X_{n-1} + (1-\theta)\theta^2 X_{n-2} + \ldots,$$

where θ is some constant such that $0 < \theta < 1$ (so all the weights are positive).

A recursive version of this formula is then

$$X_{n+1}^n = (1-\theta)X_n + \theta X_n^{n-1},$$

i.e. a weighted average of the current value (X_n) and the previous forecast for X_n.

A suitable constant, θ may be obtained by minimising errors for forecasts of known values in the series.

In practice, a starting value such as X_t^{-t-1} is needed, where t is some past time.

Remark 87. *Note that this method does not assume a particular model for the series* (X_t).

Geometric interpretation of exponential smoothing Asymptotically, X_{n+h}^n is the best least square approximation (for exponential weighting) of x by the constant vector $\theta\delta$, with $\delta = (1, \dots, 1) \in \mathbb{R}^n$. As the minimiser:

$$\text{argmax}_s \sum_{j=0}^{n-1} \theta^j (x_{n-j} - s)^2$$

is given by

$$\hat{s}(n) = \frac{1-\theta}{1-\theta^n} \sum_{j=0}^{n-1} \theta^j x_{n-j}$$

by writing $x = (x_0, \dots, x_n) = (\dots, 0, x_1, \dots, x_n)$, that is we set to 0 the values $x_j, j < 0$, we get to minimize

$$\sum_{j=0}^{\infty} \theta^j (x_{n-j} - \theta)^2$$

the solution is given by

$$X_{n+h}^n = (1-\theta) \sum_{j=0}^{n-1} \theta^j x_{n-j}.$$

It is not difficult to show that the two estimators are equivalent as $n \to \infty$.

We can see in the expression that the influence of the observations decreases exponentially, depending on θ, which gives the name to the method.

Example

Example 88 (Exponential smoothing and ARIMA$(0,1,1)$ models)**.** *Consider the ARIMA(0,1,1) process*

$$X_t = X_{t-1} + w_t - \theta w_{t-1},$$

where $(w_t) \sim (0, \sigma^2)$.

We have shown that

$$w_t = \sum_{i=0}^{\infty} \pi_i X_{t-i},$$

where $\pi_i = -(1-\theta)\theta^{i-1}$ *for* $i = 1, 2\ldots$ *and* $\pi_0 = 1$. *Using the Box-Jenkins approach (i.e. the approach discussed for ARMA(p,q) models), we have that*

$$\begin{aligned} X_{n+1}^n =& X_{n+1} - w_{n+1} \\ =& X_{n+1} - \sum_{i=0}^{\infty} \pi_i X_{n+1-i} \\ =& -\pi_1 X_n - \pi_2 X_{n-1} - \ldots \end{aligned}$$

Therefore, using our equation for π_i, *we get*

$$X_{n+1}^n = (1-\theta)X_n + (1-\theta)\theta X_{n-1} + (1-\theta)\theta^2 X_{n-2} + \ldots$$

which is the exponential smoothing formula, provided that θ *satisfies* $0 < \theta < 1$. *Thus, when the process* (X_t) *is ARIMA(0,1,1), exponential smoothing corresponds to the Box-Jenkins approach discussed previously.*

Exponential smoothing may be used when (X_t) has no systematic trend, so that $E[X_t] = \mu$ (a constant, maybe $\neq 0$). When $(X_t) \sim \text{ARIMA}(0,1,1)$ (without constant), exponential smoothing gives the same results as the Box-Jenkins approach if $0 < \theta < 1$. Using the recursive formula, it is possible to find X_{n+1}^n from X_n and X_n^{n-1}.

Adjustments for a finite number of past observations The exponential smoothing formula requires an infinite number of past observations, which is not possible in practice. Thus, the weights used in exponential smoothing should be modified so that they sum to one. This may be done in various ways, e.g.:

1. Increase the weight on X_1 to w, such that

$$(1-\theta) + (1-\theta)\theta + \ldots + (1-\theta)\theta^{n-2} + w = 1$$

. This gives

$$\begin{aligned} w =& 1 - (1-\theta)\left[1 + \theta + \theta^2 + \ldots + \theta^{n-2}\right] \\ =& 1 - (1-\theta)\frac{(1-\theta^{n-1})}{(1-\theta)} \\ =& \theta^{n-1}. \end{aligned}$$

2. Multiply all of the weights by a factor, f such that

$$f[(1-\theta) + (1-\theta)\theta + \ldots + (1-\theta)\theta^{n-2} + (1-\theta)\theta^{n-1}] = 1.$$

Then

$$f\left[(1-\theta)\frac{[1-\theta^n]}{(1-\theta)}\right] = 1,$$

and so

$$f = \frac{1}{1-\theta^n}.$$

Remark 89. *When* n *is large, both of these approaches give similar results.*

6.4.2 Holt's method

The *Holt method of forecasting* is widely used and may be employed when there is a **linear trend** in the series (i.e. $E[X_t] = \alpha + \beta t$ for some constant α, β). The *Holt-Winters method* (omitted) is an extension of Holt's method to include seasonal variations.

Suppose we want to find X_{n+1}^n (we are currently at time n). Holt's method sets

$$X_{n+1}^n = a_{n+1} + b_{n+1}$$

where

1. a_{n+1} is an estimate of the 'local' level of the series (i.e. ignoring the trend) at time n. Using the same approach as exponential smoothing, we will set, for some $\alpha \quad (0 < \alpha < 1)$,

   $$a_{n+1} = \alpha X_n + (1-\alpha) X_n^{n-1}$$

   so that a_{n+1} is a weighted average of X_n and of the forecast from the previous time point.

2. b_{n+1} is an estimate of the trend (so the increase from time n). We will let for some $\gamma \quad (0 < \gamma < 1)$

   $$b_{n+1} = \gamma(a_{n+1} - a_n) + (1-\gamma) b_n$$

   so that b_{n+1} is a weighted average of the change in 'level' from time n to time $n+1$ $(a_{n+1} - a_n)$ and the estimate of the trend at time n (b_n).

We will show that Holt's method agrees with the usual (Box-Jenkins) method if

1. The underlying time series is ARIMA(0,2,2) (without constant).
2. α and γ are chosen suitably.

But Holt's method can be used more generally, choosing α and γ using error minimisation methods. As with exponential smoothing, a particular model is not assumed for the data.

Theorem 90. *A recursive formula for Holt's method is given by*

$$X_{n+1}^n = \alpha(1+\gamma) X_n - \alpha X_{n-1} + (2 - \alpha - \alpha\gamma) X_n^{n-1} - (1-\alpha) X_{n-1}^{n-2}$$

Remark 91. *in practice, to use this recursive formula, we would need two starting values, e.g.* X_0^{-1} *and* X_{-1}^{-2}.

Example 92 (Holt's method and ARIMA$(0,2,2)$ model)**.** *Consider the ARIMA(0,2,2) model*

$$X_t = 2X_{t-1} - X_{t-2} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2},$$

where θ_1 *and* θ_2 *are constants and* $\{w_t\} \sim WN(0, \sigma^2)$.

To find a recursive formula for X_{n+1}^n *using the usual (Box-Jenkins) method, set* $w_{n+1} = 0$ *and write*

$$X_{n+1}^n = 2X_n - X_{n-1} + \theta_1 w_n + \theta_2 w_{n-1}.$$

Now note that

$$w_n = X_n - X_n^{n-1} \text{ and } w_{n-1} = X_{n-1} - X_{n-1}^{n-2},$$

and hence

$$X_{n+1}^n = 2X_n - X_{n-1} + \theta_1 w_n + \theta_2 w_{n-1}$$

$$= 2X_n - X_{n-1} + \theta_1(X_n - X_n^{n-1}) + \theta_2(X_{n-1} - X_{n-1}^{n-2})$$

$$= (2+\theta_1)X_n - \theta_1 X_n^{n-1} - (1-\theta_2)X_{n-1} - \theta_2 X_{n-1}^{n-2}.$$

This recursive formula has the same form as the recursive formula for the Holt method if we have that

1. $\alpha(1+\gamma) = (2+\theta_1)$
2. $\alpha = (1-\theta_2)$
3. $(2-\alpha-\alpha\gamma) = -\theta_1$
4. $(1-\alpha) = \theta_2$.

From (2) and (4), we see that $\alpha = (1-\theta_2)$. Using (1), we get that

$$\gamma = \frac{(2+\theta_1)}{(1-\theta_2)} - 1 = \frac{1+\theta_1+\theta_2}{1-\theta_2}.$$

Substituting these equations for α and γ into $(2-\alpha-\alpha\gamma)$, we get $-\theta_1$ (exercise: check), so these equations are consistent with (3).

We see that if $\alpha = (1-\theta_2)$ and $\gamma = 1+\theta_1+\theta_2/1-\theta_2$, the Holt method gives the same recursive formula as the Box-Jenkins approach, and so for $ARIMA(0,2,2)$ models, the two forecasting approaches are equivalent.

7 Seasonal variation

Time series may exhibit seasonal variation. Consider, for example, temperature readings or sales figures. An example is given in Figure 6. These variations can have several origins: meteorology, recurring events, hysteresis in the process generating the data, etc.

To model time series with seasonal variation, the data can be transformed by removing the seasonal component. The series can then be modelled using the techniques for stationary series. Forecasts can be made by forecasting the transformed stationary series, and then inverting any transformations made. We will go through an example that has quarterly seasonal variations. The methods can be adjusted to consider other types of seasonal variation, e.g. monthly.

Consider a quarterly time series, such that time is measured in quarter years. E.g. It might be suspected that there is seasonal variation in this data. In a cool country, demand for electricity might be higher in the winter than in the summer.

A plot of the data is given in Fig. 7. It seems that there is both seasonal variation and a trend.

Let time, t be measured in quarter years since the beginning of the observation period (which lasts for n quarters). For a series with quarterly seasonal variation, we let

$$x_t = u_t + e'_t + s_t \quad \text{for} \quad 1 \le t \le n,$$

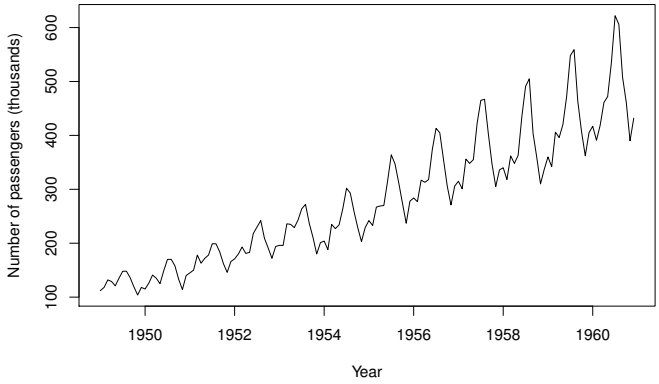
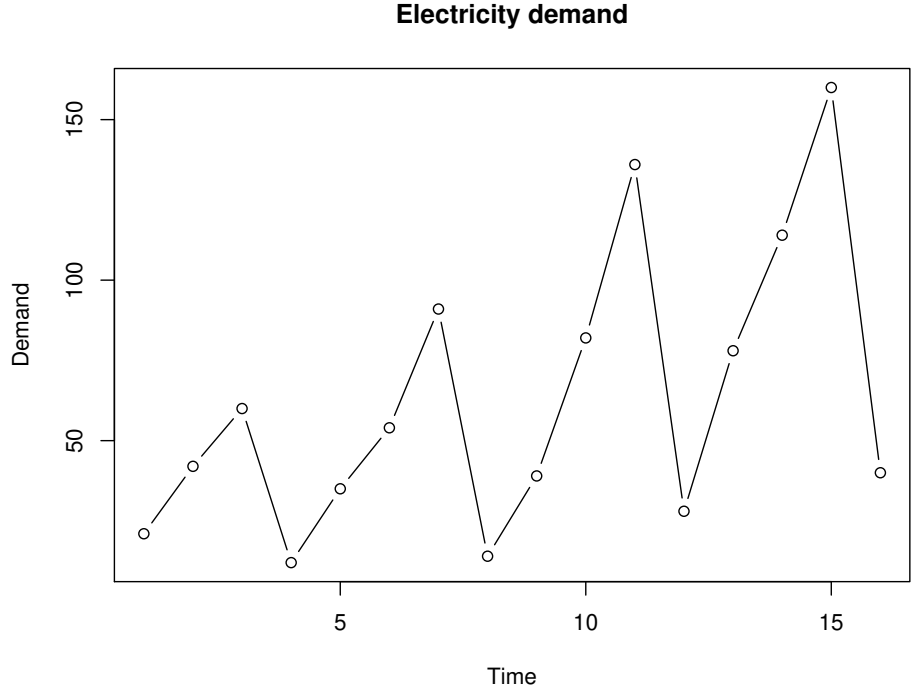


Figure 6: The Box-Jenkins data, with monthly totals of international airline passengers in the USA, from January 1949 to December 1960.

Months	2003	2004	2005	2006
Jan-Mar	21	35	39	78
Apr-Jun	42	54	82	114
Jul-Sept	60	91	136	160
Oct-Dec	12	14	28	40

Figure 7: Quarterly demand for electricity in a government building. Data provided by a former MSc in OR student.



be the raw observations at time t, where u_t is the 'trend' at time t, e'_t is a random error with zero mean and s_t is the seasonal variation at time t.

It is assumed that s_t takes the values s^I, s^{II}, s^{III} and s^{IV} in quarters I, II, III and IV respectively. By convention, we assume that

$$s^I + s^{II} + s^{III} + s^{IV} = 0.$$

The usual procedure is first to smooth the raw data. There are various methods of doing this (one of which we will discuss later).

For quarterly seasonal variation we will use the 5-term adjusted-average formula with coefficients

$$\left(\frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}\right),$$

which may be applied for all t, except the two values in the upper tail and the two values in the lower tail. The 5-term adjusted-average formula with these coefficients sets

$$y_t = \frac{1}{8}x_{t-2} + \frac{1}{4}(x_{t-1} + x_t + x_{t+1}) + \frac{1}{8}x_{t+2} \quad (3 \leq t \leq n-2)$$

and

$$e_t = x_t - y_t \quad (3 \leq t \leq n-2).$$

To calculate the seasonal variation in each quarter, we start by taking the average of e_t over all observations that relate to the first quarter, calling this s^I_*. Similarly, we obtain s^{II}_*,s^{III}_* and s^{IV}_* for quarters two, three and four.

To ensure that the seasonal variations add to zero, we set

$$\begin{aligned} s^I &= s^I_* - \text{adj} \\ s^{II} &= s^{II}_* - \text{adj} \\ s^{III} &= s^{III}_* - \text{adj} \\ s^{IV} &= s^{IV}_* - \text{adj} \end{aligned}$$

where adj $= \frac{1}{4}(s^I_* + s^{II}_* + s^{III}_* + s^{IV}_*)$

The 'de-seasonalised' (or 'seasonally adjusted') series may be found by subtracting the appropriate seasonal variation from x_t (e.g. if t is in quarter 1, $s(t) = s^I$). This gives a series of the form $z_t = u_t + e'_t$. The trend u_t can be removed using other methods (e.g. differencing, as in ARIMA models or filtering, see next lecture).

Remark 93. *for data with seasonal variation which is not quarterly, another smoothing method should be used. For example, for monthly seasonal variation, a common choice for smoothing is the adjusted average formula*

$$y_t = \frac{1}{24}x_{t-6} + \frac{1}{12}(x_{t-5} + \ldots + x_t + \ldots x_{t+5}) + \frac{1}{24}x_{t+6}.$$

Example 94 (Electricity consumption data)**.** *For the data concerning quarterly electric use, we get*

t	x_t	y_t	e_t	*Quarter*
1	*21*	x	x	*I*
2	*42*	x	x	*II*
3	*60*	*35.50*	*24.50*	*III*
4	*12*	*38.75*	*-26.75*	*IV*
5	*35*	*44.13*	*-9.13*	*I*
6	*54*	*48.25*	*5.75*	*II*
7	*91*	*49.00*	*42.00*	*III*
8	*14*	*53.00*	*-39.00*	*IV*
9	*39*	*62.13*	*-23.13*	*I*
10	*82*	*69.50*	*12.50*	*II*
11	*136*	*76.13*	*59.88*	*III*
12	*28*	*85.00*	*-57.00*	*IV*
13	*78*	*92.00*	*-14.00*	*I*
14	*114*	*96.50*	*17.50*	*II*
15	*160*	x	x	*III*
16	*40*	x	x	*IV*

This leads to $s^I_* = -15.424$, $s^{II}_* = 11.92$, $s^{III}_* = 42.13$ $s^{IV}_* = -40.92$ *and adj* $= -0.57$. *As a result*

$$\begin{aligned} s^I &= -14.85 \\ s^{II} &= 12.49 \\ s^{III} &= 42.70 \\ s^{IV} &= -40.35. \end{aligned}$$

8 Removing trends

We have so far described the 'trend' of a time series as the long term change in mean, $\mathbb{E}[X_t]$. More formally, we will use the notation

$$X_t = u_t + s_t + E_t \quad \text{for} \quad t \in \mathbb{Z},$$

where u_t is the 'trend' at time t, s_t is the seasonal component at time t and E_t is a random error with zero mean. It is assumed that if the seasonality has period d, then $s_t + s_{t+1} + \ldots s_{t+d-1} = 0$. This model is known as the *classical decomposition model*. For example, if the series has a linear trend, we might set $u_t = a + bt$, where a and b are constants.

Above we saw how to estimate the seasonal component s_t, based on observations $x_1, \ldots x_n$. We will now discuss methods of removing the trend, u_t.

The idea is that by removing the deterministic components s_t and u_t, we will hopefully be left with a stationary stochastic process E_t. Then, the stationary models we have already studied (ARMA) can be fitted to E_t and used to estimate and forecast the process (adding the trend and seasonality back in as appropriate). We have already discussed the use of differencing (ARIMA models) to remove trend. There are two other common procedures for trend removal (known as *detrending*). These are

1. Fitting a parametric function to the data.
2. *Filtering.*

In this section, we will focus on the use of filters for detrending a series.

8.1 Filtering

We first introduced filters in the first part of this lecture, as it is the main notion to define the regression equations. Here we propose several methods where we apply a filter to a time series to remove its seasonal variation.

8.1.1 Adjusted average graduation

The process of *adjusted-average graduation* (also known as a *linear filter*) converts one set of time series observations $((X_t))$ into another $(\{y_t\})$ using the formula

$$y_t = \sum_{j=\alpha}^{\beta} K_j x_{t+j},$$

where $\{K_j; \quad j = \alpha, \alpha+1, \ldots, \beta\}$ is a set of coefficients.

The values (y_t) are the **smoothed** or **graduated** observations. These smoothed observations are used as an estimate of the trend, so $\hat{u}_t = y_t$. Thus, a weighted average is used to estimate the trend. It therefore makes sense to have $\sum K_j = 1$. E.g. when adjusting for seasonal variations we used a linear filter with

$$K_{-2} = \frac{1}{8}, K_{-1} = \frac{1}{4}, K_0 = \frac{1}{4}, K_1 = \frac{1}{4}, K_2 = \frac{1}{8},$$

and then subtracted the estimated trend (y_t) from the observations (x_t) in order to estimate s_t.

Remark 95. *We may need to use adjustments to the above formula to smooth observations at the edges of the dataset.*

The length *(l) of the adjusted average formula is given by the number of coefficients, i.e.*

$$l = \beta - \alpha + 1$$

A central *formula has $\alpha = -\beta$ (so the coefficients K_j run from $j = -\beta$ to $j = \beta$, and the length is $2\beta + 1$).*

A symmetric *formula is a central formula such that*

$$K_j = K_{-j} \quad \textit{for} \quad j = 1, 2, \ldots \beta$$

8.1.2 In practice

Suppose we have an observed series $x_1, \ldots x_n$, with seasonal variation of period d (e.g. for quarterly data the period is 4) and a trend. Using the classical decomposition model, we have that

$$x_t = u_t + s_t + e_t, \quad \text{for} \quad t = 1, \ldots n.$$

Brockwell and Davis (2002) suggest the following steps to remove seasonality and trend:

1. Use an adjusted average formula of the following form

$$y_t = \begin{cases} \frac{1}{d}\left(\frac{1}{2}x_{t-q} + x_{t-q+1} + \ldots + \frac{1}{2}x_{t+q}\right) & \text{for} \quad d = 2q, \quad q < t < n - q, \\ \frac{1}{d}\left(x_{t-q} + x_{t-q+1} + \ldots + x_{t+q}\right) & \text{for} \quad d = 2q+1, \quad q+1 < t < n - q. \end{cases}$$

   to obtain an initial estimate of the trend, $\hat{u}_t = y_t$.

2. Subtract $\hat{u}_t$ from the observations x_t.

3. Estimate the seasonal effects as done in the previous lecture to obtain $\hat{s}_t$ (adjusted so that $\sum_{t=1}^{d} \hat{s}_t = 0$).

4. Subtract the estimates of the seasonal effects to obtain a new series $x'_t = x_t - \hat{s}_t$.

5. Use a suitable adjusted average formula to estimate the trend of the new series $\{x'_t\}$ (let this new estimate now be $\hat{u}_t$).

6. Subtract this estimate of the trend from (x'_t), to obtain a new series $x''_t = x_t - \hat{s}_t - \hat{u}_t$.

7. Model this new series (x''_t) with a stationary process.

Note that steps 1 to 4 are equivalent to those previously discussed for the removal of trend for quarterly seasonal variation.

Assuming that the seasonal variation has been removed from the observations $x_1, \ldots x_n$, we have that

$$\begin{aligned} y_t =& \sum_{j=\alpha}^{\beta} K_j x_{t+j} = \sum_{j=\alpha}^{\beta} K_j (u_{t+j} + e_{t+j}) \\ =& \sum_{j=\alpha}^{\beta} K_j u_{t+j} + \sum_{j=\alpha}^{\beta} K_j e_{t+j} \\ :=& u'_t + e'_t \end{aligned}$$

To use y_t as an estimator of the trend at time t, we would like that $e'_t = \sum_{j=\alpha}^{\beta} K_j e_{t+j} \approx 0$, and that the (e'_t) are somehow 'smaller' and 'smoother' than (e_t); and that u'_t is equal to, or close to u_t, i.e. so that there is *no distortion of the trend*.

8.2 Distortion of trend

In practice, it is possible to choose the coefficients $\{K_j\}$ so that specific trends (k-th order polynomials) are not distorted by the filter. We will first derive some conditions on the coefficients that result in filters that do not distort linear trends (i.e. trends of the form $a + bt$ for constants a and b). Assume that the trend $u_t = a + bt$. In order for this trend to not be distorted, we must have that $u'_t = u_t$ for $t = 1, \ldots n$. Substituting $u_t = a + bt$ into the formula for u'_t we get

$$\begin{aligned} u'_t =& \sum_{j=\alpha}^{\beta} K_j u_{t+j} = \sum_{j=\alpha}^{\beta} K_j (a + b(t+j)) \\ =& (a + bt) \sum_{j=\alpha}^{\beta} K_j + b \sum_{j=\alpha}^{\beta} j K_j. \end{aligned}$$

So if the coefficients are such that

1. $\sum_{j=\alpha}^{\beta} K_j = 1$ and
2. $\sum_{j=\alpha}^{\beta} j K_j = 0$,

we will have that

$$u'_t = a + bt = u_t,$$

and the filter will allow the linear trend to pass without distortion.

Note that if the filter is symmetric, then $\alpha = -\beta$ and $K_j = K_{-j}$ so condition (2) is automatically satisfied.

The adjusted-average formula used for the example on quarterly seasonal variation had coefficients $\left(\frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}\right)$. This formula is symmetric and has

$$\sum_{j=-2}^{2} K_j = 1.$$

Therefore, this filter does not distort linear trend, and is 'exact on straight lines'.

Example 96 (Cubic trend)**.** *Suppose that the trend is cubic, so that* $u_t = a + bt + ct^2 + dt^3$*. We now have that*

$$\begin{aligned} u'_t =& \sum_{j=\alpha}^{\beta} K_j u_{t+j} = \sum_{j=\alpha}^{\beta} K_j (a + b(t+j) + c(t+j)^2 + d(t+j)^3) \\ =& \sum_{j=\alpha}^{\beta} K_j (a + bt + bj + c(t^2 + 2tj + j^2) + d(t^3 + st^2 j + 3j^2 t + j^3) \\ =& (a + bt + ct^2 + dt^3) \sum K_j + (b + 2ct + 3t^2 d) \sum j K_j \\ & + (c + 3td) \sum j^2 K_j + d \sum j^3 K_j \end{aligned}$$

So, we will have $u'_t = a + bt + ct^2 + dt^3$ *if and only if the following four conditions hold:*

1. $\sum_{j=\alpha}^{\beta} K_j = 1$
2. $\sum_{j=\alpha}^{\beta} j K_j = 0$
3. $\sum_{j=\alpha}^{\beta} j^2 K_j = 0$
4. $\sum_{j=\alpha}^{\beta} j^3 K_j = 0$

For symmetric filters conditions (ii) and (iv) are automatically satisified, so we need only check conditions (i) and (iii). Similar conditions can be obtained for quadratic trends and trends which are polynomials of higher orders.

8.2.1 Spencer's 15 point average filter

Spencer's 15 point average filter is a symmetric filter with coefficients

$$\begin{aligned} &(K_0, K_{\pm 1}, K_{\pm 2}, K_{\pm 3}, K_{\pm 4}, K_{\pm 5}, K_{\pm 6}, K_{\pm 7}) = \\ &\frac{1}{320}(74, 67, 46, 21, 3, -5, -6, -3). \end{aligned}$$

This filter is symmetric so we only need to check conditions (i) and (iii) to check whether the filter is exact on cubics.
We have that

- $\sum_{j=\alpha}^{\beta} K_j = 1$ so condition (i) holds.
- $\sum_{j=\alpha}^{\beta} j^2 K_j = 2\left(\sum_{j=1}^{\beta} j^2 K_j\right)$ so condition (iii) is satisfied if $\sum_{j=1}^{\beta} j^2 K_j = 0$. We have

  $$\frac{1}{320}(1 \times 67 + 2^2 \times 46 + 3^2 \times 21 + 4^2 \times 3 + 5^2 \times -5 + 6^2 \times -6 + 7^2 \times -3) = 0,$$

  so condition (iii) holds.

8.2.2 The problem of the tails

The central filters discussed can not be applied for observations in the 'tails' of the dataset (i.e. observations at the start and end of the dataset).

Instead, we must use non-central filters, that use only present and past values of x_t. This problem is particularly important for forecasting.

A common example of a non-central filter is exponential smoothing (see forecasting section), where

$$y_t = \sum_{j=0}^{\infty} (1-\theta)\theta^j x_{t-j} \quad \text{for} \quad 0 < \theta < 1.$$

Methods for adjusting the weights for a finite number of past observations can be used here.

8.3 Fitting a polynomial

Another method of removing and estimating a trend using the classical decomposition model is by fitting a k-th order polynomial to the observations $x_1, \dots x_n$. To do this, let the trend $u_t = a_0 + a_1 t + \dots a_k t^k$. Least squares estimation can be used to find estimates of the coefficients $a_0, a_1, \dots a_k$. This involves minimizing the sum of squares, given by

$$\sum_{t=1}^{n} (x_t - u_t)^2 .$$

The fitting of such a polynomial should be done after any seasonal adjustments have been made to the series.

8.4 Differencing

If an estimate of the trend is not required, detrending can be carried out using differencing. As we have already seen, if the observed series (X_t) is thought to have a polynomial trend of degree k, then the k-th order differences $\Delta^k x_t$ should be considered. For example, if the trend is quadratic, with $u_t = a + bt + ct^2$, then assuming that $x_t = u_t + e_t$ for stationary $\{e_t\}$,

$$\begin{aligned}
\Delta^2 x_t =& x_t - 2x_{t-1} + x_{t-2} \\
=& u_t + e_t - 2(u_{t-1} - e_{t-1}) + u_{t-2} + e_{t-2} \\
=& a + bt + ct^2 - 2a - 2b(t-1) - 2c(t-1)^2 \\
& + a + b(t-2) + c(t-2)^2 + (e_t - 2e_{t-1} + e_{t-2}) \\
=& (a - 2a + a + 2b - 2c - 2b + 4c) + t(b - 2b + 4c + b - 4c) \\
& + t^2(c - 2c + c) + e'_t \\
=& 2c + e'_t, \quad \text{where} \quad e'_t = \Delta^2 e_t.
\end{aligned}$$

Thus, the second differences are a stationary process, with constant of $2c$. Differencing in this way should be applied after any seasonal adjustments have been made.

The technique of differencing can also be used to remove seasonality in a series. If observations $x_1, \dots x_n$ are thought to have seasonality of period d, then the difference

$$(1 - B^d)x_t = x_t - x_{t-d}$$

should be considered. Assume we have

$$x_t = u_t + s_t + e_t,$$

with s_t representing the deterministic seasonal effect with period d, so that $s_t = s_{t-d}$, and $\{e_t\}$ stationary. This means that

$$(1 - B^d)x_t = u_t + s_t + e_t - u_{t-d} - s_{t-d} - e_{t-d} = (u_t - u_{t-d}) + (e_t - e_{t-d}).$$

The term $u_t - u_{t-d}$ is the trend component of the differenced series. This can be estimated using any of the methods discussed. The term $e_t - e_{t-d}$ is a stationary series. Note that $(1 - B^d) \neq \Delta^d = (1 - B)^d$.

Example 97 (Seasonality and trend)**.** *We will consider an example concerning Edinburgh house prices from the first quarter of 1993 to the first quarter of 2005. The following gives the first few entries of a dataset with quarterly measurements of the average house price (in £) in Edinburgh.*

Quarter	*Year*	*Price (£)*
Q1	*1993*	*57607.90*
Q2	*1993*	*61504.50*
Q3	*1993*	*60762.60*
Q4	*1993*	*64394.70*
Q1	*1994*	*62848.60*
Q2	*1994*	*67671.20*
Q3	*1994*	*64035.40*
Q4	*1994*	*67702.60*

A plot of the raw data is given in Figure 8. We see a clear trend but it is not clear whether there is any seasonal variation. We will assume the classical decomposition model of

$$x_t = u_t + s_t + e_t,$$

where x_t is the observed house price at time t, u_t is the trend, s_t is the seasonal effect and e_t is a stationary series with mean zero. As we have quarterly data, we would expect seasonal variation to have period 2, so we set $d = 4$, and $s_t = s_{t-d}$ for all t.

The first few lines of R output are

Date	*Price*	ERR_1	SAS_1	SAF_1	STC_1
Q1 1993	*57607.90*	*1794.47*	*60798.24*	*-3190.34*	*59003.78*
Q2 1993	*61504.50*	*-3449.07*	*56490.89*	*5013.61*	*59939.96*
Q3 1993	*60762.60*	*718.42*	*62530.73*	*-1768.13*	*61812.31*
Q4 1993	*64394.70*	*1156.80*	*64449.83*	*-55.13*	*63293.04*
Q1 1994	*62848.60*	*1520.51*	*66038.94*	*-3190.34*	*64518.44*
Q2 1994	*67671.20*	*-2216.33*	*62657.59*	*5013.61*	*64873.92*

The values in the SAF_1 column are the values s^I, s^{II}, s^{III} and s^{IV}. Note that the two values shown in Q1 are the same.

The values in SAS_1 are $x_t - \hat{s}_t$. R has estimated the trend u_t using an adjusted average formula with coefficients $\frac{1}{9}(1, 2, 3, 2, 1)$, with adjustments in the tails (applied to the seasonally adjusted data). This formula is exact on straight lines. This estimated trend is given in the STC_1 column. We could use another formula for this if we wanted. The ERR_1 column is $x_t - \hat{s}_t - \hat{u}_t$, i.e. it is the estimates $\hat{e}_t$. These should hopefully form a stationary series.

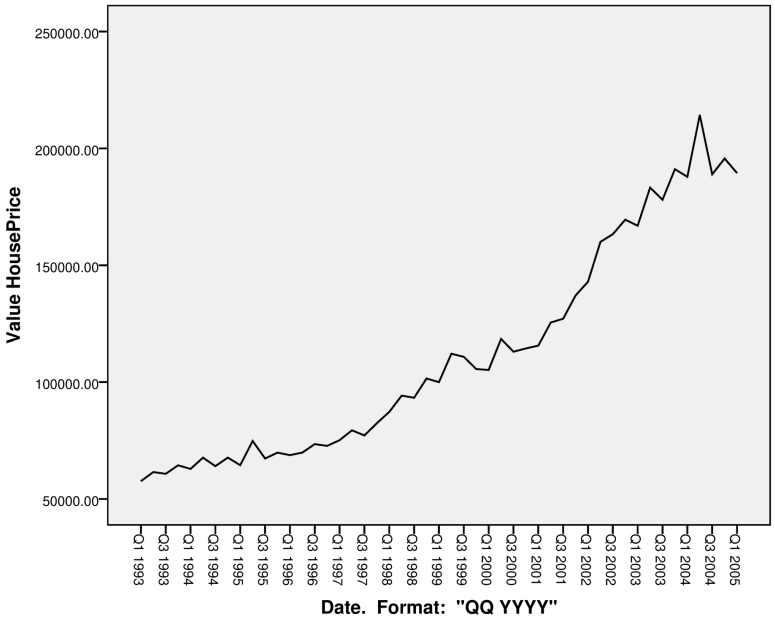


Figure 8: Edinburgh house prices between first quarter 1993 and 2005.

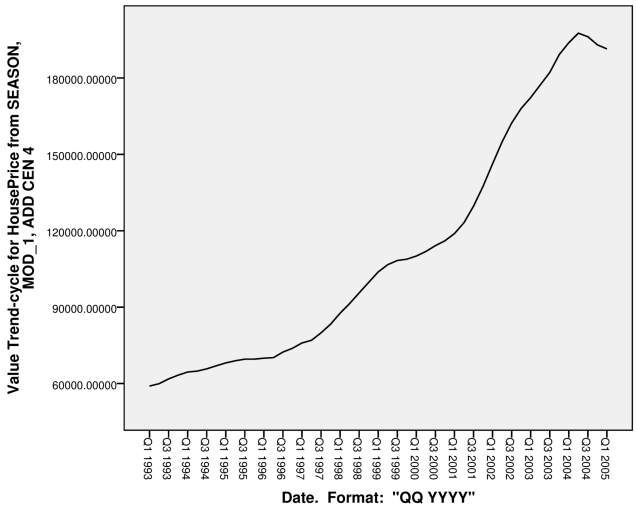


Figure 9: House price in Edinburgh, estimated trend

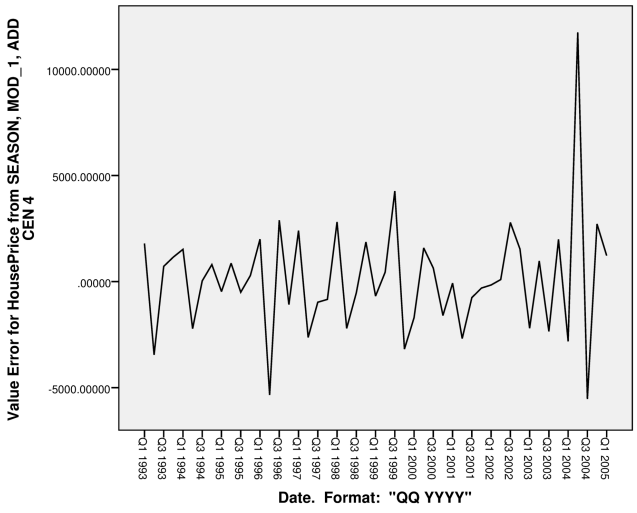


Figure 10: House price in Edinburgh, error plot

We can plot the estimated trend, to help us see the general change in house prices in Edinburgh, without the seasonal effect. We see (unsurprisingly) that house prices rose fairly steadily over the time period.

We can also plot the errors, in Figure 10. They seem fairly stationary, although there was a big error around Q2 of 2004. We might need to investigate this.

The Ljung-Box statistics for the errors imply that they are not white noise. We should therefore investigate possible stationary models to fit to the error terms.

A correlogram of the errors is presented in Figure 11.

A PACF for the errors is shown in Figure 12. It looks like an AR(1) model might be appropriate. What would you estimate for the parameter α_1*?*

Part III
Multivariate time series: VARMA and GARCH models

In this part we introduce two models, VARMA and GARCH that can be used when the data is multivariate.

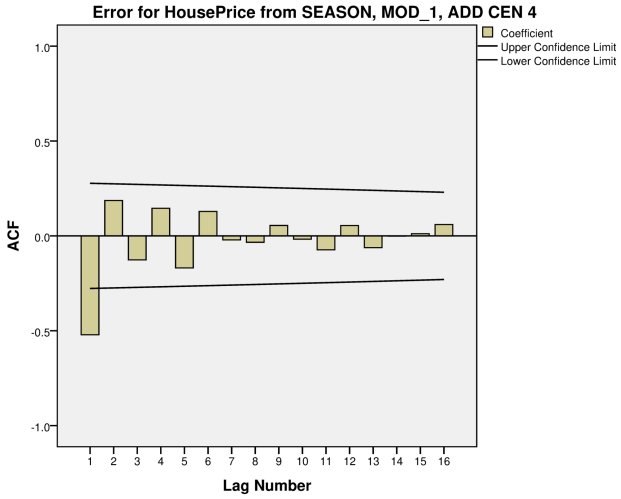


Figure 11: House price in Edinburgh, correlogram

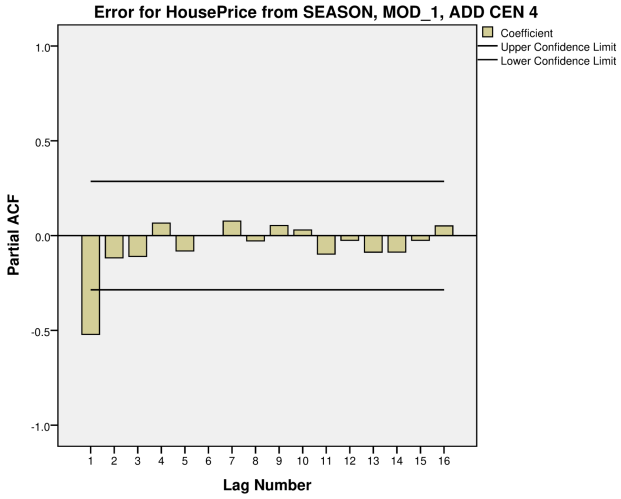


Figure 12: House price in Edinburgh, PACF plot

9 VARMA models

ARMA modelling may be extended to multivariate cases, but is much more complicated. One simple case is the *vector autoregressive (VAR)*, where with Y_t a $k \times 1$ vector, we have

$$Y_t = \mu_{k\times 1} + \sum_{j=1}^{p} \Phi_j Y_{t-j} + w_t, \quad w_t \overset{\text{iid}}{\sim} \mathcal{N}(0_{k\times 1}, \Sigma_{k\times k}),$$

where $\Phi_1, \ldots, \Phi_p$ are $k \times k$ matrices and $\Phi_p \neq 0$; such models have a lot of parameters. The VMA models can be defined likewise. This extends to the vector autoregressive moving average (VARMA) model,

$$Y_t = \mu_{k\times 1} + \sum_{j=1}^{p} \Phi_j Y_{t-j} + w_t + \sum_{j=1}^{p} \Theta_j w_{t-j} \quad w_t \overset{\text{iid}}{\sim} \mathcal{N}(0_{k\times 1}, \Sigma_{k\times k})$$

where $\Theta_1, \ldots, \Theta_q$ are also $k \times k$ matrices with $\Theta_q \neq 0$.

We define the AR and MA operators $\Phi(B)_{k\times k} = I - \Phi_1 B - \Phi_2 B^2 - \cdots \Phi_p B^p$ and $\Theta(B)_{k\times k} = I + \Theta_1 B + \Theta_2 B^2 + \cdots \Theta_q B^q$; the process is causal if the roots of $|\Phi(z)|$ lie outside $\mathbb{D}$, and is invertible if the roots of $|\Theta(z)|$ lie outside $\mathbb{D}$

Complicated conditions are needed to ensure uniqueness and identifiability, but can be avoided by fitting only VAR models—see Tsay (2014) *Multivariate Time Series Analysis with R and Financial Applications*. John Wiley.

9.1 Fitting VARMA models: Example on EU markets return

DEPLACER DANS LES TP.

We can try fitting a VARMA model to the EU stock market dataset presented in Figure 13. For that we can use the `VARMA` function of the `MTS` package.

- `base::ar`
- `MTS::VAR, VMA, VARMA`

If we fit a VAR model to the European stock markets, AIC gives $p = 1$ and the estimated matrix of coefficients $\hat{\Phi}_1$ and their standard errors are

```
      > VAR.model <- VARMA(returns) ## first calculate returns from data
Constant term:
Estimates:  0.0007532118 0.000823424 0.0005525078 0.000473503

       DAX       SMI       CAC    FTSE
DAX   0.00497 -0.09642  0.03919 0.0483
SMI  -0.00738 -0.00682  0.03611 0.0674
CAC  -0.02784 -0.11229  0.06117 0.0939
FTSE -0.01084 -0.08922 -0.00395 0.1655
```

Here $n = 1860$ so $n^{-1/2} = 0.023$ helping us assess which of these coefficients differ significantly from zero:

- all react positively to a jump in the FTSE the previous day.
- all react negatively to a jump in the SMI the previous day.
- all are somewhat decoupled from the DAX.

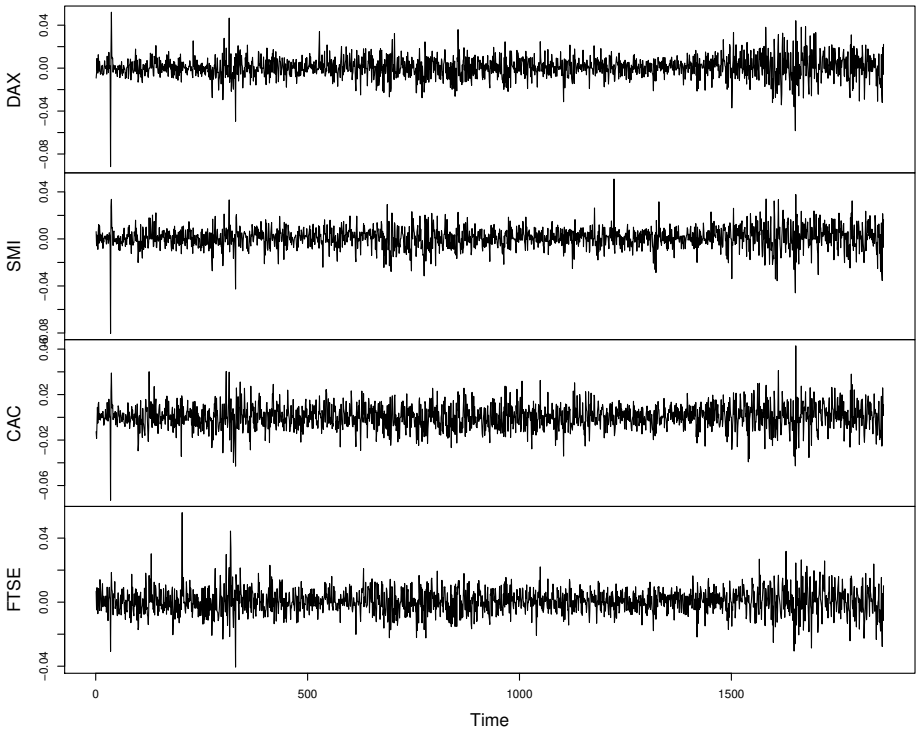


Figure 13: Several EU markets return

10 GARCH models

Many time series show changes in the variance as well as in the mean; this is particularly prominent in financial time series, but arises in many other contexts also. However (stationary, invertible, causal) *Gaussian* (V)ARMA models satisfy, if $(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^T & \Sigma_Y \end{pmatrix}\right)$ then $\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}_Y \Sigma_{XY}^T \Sigma_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_X), \Sigma_Y - \Sigma_{XY}^T \Sigma_X^{-1} \Sigma_{XY}\right)$,

$$Y_t \mid Y_{-t} = y_{-t} \sim \mathcal{N}(\mu + \Sigma_{t,-t}\, \Sigma_{-t}^{-1}\, (y_{-t} - \mu), \Sigma_{t,t} - \Sigma_{t,-t}\, \Sigma_{-t}^{-1}\, \Sigma_{-t,t}),$$

in an obvious notation: the conditional variance of Y_t given the preceding observations Y_{-t} does not depend on their values, and so does not vary with time.

This prompts the search for models that do allow such dependence. A prominent class of such models is the **generalised autoreressive conditionally heteroscedastic (GARCH)** class, given by

$$Y_t = \sigma_t\, \varepsilon_t, \quad \sigma_t^2 = \alpha_0 + \sum_{j=1}^{m} \alpha_j Y_{t-j}^2 + \sum_{j=1}^{r} \beta_j \sigma_{t-j}^2, \quad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

and labelled GARCH(m, r). To avoid negative variances we take $\alpha_j, \beta_j \geq 0$.

A GARCH model is not unlike an ARMA model in the variances σ_t^2, and the fitting approach is similar. If $\beta_1 = \cdots = \beta_r = 0$, we have an ARCH$(m)$ model.

Typically in practice m, r are small: often we take $m = r = 1$

Definition 98. For $t \in \mathbb{Z}$, let $\mathcal{H}_t$ denote the entire history of the process $\{Y_t\}$ up to time t.

Lemma 99. *A GARCH model has zero mean, and satisfies $E(Y_t \mid \mathcal{H}_{t-1}) = 0$; thus it is an uncorrelated sequence: $cor(Y_{t+h}, Y_t) = 0$. The quantity Y_t is called a **martingale difference**.*

Lemma 100. *An ARCH(1) model with standard Gaussian innovations, $\{\varepsilon_t\} \overset{iid}{\sim} \mathcal{N}(0,1)$, may be written*

$$Y_t^2 = \alpha_0 + \alpha_1\, Y_{t-1}^2 + V_t,$$

where $V_t = \sigma_t^2(\varepsilon_t^2 - 1)$, so the $\{V_t\}$ are non-Gaussian white noise. Thus the process is stationary and causal if $0 \leq \alpha_1 < 1$. In this case we have $var(Y_t) = \alpha_0/(1 - \alpha_1)$ and if in addition $3\alpha_1^2 < 1$,

$$\frac{E(Y_t^4)}{var(Y_t)^2} = 3\, \frac{1 - \alpha_1^2}{1 - 3\alpha_1^2} \geq 3.$$

The implication of this last result is that the marginal distribution of the ARCH(1) model has heavier tails than does the normal.

10.1 (G)ARCH theory

For the ARCH(1) model:

- If $0 \leq \alpha_1 < 1$, then $\{Y_t\}$ is white noise and its unconditional distribution is symmetrically distributed around zero. This unconditional distribution is leptokurtic (has heavier tails than the normal).

- If in addition $3\alpha_1^2 < 1$, then the process follows $\{Y_t^2\}$ follows a causal AR(1) model with ACF $\rho_h = \alpha_1^{|h|}$.

– If $3\alpha_1^2 \geq 1$ but $\alpha_1 < 1$, then $\{Y_t^2\}$ is strictly stationary with infinite variance.

– We have already seen how to estimate the parameters α_0, α_1 using—the Markov property, and—least squares.

Likewise the GARCH(1,1) has an ARMA(1,1) representation

$$Y_t^2 = \alpha_0 + (\alpha_1 + \beta_1)\, Y_{t-1}^2 + V_t - \beta_1\, V_{t-1},$$

where V_t is defined as before.

Estimation for GARCH(m, r) is performed by supposing that $\sigma_1^2 = \cdots = \sigma_r^2 = 0$, and building the likelihood for $Y_{r+1}, \ldots, Y_t$.

Example 101 (FTSE)**.** *The Financial Times Stock Exchange Index, 1991–1998.*

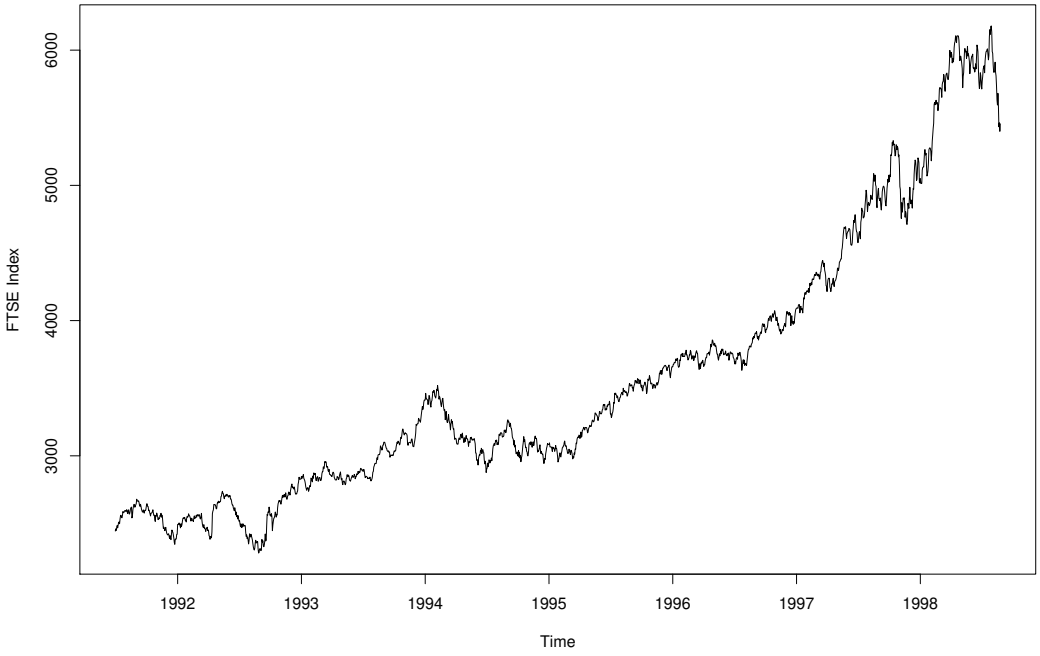


Figure 14: FTSE value

Remark 102. *(G)ARCH models have the following weaknesses*

– *financial markets react more strongly to negative shocks, but in a GARCH model σ_t reacts identically to positive and negative shocks;*

– *they react slowly to large shocks in the return, so may over-predict volatility;*

– *strong restrictions on the parameters are needed for stationarity and finite variance;*

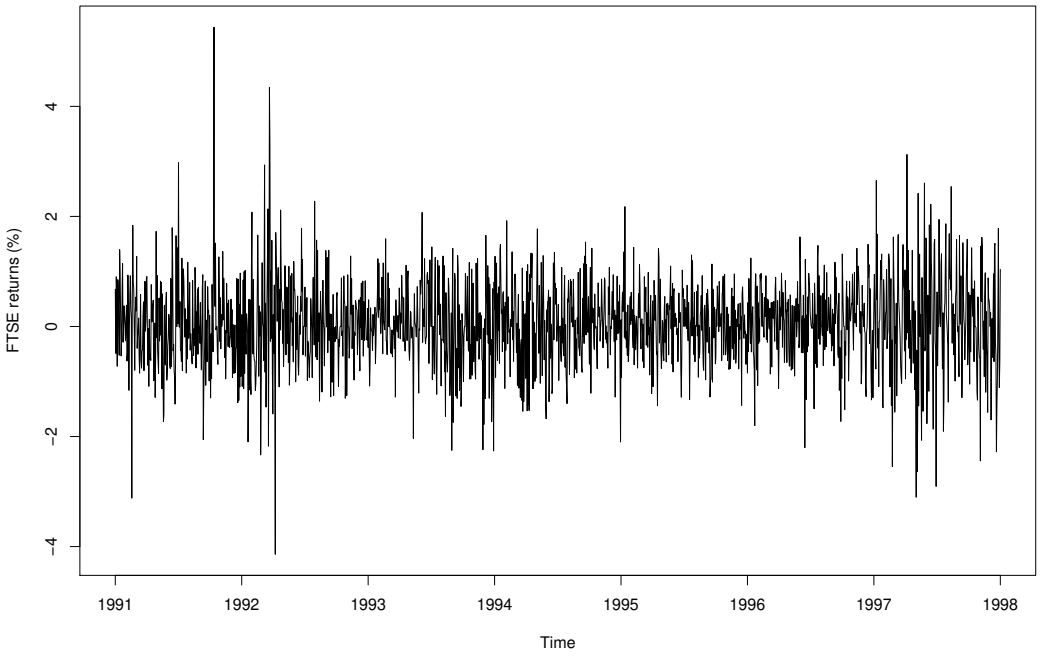


Figure 15: FTSE return

– *they may describe the second-order properties, but they are not based on a financial/economic theory so give no insight into why a series behaves as it does;*

10.2 Some extensions

Exponential GARCH (EGARCH) The EGARCH model allows the variance to depend on the sign of the series, for example giving

$$\sigma_t^2 = \sigma_{t-1}^{2\alpha} \exp\left[\alpha_* + \left\{\gamma + \theta \,\mathrm{sign}(Y_{t-1}) \frac{|Y_{t-1}|}{\sigma_{t-1}}\right\}\right]$$

for suitable constants $\alpha, \alpha_*, \gamma, \theta$; we expect that $\theta < 0$ if negative shocks have higher impacts.

Integrated GARCH GARCH models have been extended to *integrated GARCH (IGARCH)* models, e.g.,

$$\sigma_t^2 = \alpha_0 + \beta_1 \sigma_{t-1}^2 + (1 - \beta_1)\, Y_{t-1}^2,$$

which is analogous to an ARIMA model, in that past volatility shocks persist.

11 Stochastic Volatility models

A useful division of time series models is into *observation-driven* models and *parameter-driven* models.

The difference is more clear on the following example:

$$Y_t \mid Z_t \sim \mathcal{N}\left(\mu_t, \sigma_t^2\right), \qquad t \in \mathbb{Z}.$$

The distribution of $Y_t \mid Z_t$ is normal with mean μ_t and variance σ_t^2 both functions of Z_t.

Observation-driven models These models take Z_t to be a function of past observations. Examples:

– AR(p) model with $\mu_t = \phi_1\, y_{t-1} + \cdots + \phi_p\, y_{t-p}, \quad \sigma^2 = \sigma^2$;

– ARCH(m) model with $\mu_t = 0, \quad \sigma_t^2 = \alpha_0 + \alpha_1\, y_{t-1}^2 + \cdots + \alpha_p\, y_{t-m}^2$;

– AR(1)-ARCH(1) model, with $\mu_t = \mu + \phi(y_{t-1} - \mu), \quad \sigma_t^2 = \alpha_0 + \alpha_1 (y_{t-1} - \mu)^2$.

These models are particularly adapted to the representation of volatility:

– the likelihood is easily computed, so estimation, testing, model-checking are easy;

– finance theory is often specified through one-step-ahead movements, defined with respect to economic agents' information;

– they parallel the very successful ARMA-type models.

Parameter driven models These models take Z_t to be a function of some unobserved or latent component, e.g., with

$$\mu_t = 0, \qquad \log \sigma_t^2 = \gamma_0 + \gamma_1 \log \sigma_{t-1}^2 + \eta_t \qquad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\eta^2),$$

so that the unobserved log-volatility process $\log \sigma_t^2$ satisfies an AR(1) model

These models have good properties as well:

- properties are easier to find, manipulate and generalize to the vector case;
- have simpler representations in continuous time (e.g., as diffusions that satisfy suitable SDEs), which makes them natural in finance where much theory is based on stochastic differential equations;

but parameter-driven models don't have simple forms for their density $f(y_t \mid \mathcal{H}_t)$, so can't easily write down likelihood ... so inference becomes more difficult. Fore more details on these models see Kim, Shephard and Chib (1998), *Review of Economic Studies*, 351–393.

12 Simple stochastic volatility model

Consider

$$Y_t = \sigma_t \, \varepsilon_t, \qquad h_t = \log \sigma_t^2, \qquad h_{t+1} = \gamma_0 + \gamma_1 \, (h_t - \gamma_0) + \sigma_\eta \, \eta_t$$

where $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ independent of $\eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

Thus $\{h_t\} \sim$ AR(1), and if $|\gamma_1| < 1$, then we have stationary distribution

$$h_t \sim \mathcal{N}\left(\gamma_0, \frac{\sigma_\eta^2}{1-\gamma_1^2}\right),$$

with long excursions above or below γ_0 if $\gamma_1 \approx 1$. (why?)

Inference on the model Inference for this model has close link to state space models, with the difference that likelihood inference can be replaced by *quasi-likelihood* inference using the relation $\log Y_t^2 = h_t + \varepsilon_t^2$, where (with ε_t normal), we have $E(\log \varepsilon_t^2) = -1.27$ and $\text{var}(\log \varepsilon_t^2) = 4.93$, but this doesn't work well. It is also possible to use Bayesian inference and Markov chain Monte Carlo (MCMC) to estimate the time series h.

We will now explain how explain MCMC and Bayesian inference in the context of an AR(1) model with measurement error, before returning to stochastic volatility modelling.

Bayesian inference *Bayesian inference* requires that a prior density be specified for every unknown parameter in a statistical problem, and then just applies the laws of probability to obtain the **posterior density** for all unknowns, conditional on the observed data.

Given the data model $f(y \mid \theta) \equiv f(y; \theta)$ and prior density $\pi(\theta)$ for a parameter $\theta_{m \times 1}$, Bayes' theorem posterior density

$$\pi(\theta \mid y) = \frac{\pi(\theta) \, f(y \mid \theta)}{\int \pi(\theta) \, f(y \mid \theta) \, \mathrm{d}\theta}$$

which contains all information about θ, conditional on the observed data.

We often use the *Markov chain Monte Carlo* simulation to generate samples from $\pi(\theta \mid \theta_{-i})$

We presend in Algorithm 1 the Gibbs sampler. Under suitable conditions and for large R, the distribution of the sequence $\theta^{(1)}, \ldots, \theta^{(R)}$ approximates $\pi(\theta \mid y)$, so we can use the $\theta^{(r)}$ to estimate properties of the posterior density (often dropping the initial transient part of the sequence—knowns as **burn-in.**)

Ket $\theta^{(0)}$;
for $r = 1, \ldots, R$ **do**
  set $\theta^{(r)} = \theta^{(r-1)}$,
  for $i = 1, \ldots, m$, generate $\theta_i^* \sim \pi(\theta_i \mid \theta_{-i}^{(r)})$ and replace $\theta_i^{(r)}$ by θ_i^*
  Let θ_{-i} denote the elements of θ without θ_i, for $i = 1, \ldots, m$. Suppose that it is possible to sample from the distributions $\pi(\theta \mid \theta_{-i})$, for $i = 1, \ldots, m$.

Algorithm 1: Gibbs sampler

Example 103 (Example: AR(1) with measurement error)**.** – *Suppose* $\{\eta_t\}$ *follows an AR(1) process with parameter* α *and innovation variance* τ^2*, and* $Y_t \mid \eta_1, \ldots, \eta_n \overset{ind}{\sim} \mathcal{N}(\eta_t, \sigma^2)$*, giving an* $AR(1)$ *with measurement error.*

– *Here* $\theta = (\alpha, \tau^2, \sigma^2, \eta)$ *has dimension* $m = n + 3$.

– *For simplicity we take independent prior densities*

$$\eta_1 \sim \mathcal{N}(0, \alpha^2), \quad \alpha \sim \mathcal{N}(0, b^2), \quad \tau^2 \sim IG(c, d), \quad \sigma^2 \sim IG(e, f)$$

where $IG(c, d)$ *denotes the inverse gamma density* $d^c x^{-c-1} e^{-d/x}/\Gamma(c)$*, for* $x > 0$ *and* $c, d > 0$.

The full conditional densities $\pi(\theta_i \mid \theta_{-i})$ required to run a Gibbs sampler are:

$$\alpha \mid \text{rest} \sim \mathcal{N}(B_\alpha/A_\alpha, 1/A_\alpha), \quad A_\alpha = \frac{1}{b^2} + \frac{1}{\tau^2}\sum_{t=1}^{n-1} \eta_t^2, \quad B_\alpha = \frac{1}{\tau^2}\sum_{t=2}^{n} \eta_t\, \eta_{t-1},$$

$$\tau^2 \mid \text{rest} \sim IG\left\{ c + \frac{n-1}{2}, d + \frac{1}{2}\sum_{t=2}^{n} (\eta_t - \alpha\, \eta_{t-1})^2 \right\},$$

$$\sigma^2 \mid \text{rest} \sim IG\left\{ e + \frac{n}{2}, f + \frac{1}{2}\sum_{t=2}^{n} (y_t - \eta_t)^2 \right\},$$

$$\eta_1 \mid \text{rest} \sim \mathcal{N}(B_1/A_1, 1/A_1), \quad A_1 = \frac{1}{a^2} + \frac{\alpha^2}{\tau^2} + \frac{1}{\sigma^2}, \quad B_1 = \frac{\alpha\eta_2}{\tau^2} + \frac{y_1}{\sigma^2}, \sum_{t=1}^{n-1} \eta_t^2,$$

$$\eta_t \mid \text{rest} \sim \mathcal{N}(B_t/A_t, 1/A_t), \quad A_t = \frac{1}{\sigma^2} + \frac{1+\alpha^2}{\tau^2}, \quad B_t = \frac{\alpha(\eta_{t-1} + \eta_{t+1})\eta_2}{\tau^2} + \frac{y_1}{\sigma^2}, \quad t \neq 1, n,$$

$$\eta_n \mid \text{rest} \sim \mathcal{N}(B_t/A_t, 1/A_t), \quad A_n = \frac{1}{\tau^2} + \frac{1}{\sigma^2}, \quad B_n = \frac{\alpha\eta_{n-1}}{\tau^2} + \frac{y_n}{\sigma^2}.$$

Figure 16 presents the results when this is run for $R = 10,000$ iterations, preceded by a 'burn-in' of length 200. I took $a = 100$, $b = 10$, $c = d = e = f = 0.01$, to give very vague prior information. The `R` code for the entire examples, including pictures, is 60 lines.

Remark 104. *Often the code can be accelerated by updating blocks of (conditionally) independent variables. In the example above, the most time-consuming part is the iteration over* $\eta_1, \ldots, \eta_n$*. These*

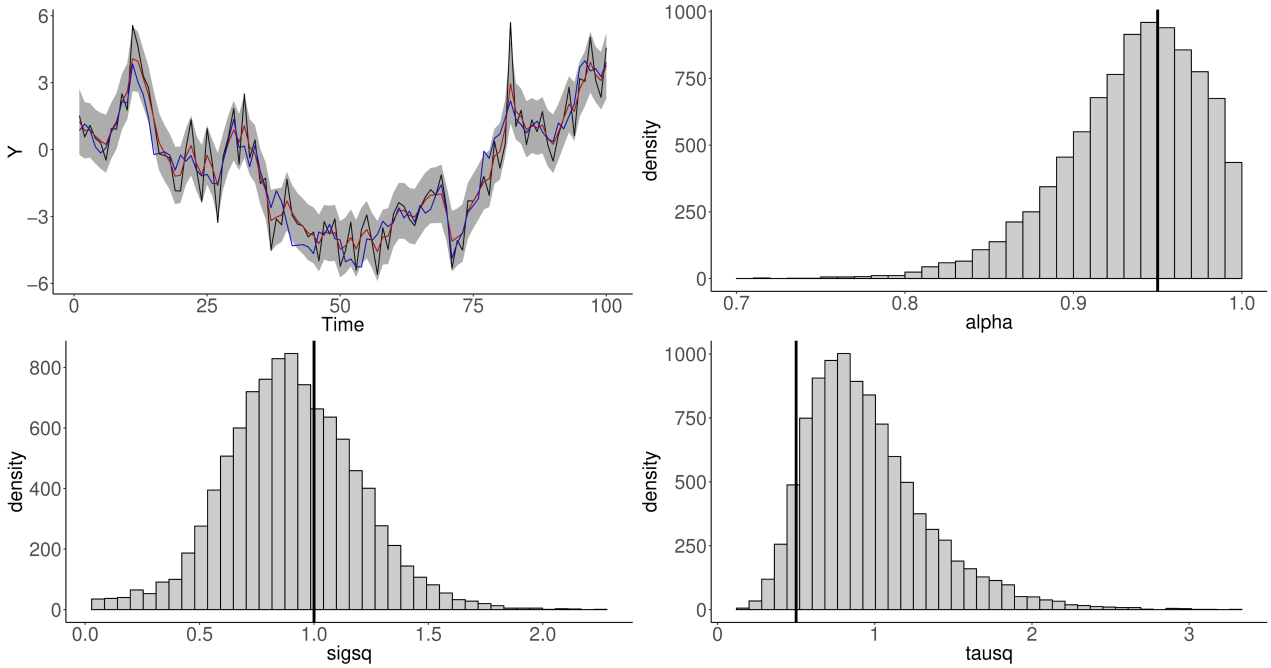


Figure 16: Top left: the **red** line is the posterior mean for η, the **grey** band are 95% pointwise credible intervals for η, the **blue** line is the true η. In the other panels the true value used is shown by a vertical line.

have a multivariate normal distribution conditional on θ, and updating them simultaneously would involve inverting an $n \times n$ matrix, but we could avoid this by updating $\eta_1, \eta_3, \ldots$, conditional on $\eta_2, \eta_4, \ldots$, and then vice verse (both conditional on θ), noting that the odd ηs are conditionally independent given the even ones, and vice versa. This replaces a loop over $\eta_1, \ldots, \eta_n$ with two steps, making the algorithm much faster.

MCMC algorithms can be developed for many models, though usually some full conditional densities $\pi(\theta_i \mid \theta_{-i})$ will be unavailable in a simple form and Metropolis-Hastings steps must be used.

*Reversible jump MCMC algorithms (Green, 1995. Biometrika, **82**, 711–732) can be used when the number of underlying states is unknown.*

This is part of a very large literature on numerical methods for inference on these models.

MCMC for SV The Gibbs sampler in this case might be, although Many variant algorithms and procedures exist; this is just the simplest—see Kim *et al.* (1998).:

- Initialise $h = (h_1, \ldots, h_n)$ and $\theta = (\gamma_0, \gamma_1, \sigma_\eta^2)$.
- For $r = 1, \ldots, R$, repeat
  1. For $t = 1, \ldots, n$, sample $h_t \sim \pi(h_t \mid h_{-t}, \gamma_0, \gamma_1, \sigma_\eta^2, y)$
  2. Sample $\sigma_\eta^2 \sim \pi(h, \gamma_0, \gamma_1, , y)$
  3. Sample $\gamma_0 \sim \pi(h, \gamma_1, \sigma_\eta^2, y)$
  4. Sample $\gamma_1 \sim \pi(h, \gamma_0, \sigma_\eta^2, y)$

A complete sweep involves cycling once through steps 1-4, and we must take $R = O(10^3)$ at least, perhaps $R = O(10^6)$, so efficient coding and fast algorithms are essential.

Computing the conditional densities here may be hard, or even impossible, but then we can use Metropolis-Hastings steps, require only the densities up to a constant of proportionality.

Once we have the output, we have to check convergence of the algorithm, and if satisfied we then use the output to estimate the posterior density $\pi(\theta \mid y)$.

Discussion Stochastic volatility modelling is an attractive approach to accounting for changing variance, because we can allow the hidden process σ^2 to vary according to economic/financial theory, including the role of exogenous variables. Nevertheless it involves more sophisticated approaches than we have used so far, based on Markov chain Monte Carlo, and these may be hard to implement. See Kim *et al.* (1998) and subsequent papers for more details, developments, etc.

13 Factor models

Popular models for financial time series posit that returns for many series react to changes in a few underlying factors, which themselves evolve according to stationary time series. To formalise this, suppose we have k assets and n time periods, and let the return on asset i in time period t be

$$r_{it} = \alpha_i + \beta_{i1} f_{1t} + \cdots + \beta_{im} f_{mt} + \varepsilon_{it}, \qquad t = 1, \ldots, n, \quad i = 1, \ldots, k,$$

where α_i is a constant expressing the intercept, f_{jt} for $j = 1, \ldots, m$ are m common factors, β_{ij} is the factor loading for asset i on factor j, and ε_{it} is the specific factor of asset i. We hope that $m \ll k$; but in some cases have $k \gg n$.

Assume that $(f_t)_{m\times 1}$ is a stationary process satisfying $E(f_t) = \mu_f$, $\text{var}(f_t) = \Sigma_f$, and that the (ε_{it}) are uncorrelated white noise series, but with $\text{var}(\varepsilon_{it}) = \sigma_i^2$. We can write

$$r_t = \alpha_{k\times 1} + \beta_{k\times m}\, f_t + \varepsilon_t, \qquad t = 1, \ldots, n,$$

in an obvious notation, and then have $\text{var}(r_t) = \beta\, \Sigma_f \beta^T + \text{diag}(\sigma_1^2, \ldots, \sigma_k^2)$.

It is common to assume that the factors are serially uncorrelated. These models are widely used but suffer from the curse of dimensionality: the number of parameters quickly becomes very large, so symmetries must be exploited to reduce the number of parameters — see *Tsay (2005), Analysis of Financial Time Series, Second edition, Wiley.*

Part IV
Frequency Analysis, Fourier

We know that autocovariance functions are symmetric positive, this comes from the covariance structure, consequence of the stationarity. In this section we will discuss an interesting link between Fourier coefficients of positive even finite measures. This allows to use spectral theory methods in Time series analysis.

This will also be of interest to understand the notion of "white noise" and filtering, that are common in signal theory.

14 Spectral measure of a Stationary Process

We need to state the following theorem, without proof first, it states that the covariance functions are Fourier transform of some measures.

Theorem 105 (Hergoltz — spectral measure and spectral density)**.** *For all $\gamma : \mathbb{Z} \to \mathbb{R}$, the two following properties are equivalent:*

1. *γ is symmetric, positive:*

$$\forall h \in \mathbb{Z},\ \gamma(-h) = \gamma(h),\ \text{and } \forall n \geq 1, \forall v \in \mathbb{R}^n,\ \sum_{j,k=1}^{n} v_j v_k \gamma(j-j) \geq 0.$$

2. *$((\gamma(h))_{h\in\mathbb{Z}}$ are the Fourier coefficients of a positive finite measure ν on $[-\pi, \pi]$. In particular, ν is even and $\gamma(0) = \nu([-\pi, \pi])$ as:*

$$\forall h \in \mathbb{Z}, \gamma(h) = \int_{[0,1]} e^{-2\pi i h u} \nu(du) \in \mathbb{R}.$$

When any of these properties is true, the measure ν is unique and is called spectral measure*, its density f if it exists is called the spectral density.*

Furthermore, when any of these properties is true and if $\gamma \in \ell^1(\mathbb{Z})$, then the spectral density can be written as:

$$\forall u \in [0, 1],\ f(u) = \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} \gamma(h) e^{-2i\pi h u},$$

and this function is even and continuous.

Remark 106. *In our cases, we will only work with the densities, and the part about measures can be forgotten.*

Example 107 (Spectral density of an MA(1))**.** *Let X be a $MA(1)$ rocess defined for $w \sim (0, \sigma^2)$ by $X_t = \theta X_{t-1} + w_t$. Then its covariance is $\gamma_X = \sigma^2(1+\theta^2)\mathbf{1}_{h=0} + \sigma^2\theta\mathbf{1}_{h=\pm 1} \in \ell^1(\mathbb{Z})$, and its spectral density is then:*

$$\forall u \in [0,1], \quad f(u) = \frac{1}{2\pi} \sum_{h=-1,0,1} \gamma_X(h) e^{-2i\pi hu} = \frac{\sigma^2}{2\pi}(1 + \theta^2 + 2\theta\cos(2\pi u)).$$

Definition 108 (Spectral density of a stationary process)**.** Let $X = (X_t)_{t\in\mathbb{Z}}$ be a stationary process with autocovariance function γ_X then the spectral measure of X, noted ν_X is the positive measure whose Fourier transform is γ_X as defined in Theorem 105. If $\gamma_X \in \ell^1(\mathbb{Z})$ then the spectral density of X, noted f_X is the density of ν_X.

Example 109 (White noise spectral density)**.** *Let $w \sim (0, \sigma^2)$, then $\gamma_w(h) = \sigma^2\mathbf{1}_{h=0}$ and thus, w has a spectral density:*

$$f_w(u) = \frac{\sigma^2}{2\pi}.$$

From there, we see why the term "white noise" has been chosen: every frequency of the spectrum appears with equal weight.

The question is now to use these results in practice. Firstly, we can note that filtering of a process that has a spectral density also has a spectrald density.

Theorem 110 (Spectre and filtering)**.** *Let X be a stationary process with spectral density f_X and let $\alpha \in \ell^1(\mathbb{Z})$, then $F_\alpha X = P_\alpha(B)X = Y$ also has a spectral density that writes, with $P_\alpha(B) = \sum_{j\in\mathbb{Z}} \alpha_j B^j$:*

$$\forall u \in [0,1], \ f_Y(u) = |P_\alpha(e^{-2i\pi u})|^2 f_X(u).$$

Example 111 (Spectral density of ARMA process)**.** *Let X be a regular invertible causal $ARMA(p,q)$ process, verifying*

$$\Phi(B)X = \Theta(B)w, \quad w \sim (0, \sigma^2),$$

with $\Phi(B) = 1 - (\phi_1 B + \cdots + \phi_p B^p)$, and $\Theta(B) = 1 + \theta B + \cdots + \theta_q B^q$, such that $\Theta(B)/\Phi(B) = \sum_k \psi_j B^j$ is the Wold decomposition. Then the spectral density of the ARMA process, using the result of Example 109, is

$$\forall u \in [0,1], \quad f_X(u) = \frac{\sigma^2|\Theta(e^{-2i\pi u})|^2}{2\pi|\Phi(e^{-2i\pi u})|^2}.$$

In some cases, this function is more usable in practice than the covariance formula.

14.1 Periodogram

A natural way to use the previous concepts is to compute the empiriral spectral density associated with the observations of a time series.

Definition 112 (Periodogram)**.** If a time series $x_1, \ldots, x_n$ is an equally spaced time series, its *periodogram ordinate* is the function

$$\forall 0 < \omega < 1/2,\ I(\omega) = n^{-1}\left[\left(\sum_{t=1}^{n} y_t \sin(2\pi\omega t)\right)^2 + \left(\sum_{t=1}^{n} y_t \cos(2\pi\omega t)\right)^2\right]$$

1. The periodogram is the plot of $I(j/n)$ representing the Fourier frequencies $2\pi j/n$, $j = 1, \ldots, m = \lceil (n-1)/2 \rceil$, $I(1/2)$ is included only if n is even. By default, `R` plots the *log-periodogram* $log(I)$.
2. The *cumulative periodogram* is a plot of the C_j against the frequencies j/n where

$$C_r = \frac{\sum_{j=1}^{r} I(j/n)}{\sum_{\ell=1}^{m} I(\ell/n)}.$$

15 Frequency analysis

Example 113 (Sine wave with white noise)**.** *First, we can have a look at the following example, where the time series is a sine with white noise. In Figure 17 the noise has a variance* 0.25 *and frequency* 1/200*. In Figure 18 the noise has variance* 1 *and freqiency* 1/20*.*

Example 114 (Periodogram for an AR(1) process)**.** *Figures 19 and 20 represent two periodograms for AR processes with different parameters.*

15.1 Periodogram in practice, example: Litenizing hormon

We represent in Figure 21 the dataset (top left), the periodogram (top right), the possible Fourier series that are the base for the decomposition (bottom left) and the cumulative periodogram (bottom right).

Low frequency variation (trend) appears at the left of the periodogram, and high frequency variation (rapid oscillations) appears at the right. The rationale for considering only the frequencies $\omega = 2\pi j/n$ is that

$$\sum_{t=1}^{n} y_t^2 = I(0) + 2\sum_{j=1}^{m} I(j/n) + I(1/2), \tag{6}$$

with $I(1/2)$ included only if n is even. Thus the periodogram decomposes the total variability $\sum y_t^2$ of the data into components associated with each of these frequencies, plus one for the grand mean $I(0) = n\bar{y}^2$ which we ignore because it is not periodic.

The rationale for plotting the log periodogram is that the periodogram ordinates are (roughly) exponentially distributed, and the log-transformation is variance-stabilising for the exponential distribution. A rough significance scale for the log-periodogram is shown by the vertical line on its right.

The cumulative periodogram provides a visual test of whether the series white noise. We compare C_r with its expected value r/m: a large value of the *Kolmogorov-Smirnov statistic* $D = \max|C_r - r/m|$ suggests that the underlying series is not white noise. The test involves seeing whether the cumulative periodogram falls outside a diagonal band, whose width determines the size of the test.

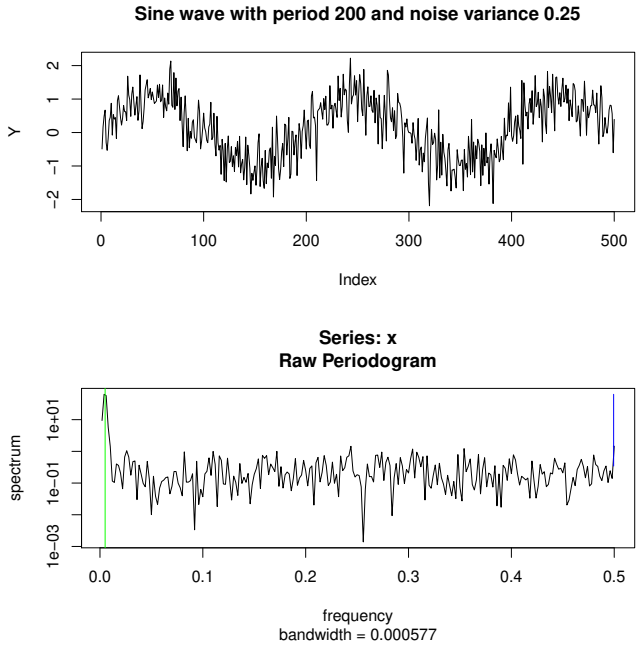


Figure 17: Top: data from a simulated sine wave with added white noise. Bottom: log periodogram with red horizontal line showing noise variance $\sigma^2 = 0.25$, and a green vertical line showing the signal frequency 1/200. The blue line shows the width of a 95% confidence interval for the true value at each point.

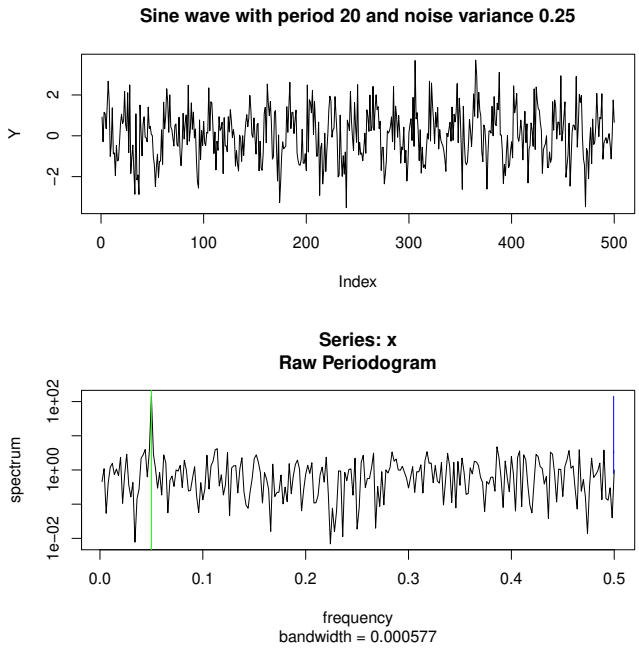


Figure 18: Top: data from a simulated sine wave with added white noise. Bottom: log periodogram with red horizontal line showing noise variance $\sigma^2 = 1$, and a green vertical line showing the signal frequency 1/20. The blue line shows the width of a 95% confidence interval for the true value at each point.

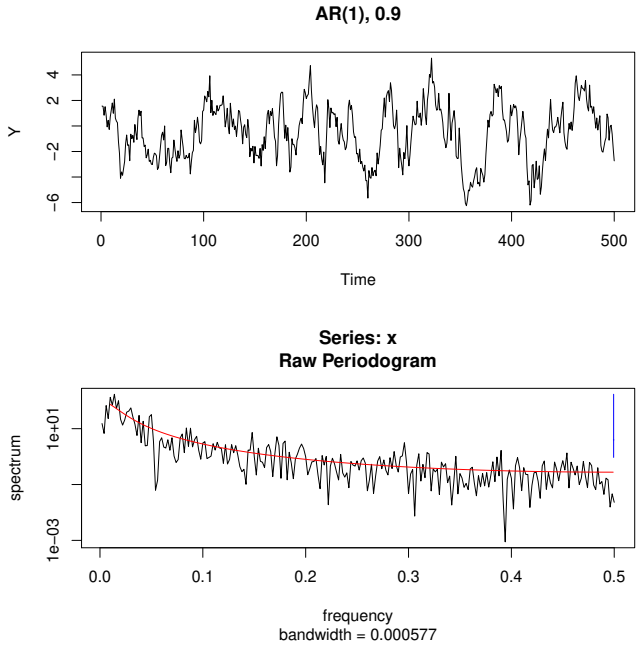


Figure 19: Data from a simulated AR(1) model with parameter 0.9, with log-periodogram and theoretical value (in red). The blue line shows the width of a 95% confidence interval for the true value at each point. The log scale on the vertical axis means there is a very large change in the periodogram itself.

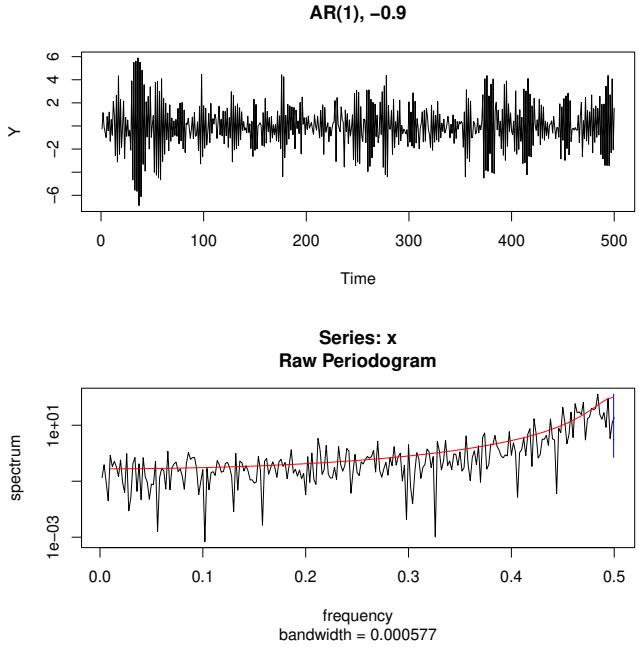


Figure 20: Data from a simulated AR(1) model with parameter -0.9, with log-periodogram and theoretical value (in red).

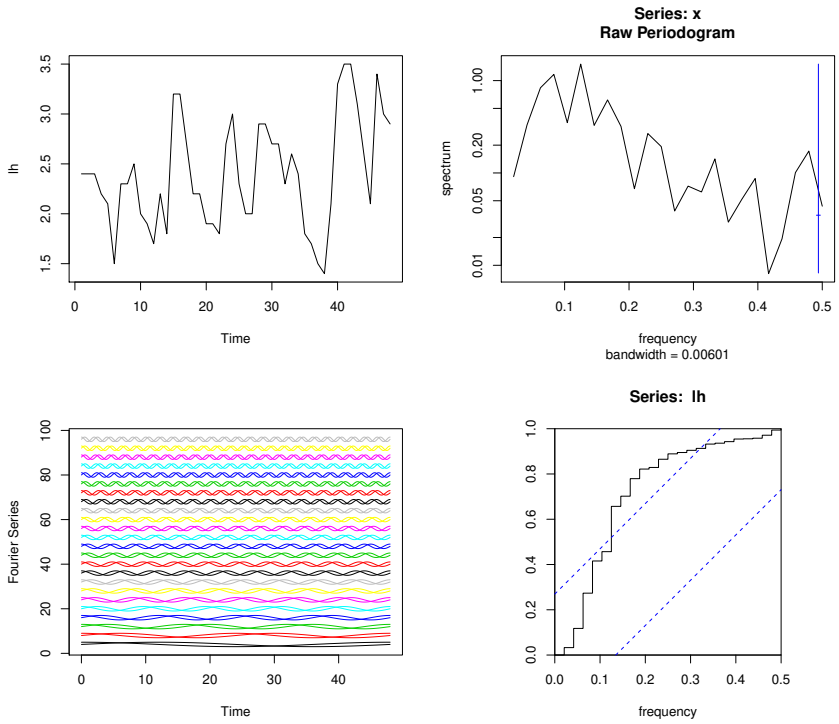


Figure 21: Lutenizing hormon in $n = 48$ successive blood samples from a woman, taken at 10 min intervals, and associated periodograms.

15.1.1 Properties of the periodogram

Theorem 115. *If* $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$*, then all the periodogram ordinates are independent, and*

(a) $I(1/n), \ldots, I(m/n)$ *are exponential random variables with mean* σ^2*;*

(b) if n *is even,* $I(1/2) \sim \sigma^2 \chi^2_1$*; finally,*

(c) the cumulative periodogram ordinate

$$C_r = \frac{\sum_{j=1}^{r} I(j/n)}{\sum_{l=1}^{m} I(l/n)}, \quad r = 1, \ldots, m,$$

has a beta$(r, m-r)$ *distribution, and so has mean and variance* r/m*,* $r(m-r)/m$

Part (a) adds that if a white noise is Gaussian, we get the independence independence.

15.2 Frequency domain analysis

Here are some referencest to frequency domain analysis

– Excellent introduction: *Bloomfield (2000) Fourier Analysis of Time Series, Second edition, Wiley*

– The bible is *Priestley (1981) Spectral Analysis and Time Series, Volumes 1 and 2, Academic Press* which also deals with spectra for non-stationary series.

– See also *Persival and Walden (1993) Spectral Analysis for Physical Applications: Multitape and Conventional Univariate Techniques, CUP*

– More modern approach is use of *wavelets*, which decompose series in orthogonal functions that are local in time: see *Percival and Walden (2000) Wavelet Methods for Time Series Analysis, CUP*.

Part V
State Space models

State space models, also called *structural models*, provide a very general approach to time series modelling. They originated in the 1960s in the control theory/electrical engineering literature, then were taken up for economic modelling, and are now used in many application areas. They allow maximum likelihood estimation for the model parameters and easy treatment of missing data

The basic model presupposes that the observed data follow a normal linear model that depends upon an unobservable underlying state, so the whole theory hinges upon computations for the multivariate normal distribution.

The theory extends to much wider classes of models, often using Monte Carlo methods known as *particle filters*, which are a very active topic of research, with applications in finance, genetics, atmospheric science, etc.

16 Basic ideas

An unseen *state process* (μ_t) follows a Markov model, so the distribution of μ_{t+1} depends only on μ_t. The *observed process* Y_t is such that Y_t depends only on μ_t. Hence we have

$$\begin{aligned} \text{State equation :} \quad & \mu_t \sim K(\mu_t \mid \mu_{t-1}) \\ \text{Observation equation :} \quad & y_t \sim g(y_t \mid \mu_t) \end{aligned}$$

where the kernel K determines the evolution of $\{\mu_t\}$ and g determines the observation process in terms of μ_t. In general K and g might vary with time t. We can write this :

Let $\pi_{s|t}$ denote the conditional density of the underlying state μ_s given observations $Y_1, \ldots, Y_t$.

– Just after time $t-1$, we have observed $Y_1, \ldots, Y_{t-1}$, and our information about the unobserved state of the system is summarised in $\pi_{t-1|t-1}$. At that point our density for μ_t is given by

$$\pi_{t|t-1}(\mu_t) = \int K(\mu_t \mid \mu_{t-1})\, \pi_{t-1|t-1}(\mu_{t-1}) d\mu_{t-1}.$$

– At time t we observe Y_t, and then can update $\pi_{t|t-1}$ to

$$\pi_{t|t}(\mu_t) = \frac{\pi_{t|t-1}(\mu_t)\, g(y_t \mid \mu_t)}{\int \pi_{t|t-1}(\mu_t)\, g(y_t \mid \mu_t) d\mu_t}.$$

– We repeat these **prediction** and **filtering (or data assimilation)** steps each time a new observation arrives.

– The steps can be performed explicitly in very few cases, so usually numerical approximations are needed.

– The normal model is an exception, since in that case both steps can be performed analytically.

An interesting case is when the observations and noises are normal, we will discuss that in the remainder.

16.0.1 Reminder: Multivariate normal distribution

Definition 116. An n-dimensional *multivariate normal* random variable $X = (X_1, \ldots, X_n)^T$ with mean $\mu_{n\times 1}$ and covariance matrix $\Sigma_{n\times n}$ has density

$$f(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \qquad x,\mu \in \mathbb{R}^n;$$

we write $X \sim \mathcal{N}_n(\mu,\Sigma)$. We assume that the distribution is not degenerate, in which case Σ is positive definite, then its determinant $|\Sigma| > 0$.

Lemma 117. *If $X \sim \mathcal{N}_n(\mu,\Sigma)$, $Y_1^T = (X_1,\ldots,X_q)$ and $Y_2^T = (X_{q+1},\ldots,X_n)$, and we write*

$$X = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}_n\left\{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right\}$$

then

– *the marginal distribution of Y_1 is $\mathcal{N}_q(\mu_1,\Sigma_{11})$*

– *the conditional distribution of Y_1 given $Y_2 = y_2$ is*

$$\mathcal{N}_q(\mu_1 + \Sigma_{12}\,\Sigma_{22}^{-1}\,(y_2-\mu_2), \Sigma_{11} - \Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{21})$$

16.1 Normal model for state space models

In order to study state space models, we will base ourselves on the following lemma. Normal distributions have nice properties, in particular here concerning their conditional distributions.

Lemma 118. *Let $X_{n\times 1}, Y_{m\times 1}$ and $Z_{p\times 1}$ have a joint multivariate normal distribution and suppose that their respective variance matrices $\Sigma_{XX}, \Sigma_{YY}, \Sigma_{ZZ}$ are nonsingular and that $cov(Y,Z) = \Sigma_{YZ} = 0$. Then in an obvious notation, we have*

(a) $E(X \mid Y = y) = \mu_X + \Sigma_{XY}\,\Sigma_{YY}^{-1}\,(y-\mu_Y)$,

(b) $var(X \mid Y = y) = \Sigma_{XX} - \Sigma_{XY}\,\Sigma_{YY}^{-1}\,\Sigma_{YX}$,

(c) $E(X \mid Y = y, Z = z) = E(X \mid Y = y) + \Sigma_{XZ}\,\Sigma_{ZZ}^{-1}\,(z-\mu_z)$

(d) $var(X \mid Y = y, Z = z) = var(X \mid Y = y) - \Sigma_{XZ}\,\Sigma_{ZZ}^{-1}\,\Sigma_{ZX}$

Given the previous lemma, we will be able to:

– given data up to time t, we predict the data at time $t+1$ using the conditional mean of Y_{t+1} given $Y_1, \ldots, Y_t$;

– if Y_s is missing, we replace it by its conditional mean given the observed data;

– to smooth noise out of the observed series, we consider the conditional mean of Y_s given all the data before and after it.

However the state space algorithms are designed for iterative computation and so look more complicated...

16.2 An example: Local trend model

The general derivation of the state space equations is algebraically messy, so we illustrate the ideas using a simple special case, the so-called *local trend model* (or *local level model*).

All linear state space models involve two equations, the *state equation*, which determines the evolution of an underlying unobserved state, and the *observation equation*, which determines how the observed data are related to the state.

The local trend model is defined by

$$\begin{aligned} \text{State equation :} \quad & \mu_{t+1} = \mu_t + \eta_t, \qquad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\eta^2) \\ \text{Observation equation :} \quad & y_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \end{aligned}$$

where the η_t and ε_t are all mutually independent.

Thus the state μ_t performs a random walk, but is observed only with error.

Figure 22 represents such a trajectory from this model.

17 Filtering, prediction and smoothing

The natural question after presenting these models is to propose inference methods and prediction routines.

Three problems are typically tackled:

– *filtering*: we estimate μ_t using $\mathcal{H}_t$

– *prediction*: we forecast μ_{t+h} for $h > 0$ using $\mathcal{H}_t$

– *smoothing*: we estimate μ_t using $\mathcal{H}_n$

to which we might add *identification* of the model structure, *estimation* of the unknown parameters, and *diagnosis* of model failure!

Analogy: you are trying to read bad handwriting on a blackboard during a lecture:

– *filtering* is deciphering the current word, using the lecture contents thus far;

– *prediction* is guessing what the next words will be, using the lecture contents thus far;

– *smoothing* is deciphering a word in the middle of the board, using the entire lecture.

We suppose that data $y_1, \ldots, y_n$ are available, let $\mathcal{H}_t = \{y_1, \ldots, y_n\}$ denote the information available at time t, and assume (for now) that the model, including the parameter values, is known We aim to make inference about the underlying states $\mu_1, \ldots, \mu_n$

In this section, we will note:

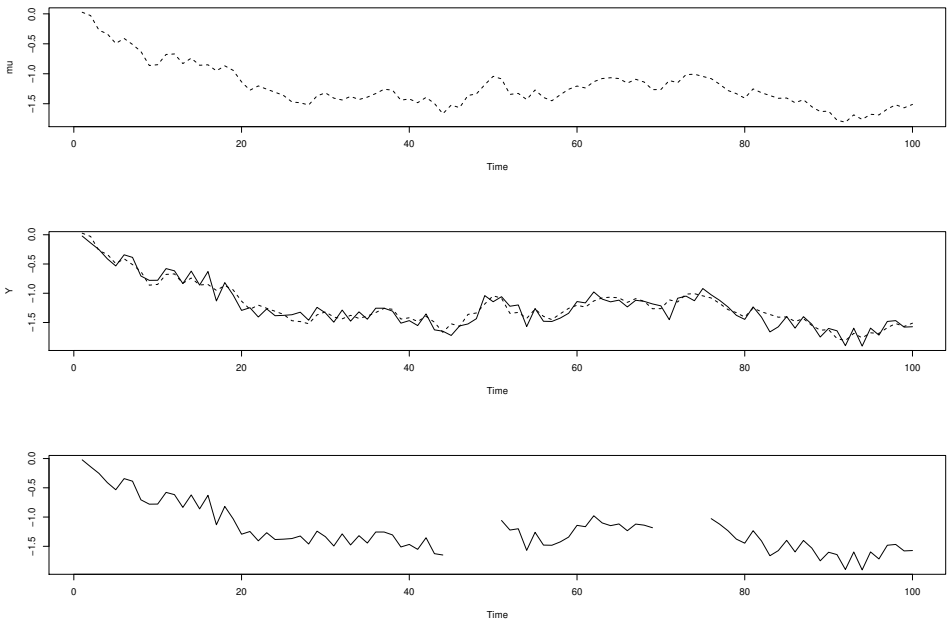


Figure 22: Upper panel: evolution of underlying state according to state equation; middle panel: evolution of state equation, with observed data; lower panel: observed data, with two sections missing.

– $\mu_{t|j}$ and $\Sigma_{t|j}$ denote the conditional mean and variance of μ_t given $\mathcal{H}_t$

– $y_{t|j}$ denote the conditional mean of y_t given $\mathcal{H}_t$

– $v_t = y_t - y_{t|t-1}$ and $V_t = \text{var}(v_t \mid \mathcal{H}_{t-1})$ be the one-step ahead forecast error and its variance given $\mathcal{H}_{t-1}$.

The observation equation gives

$$y_{t|t-1} = \text{E}(y_t \mid \mathcal{H}_{t-1}) = \text{E}(\mu_t + \varepsilon_t \mid \mathcal{H}_{t-1}) = \text{E}(\mu_t \mid \mathcal{H}_{t-1}) \qquad \text{, so}$$

$$v_t = y_t - y_{t|t-1} = y_t - \mu_{t|t-1}, \quad V_t = \text{var}(y_t - \mu_{t|t-1} \mid \mathcal{H}_{t-1}) = \cdots = \Sigma_{t|t-1} + \sigma^2,$$

and likewise conditioning on $\mathcal{H}_{t-1}$ shows that v_t is independent of $y_1, \ldots, y_{t-1}$, because

$$\text{E}(v_t) = 0, \quad \text{cov}(v_t, y_j) = 0, \quad j < t$$

Thus conditional on $\mathcal{H}_{t-1}$, knowing v_t is equivalent to knowing y_t, so we can write $\mathcal{H}_t = \{\mathcal{H}_{t-1}, y_t\} \equiv \{\mathcal{H}_{t-1}, v_t\}$.

The forecast error v_t is independent of $\mathcal{H}_{t-1}$, so $\text{var}(v_t) = \text{var}(v_t \mid \mathcal{H}_{t-1})$.

17.1 Kalman filter

In the normal case, we can introduce a (relatively) simple filter that allows us to predict observations.

Lemma 119. *For the local trend model, conditional on $\mathcal{H}_{t-1}$, we have*

$$\begin{pmatrix} \mu_t \\ v_t \end{pmatrix}_{\mathcal{H}_{t-1}} \sim \mathcal{N}_2 \left\{ \begin{pmatrix} \mu_{t|t-1} \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{t|t-1} & \Sigma_{t|t-1} \\ \Sigma_{t|t-1} & V_t \end{pmatrix} \right\}$$

and thus the conditional distribution of $\mu_t \mid \mathcal{H}_t$ is normal with mean and variance

$$\mu_{t|t} = \mu_{t|t-1} + \Sigma_{t|t-1} v_t / V_t = \mu_{t|t-1} + K_t \, v_t, \qquad \Sigma_{t|t} = \Sigma_{t|t-1} \, (1 - K_t)$$

where $K_t \in (0, 1)$ is called the Kalman gain. *Moreover, based on $\mathcal{H}_t$, we have*

$$\mu_{t+1|t} = \mu_{t|t}, \qquad \Sigma_{t+1|t} = \Sigma_{t|t} + \sigma_\eta^2.$$

The *Kalman filter* algorithm for the local trend model based on $y_1, \ldots, y_n$ entails assuming that $\mu_1 \sim \mathcal{N}(\mu_{1|0}, \Sigma_{1|0})$, and then for $t = 1, \ldots, n$ iterating the computations

$$\begin{aligned} v_t &= y_t - \mu_{t|t-1} \\ V_t &= \Sigma_{t|t-1} + \sigma^2 \\ K_t &= \Sigma_{t|t-1} / V_t \\ \mu_{t+1|t} &= \mu_{t|t-1} + K_t \, v_t \\ \Sigma_{t+1|t} &= \Sigma_{t|t-1} \, (1 - K_t) + \sigma_\eta^2 \end{aligned} \tag{7}$$

We take arbitrary initial values of 0 and 1 for $\mu_{1|0}$ and $\Sigma_{1|0}$, and need guesses for σ^2 and σ_η^2 For the filter, we then iterate forward:

t	y_t	$\mu_{t\|t-1}$	v_t	$\Sigma_{t\|t-1}$	V_t	K_t
1	y_1	0		1		
2	y_2					
3	y_3					
$\vdots$	$\vdots$					
$n-1$	y_{n-1}					
n	y_n					

Example 120 (Nile data)**.** *To explore this filter, we propose as example the* `Nile` *dataset from the* `datasets` *package of* `R`*. This dataset contains measurements of the annual flow of the river Nile at Aswan (formerly 'Assuan'), 1871-1970, in* 10^8 m^3*.. Figure 23 represents the data and filtered estimates associated.*

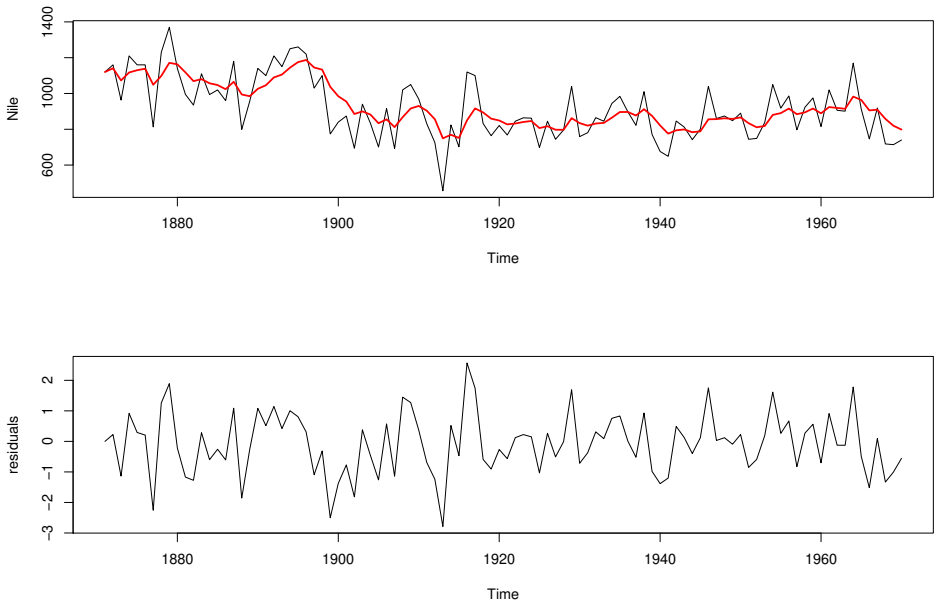


Figure 23: Nile Data. Upper panel: data and filtered estimate of stated variable (red). Bottom panel: residuals

You can observe in Fig 24 the time series diagnostic for the fit of the state space model.

17.1.1 Forecasting errors

Given the initial values $\Sigma_{1|0}$ and $\mu_{1|0}$, which are independent of the data, we use the data to write *one-step ahead forecast errors* as

$$
\begin{aligned}
v_1 &= y_1 - \mu_{1|0} \\
v_2 &= y_2 - \mu_{2|1} = y_2 - \mu_{1|0} - K_1(y_1 - \mu_{1|0}) \\
v_3 &= y_3 - \mu_{3|2} = y_3 - \mu_{1|0} - K_2(y_2 - \mu_{1|0}) - K_1(1 - K_2)(y_1 - \mu_{1|0}) \qquad \text{and so on.}
\end{aligned}
$$

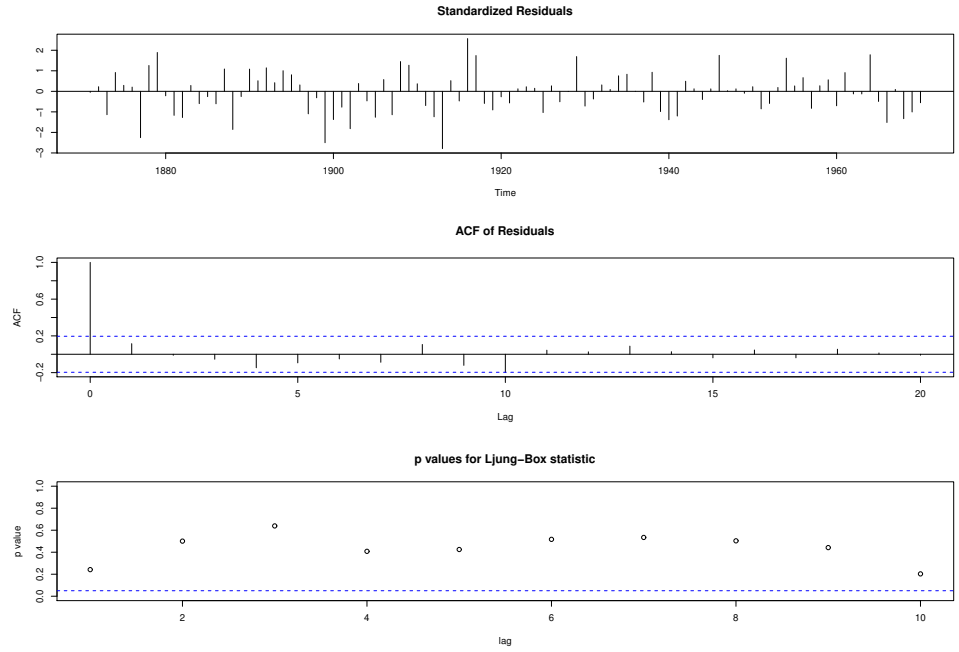


Figure 24: Time series diagnostics for fit of state space model to Nile data.

We write this in matrix form as

$$v = K\,(y - \mu_{1|0}1_n), \tag{8}$$

where $v^T = (v_1, \ldots, v_n)$, $y^T = (y_1, \ldots, y_n)$, 1_n is a $n \times 1$ vectore of ones, and

$$K = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ k_{21} & 1 & 0 & \cdots & 0 \\ k_{31} & k_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ k_{n1} & k_{n2} & k_{n3} & \cdots & 1 \end{pmatrix}$$

where $k_{i,i-1} = -K_{i-1}$, $k_{ij} = -(1 - K_{i-1})(1 - K_{i-2}) \cdots (1 - K_{j+1})K_j$,
for $i = 2, \ldots, n$ and $j = 1, \ldots, i - 2$.

Equation 8 has consequences in terms of prediction.

Lemma 121. *The one-step-ahead forecast errors $v_1, \ldots, v_n$ are mutually independent normal random variables, and the matrix K provides a Cholesky decomposition of $\Sigma = var(y)$:*

$$K\Sigma K^T = diag(V_1, \ldots, V_n).$$

The following lemma suggests how to compute the likelihood as

$$f(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} f(v_t; \theta),$$

where $v_t \sim \mathcal{N}(0, V_t)$, and the variances V_t depend on the parameters $\theta = (\sigma^2, \sigma_\eta^2)$; we assume that $\mu_{1|0}$ and $\Sigma_{1|0}$ are known. Unfortunately, it is not uncommon to find variance estimates of zero, so care is needed. We can also propose paremeter estimations with the same results.

Example 122 (Return on the Nile)**.** *For the river Nile, we find* $\hat{\sigma}^2 = 15098.577$*,* $\hat{\sigma}_\eta^2 = 1469.147$*, so the changes in the state due to the 'state innovation'* η_t *are dwarfed by the 'observation error'* ε_t*.*

17.2 State error recursion

Define $x_t = \mu_t - \mu_{t|t-1}$ to be the forecast of the state variable μ_t, given $\mathcal{H}_{t-1}$. Evidently $\text{var}(x_t \mid \mathcal{H}_{t-1}) = \text{var}(\mu_t \mid \mathcal{H}_{t-1}) = \Sigma_{t|t-1}$, and the Kalman filter gives that

$$v_t = y_t - \mu_{t|t-1} = \mu_t + \varepsilon_t - \mu_{t|t-1} = x_t + \varepsilon_t.$$

Now we get a recursion for x_t, as follows:

$$\begin{aligned} x_{t+1} &= \mu_{t+1} - \mu_{t+1|t} = \mu_t + \eta_t - (\mu_{t|t-1} + K_t v_t) \\ &= x_t + \eta_t - K_t v_t = x_t + \eta_t - K_t(x_t + \varepsilon_t) \\ &= (1 - K_t)x_t + \eta_t - K_t\varepsilon_t \end{aligned}$$

and so we have 'observation' and 'state' equations

$$v_t = x_t + \varepsilon_t, \quad x_{t+1} = L_t x_t + \eta_t - K_t\varepsilon_t, \quad t = 1, \ldots, n,$$

where $L_t = 1 - K_t = 1 - \Sigma_{t|t-1}/V_t = (V_t - \Sigma_{t|t-1})/V_t = \sigma^2/V_t$ and $x_1 = \mu_1 - \mu_{1|0}$

The goal is now to estimate the state variables $\mu_1, \ldots, \mu_n$ based on $y_1, \ldots, y_n$: we seek to compute the (multivariate normal) distribution of $\mu_t \mid \mathcal{H}_n$, for each t we seek conditional mean and variance $\mu_{t|n}$ and $\Sigma_{t|n}$, here called the **smoothed state mean** and the **smoothed state variance**.

We use the following facts:

- the $v_1, \ldots, v_n$ are mutually independent functions of $y_1, \ldots, y_n$
- if $y_1, \ldots, y_n$ are fixed, then $\mathcal{H}_{t-1}$ and $v_1, \ldots, v_n$ are fixed, and vice versa.
- $v_t, \ldots, v_n$ are independent of $\mathcal{H}_{t-1}$ with mean zero and $\text{var}(v_j) = V_j$, for each $j \geq t$

We apply Lemma 118 (c) to the conditional joint distribution of $\mu_t, v_t, \ldots, v_n$ given $\mathcal{H}_{t-1}$, and obtain, with $Y \equiv \mathcal{H}_{t-1}$ and $Z = (v_1, \ldots, v_n)^T$,

$$\begin{aligned} \mu_{t|n} &= \text{E}(\mu_t \mid \mathcal{H}_{t-1}, v_t, \ldots, v_n) \\ &= \mu_{t|t-1} + \begin{pmatrix} \text{cov}(\mu_t, v_t) \\ \text{cov}(\mu_t, v_{t+1}) \\ \cdots \\ \text{cov}(\mu_t, v_n) \end{pmatrix}^T \begin{pmatrix} V_t & 0 & \cdots & 0 \\ 0 & V_{t+1} & \cdots & 0 \\ \cdots & \cdots & & \cdots \\ 0 & 0 & \cdots & V_n \end{pmatrix}^{-1} \begin{pmatrix} v_t \\ v_{t+1} \\ \cdots \\ v_n \end{pmatrix} \\ &= \mu_{t|t-1} + \sum_{j=t}^{n} \text{cov}(\mu_t, v_j) V_j^{-1} v_j \end{aligned} \tag{9}$$

To compute $\mu_{t|n}$ we need $\operatorname{cov}(\mu_t, v_j) = \operatorname{cov}(x_t, v_j)$, because $\mathrm{E}(\mu_{t|t-1}\varepsilon_j) = 0$ for any $j \geq t$, and as $\mathrm{E}(v_j) = 0$, $\operatorname{cov}(x_t, v_j) = \mathrm{E}(x_t(x_j + \varepsilon_j))$, which equals

$$\begin{aligned}
\operatorname{var}(x_t) &= \Sigma_{t|t-1}, && j = t,\\
\mathrm{E}(x_t(L_t x_t + \eta_t - K_t\varepsilon_t)) &= \Sigma_{t|t-1}L_t, && j = t+1,\\
\mathrm{E}(x_t(L_{j-1}x_{j-1} + \eta_{j-1} - K_{j-1}\varepsilon_{j-1})) &= \Sigma_{t|t-1}\prod_{i=t}^{j-1} L_i, && j = t+2,\ldots,n,
\end{aligned}$$

and consequently (9) becomes $\mu_{t|n} = \mu_{t|t-1} + \Sigma_{t|t-1}\, q_{t-1}$ where

$$q_{t-1} = \frac{v_t}{V_t} + L_t\frac{v_{t+1}}{V_{t+1}} + L_tL_{t+1}\frac{v_{t+2}}{V_{t+2}} + \cdots + \left(\prod_{j=t}^{n-1} L_j\right)\frac{v_n}{V_n}$$

We thus obtain a backwards recursion to compute the smoothed state variables:

$$q_{t-1} = V_t^{-1}v_t + L_tq_t, \quad \mu_{t|n} = \mu_{t|t-1} + \Sigma_{t|t-1}q_{t-1}, \quad t = n,\ldots,1,$$

where $q_n = 0$, and $\mu_{t|t-1}$, $\Sigma_{t|t-1}$, and $L_t = 1 - K_t$ are available from the forward pass of the Kalman filter (7)

Example 123 (Kalman filter and smoother)**.** *We take arbitrary initial values of 0 and 1 for $\mu_{1|0}$ and $\Sigma_{1|0}$, and need guesses for σ^2 and σ_η^2*

For the filter, we then iterate forward:

t	y_t	$\mu_{t\|t-1}$	v_t	$\Sigma_{t\|t-1}$	V_t	K_t	L_t	q_t	μ_t
1	y_1	*0*		*1*					
2	y_2								
3	y_3								
$\vdots$	$\vdots$								
$n-1$	y_{n-1}								
n	y_n							*0*	

For the smoother, we need an initial value of q_n (here 0), then we iterate backwards.

Similar manipulations give a recursion to compute the conditional variance $\Sigma_{t|n}$ of μ_t given $y_1,\ldots,y_n$, as follows:

$$M_{t-1} = V_t^{-1} + L_t^2M_t, \quad \Sigma_{t|n} = \Sigma_{t|t-1} - \Sigma_{t|t-1}^2M_{t-1}, \quad t = n,\ldots,1,$$

where we take initial value $M_n = 0$.

The figure on the next slide shows the difference between **filtering** and **smoothing**

Prediction of the next value is based on $\mu_{n+1|n}$, with variance $\Sigma_{n+1|n}$ for the predicted mean μ_{n+1}, and variance $\Sigma_{n+1|n} + \sigma^2$ for the next observation $y_{n+1} = \mu_{n+1} + \varepsilon_{n+1}$.

Example 124 (Again on the Nile)**.** *On the Nile dataset, we get the following (Fig. 25) estimates after smoothing.*

Missing values

– Suppose that the observations $y_{l+1},\ldots,y_{l+h}$ are missing, with $l \in \{1,\ldots,n\}$ and $h \geq 1$.

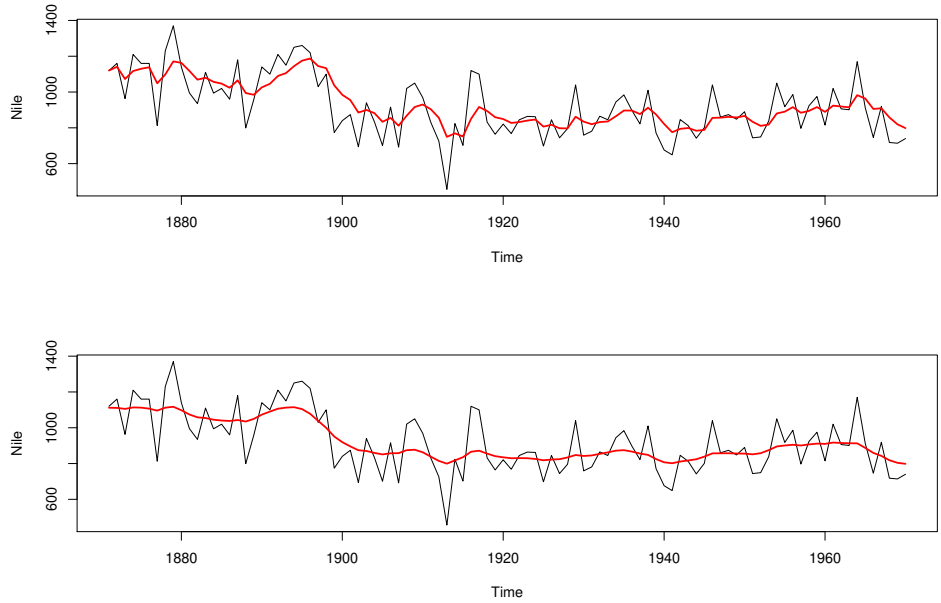


Figure 25: Nile data. Top panel: data y_t and filtered estimate of stated variable $\mu_{t|t-1}$(red). Bottom panel: data and smoothed state variable $\mu_{t|n}$

– We can use the state equation to write the state for a missing observation as

$$\mu_t = \mu_{t-1} + \eta_t = \ldots = \mu_l + \sum_{j=l+1}^{t} \eta_j$$

where the sum equals zero if $l + 1 > t - 1$.

– This implies that for missing observations, we have

$$\mathrm{E}(\mu_t \mid \mathcal{H}_{t-1}) = \mathrm{E}(\mu_t \mid \mathcal{H}_l) = \mu_{l+1|l} \quad \mathrm{var}(\mu_t \mid \mathcal{H}_{t-1}) = \mathrm{var}(\mu_t \mid \mathcal{H}_l) = \Sigma_{l+1|l} + \sigma_\eta^2,$$

and this yields the recursion

$$\mu_{t|t-1} = \mu_{t-1|t-2}$$

Thus for missing observations we apply the Kalman recursion (7) with $v_t = 0$, $K_t = 0$.

Example 125 (Crime on the Nile?)**.** *If we assume part of the data of the Nile dataset is missing, we can still estimate the underlying process, as shown in Fig 26*

17.3 Initialisation of the Kalman filter

The Kalman filter is initialised by taking $\mu_1 \sim \mathcal{N}(\mu_{1|0}, \Sigma_{1|0})$. Now we discuss the choice of $\mu_{1|0}$ and $\Sigma_{1|0}$.

Since $v_1 = y_1 - \mu_{1|0}$ and $\Sigma_1 = \Sigma_{1|0} + \sigma^2$, we get

$$\mu_{1|1} = \mu_{1|0} + \frac{\Sigma_{1|0}}{\Sigma_{1|0} + \sigma^2}(y_1 - \mu_{1|0}), \qquad \Sigma_{1|1} = \frac{\Sigma_{1|0}}{\Sigma_{1|0} + \sigma^2}\sigma^2 + \sigma_\eta^2$$

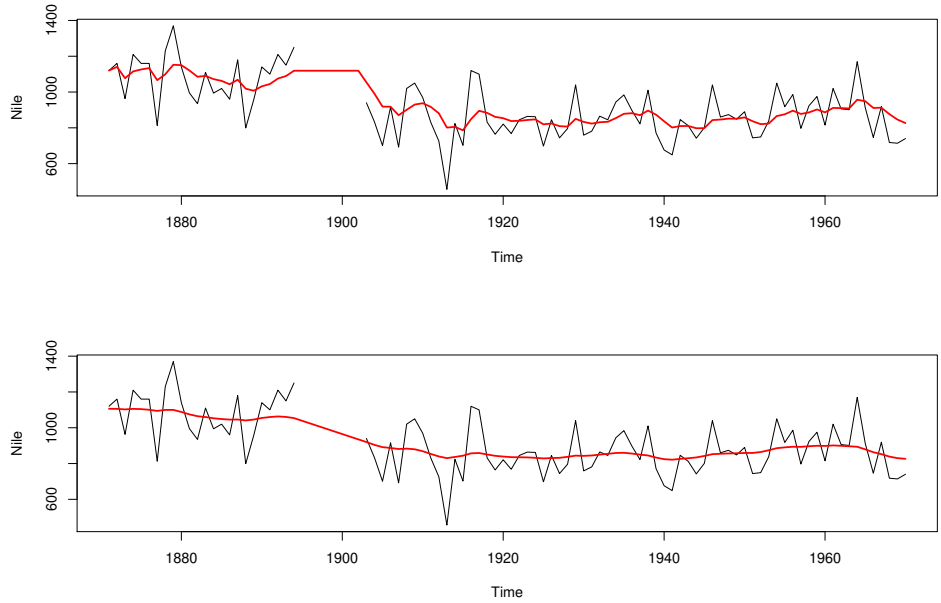


Figure 26: Nile data. Top panel: data y_t, with missing values, and filtered state variable $\mu_{t|t-1}$. Bottom panel: data, with missing values, and smoothed state variable $\mu_{t|n}$.

Thus taking a diffuse prior by letting $\Sigma_{1|0} \to \infty$ is equivalent to assuming that $\mu_1 \sim \mathcal{N}(y_1, \sigma^2)$.

This is called **diffuse initialisation** and amounts to treating y_1 as fixed.

For the state smoothing, we get

$$\mu_{1|n} = \mu_{1|0} + \frac{\Sigma_{1|0}}{\Sigma_{1|0} + \sigma^2}(v_1 + \sigma^2 q_1), \qquad \Sigma_{1|n} = \frac{\Sigma_{1|0}}{\Sigma_{1|0} + \sigma^2}\sigma^2 - \left(\frac{\Sigma_{1|0}}{\Sigma_{1|0} + \sigma^2}\right)^2 \sigma^4 M_1$$

so letting $\Sigma_{1|0} \to \infty$ gives $\mu_{1|n} = y_1 + \sigma^2 q_1$, and $\Sigma_{1|n} = \sigma^2 - \sigma^4 M_1$.

If setting $\Sigma_{1|0} = \infty$ is problematic, one may need to estimate μ_1 in the same way as the other parameters.

18 General state space model

Many models, including ARIMA models, can be written in the general state space form

$$\begin{aligned} \text{State equation :} \quad & \mu_{t+1} = T_t\,\mu_t + R_t\,\eta_t, && \eta_t \overset{\text{iid}}{\sim} \mathcal{N}_m(0, Q_t) \\ \text{Observation equation :} \quad & y_{t+1} = Z_t\,\mu_t + \varepsilon_t, && \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}_p(0, H_t) \end{aligned}$$

where the y_t have dimension $p \times 1$, the η_t have dimension $m \times 1$, and the η_t and ε_t are all mutually independent.

The matrices T_t, R_t, Q_t, Z_t, H_t are supposed known initially but may contain parameters to be estimated, and we suppose that $\mu_1 \sim \mathcal{N}(a_1, P_1)$, independently of the ε_t and η_t. The matrices Z_t and T_{t-1} may depend on $\mathcal{H}_{t-1}$, and often $R_t = I_m$.

The state equation is a first-order autoregression, and hence a first-order Markov process, and the observation equation is a linear regression model, but with correlated variables.

Essentially, all the previous computations go through again, but with a lot of linear algebra—see, for example, **Shumway and Stoffer**: *Time Series Analysis and Its Applications: With R Examples* (2006, Chapter 6)

Example 126 (ARMA as a state space model)**.** *Consider the ARMA(p, q) model written in the form*

$$y_t = \sum_{j=1}^{r-1} \phi_j y_{t-j} + \varepsilon_t + \sum_{j=1}^{r-1} \theta_j \varepsilon_{t-j}, \qquad t = 1, \ldots, n,$$

where $r = \max(p, q+1)$ and for which some of the coefficients are zero.

We set

$$Z_t = 1 \quad 0 \quad 0 \cdots 0),$$

$$\mu_t = \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \cdots + \phi_r y_{t-r+1} + \theta_1 \varepsilon_t + \cdots + \theta_{r-1} \varepsilon_{t-r+2} \\ \phi_3 y_{t-1} + \cdots + \phi_r y_{t-r+2} + \theta_2 \varepsilon_t + \cdots + \theta_{r-1} \varepsilon_{t-r+3} \\ \vdots \\ \phi_r y_{t-1} + \theta_{r-1} \varepsilon_t \end{pmatrix}$$

$$T_t \equiv T = \begin{pmatrix} \phi_1 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ \phi_{r-1} & 0 & & 1 \\ \phi_r & 0 & & 0 \end{pmatrix}, \quad R_t \equiv R = \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{r-1} \end{pmatrix}, \quad \eta_t = \varepsilon_{t+1}, \quad H_t \equiv 0$$

With the observation equation $y_t = Z_t \mu_t$, this gives the ARMA(p, q) model.

18.1 Comparison with ARIMA models

ARMA models

- are empirical models, not based on the structure of the underlying system
- provide a simple flexible black box approach to modelling, based on empirical considerations (ACF, PACF, AIC, ...)
- can deal with seasonality and trend
- are easily fitted using standard software
- can easily be used for forecasting

State space models

- enable knowledge of the underlying system to be built in
- require thought about the problem, to give an appropriate structure
- can deal with seasonality and trend

– extend in obvious ways to multivariate data

– can be fitted to a wider class of nonlinear models, using the Markov structure and particle filters

– are not so easy to understand(?)

– are not so easy to fit, because software is less widely available.

19 Conclusion

– State space modelling provides a very general approach to time series, encompassing very many time-domain approaches.

– They enable problem-specific information about the structure of the model to be built in to the formulation.

– These are models with unobserved Markovian structure, to which is added observational noise.

– The calculations involve recursive algorithms

  – which are explicit in the linear Gaussian case,
  – but require integration or other approximations to integrals in general

– They are **Formula One** time series models: not very available, not so easy to drive, but much more powerful than the family saloon, once you know how they work.

A Reminder: Normal distributions

Definition 127. Let $\mu = (\mu_1, \ldots, \mu_n)^T \in \mathbb{R}^n$ and Σ be $n \times n$ positive definite matrix with elements σ_{jk}. Then we say that the random vector $X = (X_1, \ldots, X_n)^T$ with density

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \qquad x, \mu \in \mathbb{R}^n;$$

has a **multivariate normal distribution** with expectation vector $\mu_{n\times 1}$ and covariance matrix $\Sigma_{n\times n}$.

We remind you that:

– the distribution is determined by its mean and covariances

$$\mathrm{E}(X_j) = \mu_j, \quad \mathrm{var}(X_j) = \sigma_{jj}, \quad \mathrm{cov}(X_j, X_k) = \sigma_{jk}, \quad j \neq k$$

and X_i independent of $X_j \Rightarrow \sigma_{jk} = 0$;

– if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $X_{n\times 1} = (X_1, \ldots, X_n)^T \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$;

– linear combinations of normal variables are normal:

$$a_{r\times 1} + Q_{r\times n}X \sim \mathcal{N}_r(a + Q\mu, Q\Sigma Q^T)$$

where we suppose that $|Q\Sigma Q^T| > 0$, so $r \leq n$.

A.1 Analysis of variance

Suppose $Y_{n\times 1} \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$

Let the first column of the orthogonal matrix $Q_{n\times n}$ consist of ones; hence $Q^TQ = D$, say, is diagonal and positive definite.

Now $I = D^{-1/2}Q^TQD^{-1/2} = AA^T$, say, where $A = D^{-1/2}Q^T$ is orthonormal, giving the **analysis of variance decomposition**

$$\sum_{t=1}^{n} Y_t^2 = Y^TY = Y^TIY = Y^TA^TAY = \sum_{j=1}^{n} Z_j^2,$$

where $Z = (Z_1, \ldots, Z_n)^T = AY = D^{-1/2}Q^TY \sim \mathcal{N}_n(D^{-1/2}Q^T\mu, \sigma^2 I_n)$, so

$$Z_1 = d_1^{-1/2}q_1^TY = n^{-1/2}\sum_{t=1}^{n} Y_t, \quad Z_2 = d_2^{-1/2}q_2^TY, \quad Z_3 = d_3^{-1/2}q_3^TY, \quad \ldots,$$

in an obvious notation, and $Z_j \overset{\text{ind}}{\sim} \mathcal{N}(d_j^{-1/2}q_j^T\mu, \sigma^2)$.

This implies that Z_j^2 have independent non-central $\sigma^2\chi_1^2$ distributions, with non-centrality parameters $\delta_j = (q_j^T\mu)^2/(d_j\sigma^2)$. In particular, if $q_j^T\mu = 0$, i.e., the jth column of Q is orthogonal to the mean vector μ, then $Z_j^2 \sim \sigma^2\chi_1^2$.

Different choices of Q will give different decompositions of the variation in Y.

B Reminder from your previous years: Algebra

This section is designed to be a collection of results you might/should have seen previously.

This section is very short. It contains information you should know from previous years, as well as some that you may not know, not everything is useful but don't hesitate to check the books (using the index at the end) to find additional details, maybe you'll find interesting things. Please refer to your previous course more than this section, it's only a quick reminder.

B.1 Linear Algebra

For longer details see: Peter Lax, *Linear Algebra*.

Definition 128. A vector space E on K (often we forget to mention the body, usually $K = \mathbb{R}$ or $\mathbb{C}$) is a set where we can add and invert elements without leaving the set (that is for $x, y \in E$, $-x \in E$ and $x + y \in E$), and in which we can multiply elements with a scalar $\lambda \in \mathbb{R}$ ($\lambda x \in E$).

Remark 129. *I cannot give you a full summary of linear algebra, the previous definition is very very simplified.*

Example 130. *$\mathbb{R}^n$ is a $\mathbb{R}$ vector space, $\mathbb{R}$ is a $\mathbb{R}$ vector space of dimension 1, but $\mathbb{R}$ is a $\mathbb{Q}$ vector space of infinite dimension, $\mathbb{R}[X]$ (set of polynomials with coefficients in $\mathbb{R}$) is a $\mathbb{R}$ vector space, $\mathbb{R}^+$ is not a $\mathbb{R}$ vector space, $\mathbb{Q}$ is not a $\mathbb{R}$ vector space, $\mathbb{R}$ is a $\mathbb{Q}$ vector space.*

Definition 131. A subspace of E that is also a vector space is a vector subspace of E.

Definition 132. A map $f : E \to F$ with E and F two vector spaces is a vector space morphism (also called linear application) if it preserves the structure of E onto F, that is $f(x + \lambda y) = f(x) + \lambda f(y)$. It is also called a linear map.

Example 133. *$x \in \mathbb{R}^d \mapsto 3x$ is linear, $x \mapsto x+1$ is not.*

Definition 134.
- If $f : E \to E$ is a morphism, it is called an *endomorphism*.
- If $f : E \to \mathbb{R}$ is a morphism, it is called a linear form.
- If f is injective and surjective, f is bijective and called *homeomorphism*. Only bijective mappings can be inverted, we write f^{-1} the morphism such that $f(f^{-1}(x)) = x$ and $f^{-1}(f(x)) = x$ (note that it is possible to find cases where only the left inverse or right inverse exist for non bijective mappings).

Definition 135. A set of vector $e_1, \ldots, e_n \in E$ is said to be linearly independent iif $\sum \lambda_i e_i = 0 \Rightarrow \forall i,\ \lambda_i = 0$.

Definition 136. The vector space spanned by a set of vectors $e_1, \ldots, e_n$ is written $Span(e_1, \ldots, e_n)$. It is a vector subspace of E and is constituted of all linear combinations of the vectors $e_1, \ldots, e_n$.

Example 137. *In $\mathbb{R}^3$, $Span((1,0,0))$ corresponds to a line that is a vector space.*

Definition 138. A base of E is a countable set of vectors $\mathcal{B} \subset E$ such that $\mathcal{B}$ is linearly independent and if $Span(\mathcal{B}) = E$. In this case there exists a unique way to write a $x \in E$ with elements of $\mathcal{B}$.

Definition 139. If $\mathcal{B}$ a base of E has cardinal n we say that E has dimension n. We write $dim(E) = n$.

Example 140. *$\mathbb{R}^n$ has dimension n, $\mathbb{R}[X]$ is infinite dimensional, $Span((1,0,1),(1,1,0))$ has dimension 2. $Span((1,0,1)(2,0,2))$ has dimension 1.*

Theorem 141. *A real vector space of dimension n is homeomorph to $\mathbb{R}^n$, that is there exists an homeomorphism between those spaces. More generally two vector space on the same body of the same dimension are always homeomorph.*

Theorem 142. *The rank-nullity theorem state that, if $f : E \to F$ is a linear mapping. We write $Ker(f) = \{x \in E \mid f(x) = 0\}$, and $Im(f) = \{f(x) \mid x \in E\}$, both these spaces are vector subspace of E and F respectively. Then:*

$$dim(E) = dim(Ker(f)) + dim(Im(f)).$$

We can build a vector space from two arbitrary vector space:

Definition 143. If E and F are two vector space, we can define $E \oplus F$ as the vector space of the couples (x, y) where $x \in E$ and $y \in F$ with the trivially extended operations on the couples.

We say that this is the direct sum of the vector spaces. This definition also exists for vector subspaces.

Definition 144. We say that two vector subspaces F and G of E are in direct sum if $F \cup E = \{0\}$, or equivalently if $x \in F, y \in G$ are such that $x + y = 0$ then either $x = 0$ or $y = 0$.

This allows to define vector subspace by "summing" vector subspaces.

Example 145. $Span((1,1,0)) \oplus Span((1,0,1)) = Span((1,1,0),(1,0,1))$.

B.2 Euclidean Spaces

Definition 146. An inner product (that I might call scalar product) is a bilinear, symmetric positive form. (see Lax chapter 7). It is written $\langle x, y\rangle$ or $x \cdot y$.

A real vector space associated with a scalar product is called a Euclidean space.

This structure allows to define orthogonality, symmetry, etc.

A Euclidean space is naturally normed by the norm $\|x\| = \sqrt{\langle x, x\rangle}$. This gives a topology to the space E.

The representation theorem states that any linear form f on a Euclidean space can be written as $x \mapsto \langle \alpha_f, x\rangle$ for a unique α_f.

An operator $h : E \to E$ is said orthogonal if $\langle x, y\rangle = \langle h(x), h(y)\rangle$.

Remember Cauchy-Schwarz inequality

Remark 147. *These definitions can be extended for infinite dimensional spaces, like $L^2(\mathbb{R})$ the set of functions with integrable squared value. In this case, a space associated with an inner product is called* Pre-Hilbertian*, most of the properties mentionned previously are only true in finite dimension. To get these properties back, we need the space to be complete.*

B.3 Matrices

A matrix is the representation of a linear mapping between two spaces using the effect on the basis of each space as reference. Indeed, to fully define a linear mapping it is only needed to know the effect of the mapping on the basis (and then we can deduce on any vector using linearity). To be more precise the first column of the matrix contains $f(e_1)$ written in the base of the arrival space. A consequence is that the matrix product Mx corresponds to applying the linear transformation represented by M to x.

The set of matrices with p lines and q columns with coefficients in $\mathbb{R}$ (thus representing linear mappings between a vector space of dimension p and another of dimension q) is written $\mathcal{M}_{p,q}(\mathbb{R})$.

Among common operations on matrices we remind of the transposition, if $M = (m_{ij})_{ij} \in \mathcal{M}_{p,q}(\mathbb{R})$, we define $M^\top = (m_{ji})_{ij} \in \mathcal{M}_{q,p}(\mathbb{R})$. If $p = q$, this matrix represents the adjoint linear mapping , if E is euclidean. That is the operator $f^\star$ such that $\langle f(x), y\rangle = \langle x, f^\star(y)\rangle$.

The matrix product $MN = (\sum_k m_{ik}n_{kj})_{ij}$ where $M = (m_{ij})_{ij}$ and $N_{ij} = (n_{ij})_{ij}$ is defined this way so that it represent the composition of linear mappings. That is if M represents f and N represents g, then MN represents $f \circ g$. The shape requirement on the matrices multiplied corresponds to the dimension requirement of the composition of the mappings (the space in which g arrives is the space in which f starts). This makes the matrix product *non commutative*. Furthermore there exists non null matrices M and N such that $MN = 0$. We call matrices such that $M^k = 0$ nilpotent matrices.

The inverse of a matrix M written M^{-1} is the matrix of the inverse linear mapping represented by M. It only exists if the mapping is bijective, that is if M is invertible. We write $GL_n(\mathbb{R})$ the set of invertible matrices with coefficients in $\mathbb{R}$.

Definition 148. The set of matrices representing orthogonal operators is written $O_n(\mathbb{R}) \subset GL_n(\mathbb{R})$, it is not a vector subspace. It verifies $\forall M \in O_n(\mathbb{R}),\ M^{-1} = M^\top$.

Property 149. *To find how the matrix M is represented in a different base, we need to multiply the matrix by the change of base matrices Q and P: $M_1 = Q^{-1}MP$ is the matrix representing the same morphism but in the bases Q and P. We say that M_1 and M are equivalent.*

It is so common to identify matrix and linear application that the notations blurr. In particular as a matrix can canonically represent the linear mapping in $\mathbb{R}^p$ of the form $x \mapsto Mx$.

Property 150. *It is possible to rewrite the canonical inner product on $\mathbb{R}^p$ as a matrix product: $\langle x, y\rangle = x^\top y$. We will use this notation a lot.*

Now we will mostly be interested in matrices representing endomorphism. For these matrices the basis of arrival and departure state are the same, which means the matrices are square. If we change the base of the space, we need to use the same matrix on each side, which leads to the following definition.

Definition 151. We say that two matrices M and N are *similar* if there exists G an invertible matrix such that $N = G^{-1}MG$.

This is of primary importance when studying linear transformation, as finding a base in which the mapping is simple to understand can greatly simplify the computations.

For example, the best possible matrix is the diagonal matrix, which is simpler to use in general. A matrix similar to a diagonal matrix is called *diagonalisable*.

Definition 152. The eigenvectors are the vectors e_i such that $Ae_i = \lambda_i e_i$ for some λ_i called eigenvalue. A diagonalisable matrix is a matrix for which there exists a base of eigenvectors.

Theorem 153. *Spectral theorem: if $A^\top = A$ then A is diagonalisable, and we can chose an orthogonal base of eigenvectors. In other words, there exists $O \in O_n(\mathbb{R})$ such that $O^\top AO = diag(\lambda_1, \dots, \lambda_n)$.*

The proof of this result usually is based on the Jordan decomposition of the matrix.

Theorem 154. *The eigenvalues can be found as the roots of the characteristic polynomial $det(A - \lambda I_n)$, where det is the determinant in $\mathbb{R}^n$ (I will not define the determinant here).*

Theorem 155 (Cayley-Hamilton)**.** *If χ_A is the characteristic polynomial of A, then $\chi_A(A) = 0$.*

Property 156. *The set of polynomials that annihilates a matrix is an ideal. It is generated by an element called the minimal polynomial of the matrix. We can prove that the eigenvalues of a matrix are all roots of an annihilating polynomial of the matrix. This has for consequence, along with the previous theorem, that if you have an annihilating polynomial P of degree lower than p, the eigenvalues are all roots of P.*

This result has a lot of applications. For example, if you know that your matrix A is nihilpotent of order k, then you have an annihilating polynomial A^k, and the eigenvalues can only be 0, which is normal, as otherwise the powers of the matrix would never annihilate every vector.

C Reminder from the previous years: Analysis

C.1 Fourier transform

The idea of the Fourier transform is to study a function in the space of its "frequency". Although invented for periodical signals this can be also used for any function. For periodical functions, the goal is to write the function as an (infinite) sum of sinusoids, we can extend this to the continuous case:

Definition 157. Let $f \in L^1(\mathbb{R})$, the Fourier transform of f is

$$\mathcal{F}(f)(s) = \int_{\mathbb{R}} \mathrm{e}^{-2i\pi st} f(t) \mathrm{d}t.$$

The operator $\mathcal{F} : f \mapsto \mathcal{F}(f)$ is called the Fourier transform operator. We can show that $\mathcal{F}(f)$ is continuous and bounded.

D Reminder from the previous years: Probabilty and statistics

This section is designed to be a collection of results you might/should have seen previously.

You do not need distribution and measure theory for this course, but in case you have questions on the subject, see for example Rudin, *Real and complex analysis*, or the chapters dedicated in the next references.

For probabilities, I suggest Walsh *Knowing the odds*, or Billingsley *Probability and Measure*.

For statistics, Wasserman *All of Statistics: A concise course in statistical inference*.

D.1 Probabilities

A random variable is a function from Ω a measured space called the space of possibles into E a space. This definition is very generic.

We can describe the random variable directly in E using image measure, etc. and that's what we will do. If $E = \mathbb{R}^n$ there are three cases for a random variable:

1. It has a density with respect to Lebesgue measure (sometimes we just say that it's absolutely continuous). That is we can write $P(X \in A \subset E) = \int_A f(x)\mathrm{d}x$.
2. It has density with respect to a counting measure (we often say it is discrete) that we have to precise: there exists a countable set $x_1, \ldots, x_i, \ldots$ such that: $P(X \in A \subset E) = \sum_{i=1}^{\infty} p_i \mathbf{1}_{x_i \in A}$.
3. It is not in any of the above cases, and we can prove that in this case we can write $X = pX_c + (1-p)X_d$, with X_c a random variable in the first case and X_d a random variable in the second, and with $p \in [0, 1]$.

Thus we only need to work with the two first cases. We can define the standard objects:

- The expected value $\mathbb{E}[h(X)] = \int_E h(x)f(x)\mathrm{d}x$ in the continuous case, $\sum_i p_i h(x_i)$ in the discrete case.
- the variance (only for $E = \mathbb{R}$) $Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$, the covariance $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$.
- For multidimensional variables, we define the marginal distributions as the density where we have integrated out the other variables.
- We define the covariance matrices as the matrix containing the pairwise covariances between the components of the multivariate RV. It is a matrix with n lines and n columns in X is a RV with values in $\mathbb{R}^n$

D.1.1 Independence

Definition 158. Two RV are X and Y are independent iif for all open sets A and B, $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$.

This can also be defined using the densities:

Theorem 159. *X and Y with densities f_X and f_Y are independent if their joint density $f(x, y)$ writes $f(x, y) = f_X(x)f_Y(y)$.*

A common notation if $X \sim \mathcal{L}_1$ and $Y \sim \mathcal{L}_2$ to say that the couple (X, Y) has a law where X and Y are independent is to write: $(X, Y) \sim \mathcal{L}_1 \otimes \mathcal{L}_2$. If we have n idependent replicates of the same distribution we can write $(X_1, \ldots, X_n) \sim \mathcal{L}_1^{\otimes n}$.

D.1.2 Convergence of RV

Please refer to the previous books. We use mostly almost sure convergence and convergence in distribution.

D.1.3 Characteristic function

The characteristic function is a very powerful tool.

Definition 160. If X is a RV with values in $\mathbb{R}^n$, the characteristic function is defined for $t \in \mathbb{R}^n$ as:

$$\Phi_X(t) : \mathbb{E}[\exp(-i\langle t, X\rangle)].$$

As the name indicates it characterizes the distribution of X:

Theorem 161. *If X and Y are such that $\Phi_X(t) = \Phi_Y(t)$ for all t, then X and Y have the same law.*

It can also be used to characterize convergence in distribution:

Theorem 162. *If X_n is a sequence of RV, and if for all t, $\Phi_{X_n}(t) \to_{n\to\infty} \Phi_Y(t)$, then X_n converges to Y in law.*

D.1.4 Conditional densities

The conditional density is defined similarly as the conditional probabilities $f(x \mid y) = \frac{f(x,y)}{f_y(y)}$. The expectation of the conditional density is the conditional expectation. (Note that the traditional construction of these objects is inverted: we start by constructing the conditional expectation, using σ-algebra, and then we construct the conditonal densities).

Among the important properties of conditional expectation:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]].$$

Another important result is that if f is a known function,

$$\mathbb{E}[f(X) \mid X] = f(X),$$

which is a particular case of this result:

Property 163. *If Y is $\sigma(X)$ mesurable (where $\sigma(X)$ is the σ-algebra associated with X, then we have:*

$$\mathbb{E}[Y \mid X] = Y.$$

Note that conditional expectation is theoretically defined as expectation conditonally on a σ-algebra. That is, σ-algebras represent current knowledge about the universe of possible.

We can use the conditional probabilites to ease computation, using for example the following result, which is a restatement of the definition of the conditional density:

Theorem 164 (Law of total probabilies)**.**

$$f(x, y) = f(x \mid y) f_y(y).$$

D.2 Statistics

This section contains exercises. I will *not* give a correction of those exercises.

D.2.1 Statistical models

In a statistical model, that is of the form $\{\mathcal{L}(\theta) \mid \theta \in \Theta\}$, an estimator is a statistic that *aims* at appraoching, from observations X, some quantity that can be written as a functional of the (true) parameter θ_0 that has produced the observations.

By *functionals of the parameter* θ_0, we mean θ_0 but also all functions of the form $f(\theta_0)$. In particular, it can include $\mathbb{E}[X], X \sim \mathcal{L}(\theta_0)$.

Most generaly, we are directly interested in θ_0. We call quantity of interest the unkown we aim at estimating.

Remark 165. *As often stated, all models are wrong, meaning that epistemologically the* true parameter *never exists. It is however common, when studying a statistical model to check that the estimator has properties,* if the observations come from a true distribution $\mathcal{L}(\theta_0),\ \theta_0 \in \Theta$.

Remark 166. *The first, and most exciting job as a statistician, is to* propose a model*, by discussion with the person that provides the data and wants answer on the subjects. The model should :*

- *satisfy the specialist in the sense that it contains the main assumptions, phenomenon and theories.*
- *be manageable numerically and mathematically*
- *allow to answer the questions the specialist asks.*

This subject will not be addresses here.

Definition 167 (Parametric *vs.* non parametric)**.** When Θ is included in a finite dimension vector space, the model is said *parametric*. Otherwise it is said *non parametric*.

Here, we are only interested in *parametric* models. Non parametric models however appear in bootstrap methods.

Furthermore, we will focus on *independent* samples. That is we assumes that the observations are independants, and often identically distributed.

Example 168. *You observe for the N cats of your neighbourhood the number of prey they catch in a day. You propose a Poissonian model, where all cats are independent and behave the same way : $\{\mathcal{P}(\theta)^{\otimes n} \mid \theta \in \mathbb{R}^+\}$. You are interested in estimating θ that* can be interpreted *as the mean number of preys caught per day.*

Example 169. *You play* coinche *each evening with you friends. You measure the time you spend x_i each evening. A model we can propose for n evenings is $\{\Gamma(a,b)^{\otimes n} \mid a,b>0\}$.*

Notice that (a,b) has no direct interpretable sense, the average time you will play is ab for example.

Example 170. *Planes have an unkown maximum range of ξ km, you observe the distance they travel. You can propose the following model $\{\mathcal{U}(0,\xi) \mid \xi > 0\}$.*

D.2.2 Building estimators

There are two main ways to build an estimator. The first one relies on the *maximum likelihood* technique, the second on the *moment methods*.

In this section, we assume that the model is independant and parametric, that is of the form $\{\mathcal{L}(\theta) = \mathcal{D}(\theta)^{\otimes n} \mid \theta \in \Theta \subset \mathbb{R}^d\}$. We denote $(x_1,\dots,x_n)$ the observations, and $(X_1,\dots,X_n,\dots)$ a sequence of iid realisation of $D(\theta)$.

Traditionally, an estimator of θ will be called $\hat{\theta}$ or $\tilde{\theta}$.

Remark 171. *The number of observation is* finite*, while most of the theoretical results will be asymptotical in the number of relations, that is why we study the properties of the model in therm of X_i's, and not x_i's.*

Method of moments The quantities $\mathbb{E}_\theta[X^k]$ that is $\mathbb{E}[X^k]$ where $X \sim \mathcal{L}(\theta)$, can be estimating straightforwardly thanks to the *law of large numbers*.

Theorem 172 (Law of large numbers)**.** *If $\mathbb{E}_\theta[X_1] < \infty$ (*i.e. *X is $\mathcal{L}^1$, or integrable) then,*

$$\bar{X_n} = n^{-1}\sum_{i=1}^{n} X_i \to^{\mathbb{P}\ and\ a.s.}_{n\to\infty} \mathbb{E}[X_1].$$

There exists a weak law of large number with weaker assumptions and that guarantees convergence in $\mathbb{P}$ only.

This result applied to $(f(X_1),\dots,f(X_n))$ leads to :

$$f(\bar{X_n}) = n^{-1}\sum_{i=1}^{n} f(X_i) \to^{\mathbb{P}\ and\ a.s.}_{n\to\infty} \mathbb{E}[f(X_1)].$$

If the parameter of interest can be written in such a way, $\theta = f(X)$ a simple proposition for an estimator is $\hat{\theta} = \frac{1}{n}\sum f(X_i)$.

Example 173. *In the poissonian example, $\theta = \mathbb{E}_\theta[X]$.*

In the gamma example $\mathbb{E}_{a,b}[X] = ab$, $Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = ab^2$, so that we can write :

$$\hat{b} = \mathrm{Var}[X]/\mathbb{E}[X]; \hat{a} = \mathbb{E}[X]^2/\mathrm{Var}(X).$$

In the uniform example:

$$\hat{\xi} = 2\mathbb{E}[X].$$

Exercise 174. *Prove those results, that is compute the mean and variance of the densities proposed in the models.*

Maximum likelihood The likelihood contains all the information brought on the parameter by the observations. The formal definition of likelihood requires the existence of a *Radon-Nikodym derivative* of the probability distributions of the model for all the parameter with respect to a *common* dominant measure. These issues will not be addressed and require quite a heavy knowledge of measure theory. However, notice that in some cases the likelihood can be not definite (none of the examples or exercises here).

In practice, if $f(x \mid \theta)$ is the density of $\mathcal{L}(\theta)$ with respect to some measure, the likelihood is :

$$L(\theta \mid x) = f(x \mid \theta).$$

The likelihood is a function of θ in the parameter space. It is constant in the observations as it represent how the parameter can *explain* the observations.

Example 175. *The likelihood of the three models writes :*

- $L(\theta \mid x_1, \dots, x_n) = \mathrm{e}^{n\theta}\, \theta^{\sum_i x_i} / \prod_i x_i!$.
- $L(a, b \mid x_1, \dots, x_n) = \frac{(x_1 \dots x_n)^{a-1}\, \mathrm{e}^{-\sum_i x_i / b}}{\Gamma(a)^n b^a}$.
- $L(\xi \mid x_1, \dots, x_n) = \prod_i \mathbf{1}_{x_i < \xi}$.

Exercise 176. *Compute those likelihoods. Beware the definition of the Gamma distribution can change between references.*

A proposition of estimator can be the parameter that will maximize the likelihood. In other terms, an estimation of the true parameter is the parameter that would *most probably* will lead to the observation we have observed.

To do that we have to find the maximum of the likelihood function.

Example 177. *It is in general far easier to compute the maximum of the log likelihood. As we only have to derivate a sum, rather than a product when the observations are iid.*

To formally prove that it is a maximum, we have to compute the second derivative, that should be negative (or in the multidimensional case the Hessian matrix should be negative definite). Once we have found the maxima, among the critical points we have to find the global *maximum among all the local maxima.*

Exercise 178. *Show that the maximum likelihood estimator for θ and ξ are :*

- $\hat{\theta} = \bar{X_n}$.
- $\hat{\xi} = \max_i(X_i)$.

In many cases computing the maximum likelihood is extremely difficult and requires to numerically find the maximum of the function, which is even more challenging.

Remark 179. *The existence of a maximum is not automatic. For example, if we study a (stupid) model : $\{\mathcal{N}(a+b, 1) \mid a, b \in \mathbb{R}\}$, clearly for any couple a, b such that $a + b = c$ the likelihood will be constant so that the maximum is not unique.*

Exercise 180. *Make the computation of the previous stupid model to be sure of what I'm saying.*

D.2.3 Properties of estimators

Any functional of the observations $f(x_1, \dots, x_n)$ can be an estimator. for example 1 is an estiamtor. But all estimators are not equivalent. We can distinguish several properties that are valorized for an estimator.

Convergence

Definition 181. An estimator $\hat{\theta} = f(x_1, \dots, x_n)$ is convergent if, $f(X_1, \dots, X_n) \to_{n\to\infty} \theta$, with X_i iid from $\mathcal{L}(\theta)$ for any θ. That is whatever is the true parameter, if we have an infinity of observations from the true distribution the estimator converges to the true parameter. The convergence can be $a.s.$ or in $\mathbb{P}$.

Exercise 182. *Show the convergence of all the estimators we have proposed so far.*
The most interesting must be the maximum likelihood for the uniform model.

Asymptotic normality It is one of the results that allow to build a confidence interval. The most important result being the :

Theorem 183 (CLT)**.** *If X_i are iid, with finite variance and finite mean, then :*

$$\sqrt{n}(\bar{X_n} - \mathbb{E}[X_1]) \to \mathcal{N}(0, Var(X_1).$$

This result can be used with :

Theorem 184 (Slutsky lemma)**.** *If X_n and Y_n are two sequence of rv's, such that X_n converge in distribution to X and Y_n in probability to c a constant, then, for any function f of two variables, $f(X_n, Y_n) \to f(X, c)$.*
In particular $X_n Y_n \to Xc$.

and with :

Theorem 185 (Δ method)**.** *If g is a $\mathcal{C}^1$ function, under the assumptions of the TCL we have :*

$$\sqrt{n}(g(\bar{X_n}) - g(\mathbb{E}[X_1])) \to \mathcal{N}(0, g'(\mathbb{E}[X])^2 \operatorname{Var}(X_1).$$

Building a confidence interval from these results is described in another PDF. Go there for a full example.

Exercise 186. *Compute confidence intervals for the moment estimator for the uniform model and maximum likelihood for the poisson model.*

Bias

Definition 187. The bias of an estimator $\hat{\theta}$ is $\mathbb{E}_\theta[\hat{\theta} - \theta]$ that is the mean error with resect to the true value. It can also be written : $\mathbb{E}[\hat{\theta}] = \theta$.

Usually we prefer estimator to be unbiased.

Exercise 188. *Show that :*

- *the estimator for the Poisson model is unbiased*
- *the moment estimator for the uniform distribution is unbiased*
- *the maximum likelihood for the uniform distribution is biased.*

D.2.4 Some more exercises

Exercise 189. *Let $X_1, \dots, X_n$ be an iid sample from a distribution whose density with respect to the counting measure on $\mathbb{N}$ is $f(x \mid \lambda) = \frac{(1-\lambda)^2}{\lambda} x\lambda^x$.*

- *Compute the likelihood*
- *the maximum likelihood estimator is $\frac{\bar{x}-1}{\bar{x}+1}$.*
- *compute the mean of X_1 (help : maybe it is an exponential family, or maybe you can use some techniques from L2), and show that the estimator is unbiased.*
- *compute a confidence interval.*

Exercise 190. *Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be an iid sample from a distribution with density with respect to the lebesgue measure in $\mathbb{R}^2$ of the form $f(x, y \mid \nu, \zeta) = \zeta\nu \, \mathrm{e}^{-\nu x y} \, y^{-\zeta} \mathbf{1}_{x>0, y>1}$*

- *Pronounce correctly the name of the greek letters ν and ζ.*
- *show that it is a density for $\zeta > 1$ and $\nu > 0$.*
- *compute the maximum likelihood estimator $\widehat{(\nu, \zeta)} = \left(\frac{\sum x_i y_i}{n}, \frac{\sum \log(y_i)}{n} \right)$*
- *Compute $\mathbb{E}[XY]$ and $\mathbb{E}[log(Y)]$ (help : exponential family, or rough L2 computation).*
- *study the bias of our estimators.*

D.2.5 Properties of statistics

Fisher information Several properties are valued for a statistic. First, we can be looking for statistic that *bear as much information as possible.* The "information" can be formally defined as the Fisher information :

Definition 191. If $X \sim \mathcal{L}(\theta)$, the Fisher information is $I_X(\theta) = \mathbb{E}_\theta[\left(\frac{\partial}{\partial\theta} f(X \mid \theta)\right)^2]$ or equivalently $\mathbb{E}_\theta[-\frac{\partial^2}{\partial\theta^2} f(X \mid \theta)]$.

It measures how much information is brought by X on θ. It has many properties, among which :

Properties 192. *If $X_1, \dots, X_n$ are iid, $I_{X_1,\dots,X_n}(\theta) = nI_{X_1}(\theta)$.*
If $\eta = f(\theta)$, $I_X(\theta) = \left(\frac{\partial f(\theta)}{\partial\theta}\right)^2 I_X(\eta = f(\theta))$.

This second properties comes from changes of variables. Beware when you change the parameter to use this property.

Exercise 193. *Show the change of variable.*

Exercise 194. *We define a distribution with a density with respect to Lebesgue measure of the form $f(x \mid \theta) = (\theta^2 + 1)x^{(\theta^2)} \mathbf{1}_{x\in[0,1]}$, we propose an alternative parametrization $f(x \mid \eta) = x^\eta/(\eta + 1)\mathbf{1}_{x\in[0,1]}$.*

- *Compute $\mathbb{E}[\log(X)]$ using the exponential family properties. It should be equal to $1/(\theta^2 + 1)$*
- *Compute the Fisher information for η, and θ.*
- *Compute the Fisher information for η starting from θ.*

Sufficient statistic A statistic $S(X_1, \dots, X_n)$ is said to be sufficient if $I_{S(X_1,\dots,X_n)}(\theta) = I_{X_1,\dots,X_n}(\theta)$. Which means that all the information on θ from X is contained in S.

Example 195. *$S(X_1, \dots, X_n) = (X_1, \dots, X_n)$ is sufficient.*
If X_i are iid $\mathcal{B}(p)$, then $I_{X_1,\dots,X_n}(p) = \frac{n}{p(1-p)} = I_{\sum X_i}(p)$ as $\sum X_i$ follows a Binomial distribution.

Theorem 196. *S is sufficient if and only if there exists g and h such that $f(x \mid \theta) = g(s(x), \theta)h(x)$.*

Exercise 197. *Show that in an exponential family, with natural statistic T, $\sum T(x_i)$ is a sufficient statistic.*

This result can also be used to prove that a statistic is *not* sufficient.

Property 198. *If for two samples $X \neq Y$, we have $S(X) = S(Y)$ and $L(\theta \mid X)/L(\theta \mid Y)$ is* not independent *of θ then the statistic* cannot *be sufficient.*

Example 199. *In a Poisson model, we have three observations $x = (x_1, x_2, x_3)$. The statistic $S(x) = x_1 - x_2 + x_3$ is not sufficient, as $L(\lambda \mid (1,1,1))/L(\lambda \mid (2,2,1)) = \lambda^2/4$.*

Among sufficient statistics, some are minimal.

Definition 200. A sufficient statistic S is minimal if for any sufficient statistic T, there exists g such that $S(X) = g(T(X))$.

In order to show minimal sufficiency, it can be useful to remember the following result, quite theoretical (from Bourbaki):

Theorem 201 (Factorisation theorem)**.** *Let E, F, G sets, $g : E \to F$, $f : E \to G$. There exists $h : F \to G$ such that $f = h \circ g$* if and only if *$g(x) = g(y) \Rightarrow f(x) = f(y)$.*

Example 202 (This example shows a technique not so common)**.** *In a Cauchy model, the likelihood for a n sample writes:*

$$L(\theta \mid X) = \frac{1}{\pi^n \prod_{i=1}^n (1 + (X_i - \theta)^2)}.$$

Clearly the order statistic $S(X) = X_{(1)}, \dots, X_{(n)}$ is sufficient. To prove that it is minimal, we will show that any sufficient statistic T can factorise into the order statistic. That is we have to show that if $T(X) = T(Y)$ then $S(X) = S(Y)$.

As T is sufficient, we can write: $L(\theta \mid X)/L(\theta \mid Y) = \frac{h(x)g(T(x),\theta)}{h(y)g(T(y),\theta)} = \frac{h(x)}{h(y)} = \frac{1+(x_i-\theta)^2}{1+(y_i-\theta)^2}$, as $T(x) = T(y)$ by assumption. This means that the rational function $\theta \mapsto \frac{1+(x_i-\theta)^2}{1+(y_i-\theta)^2}$ does not depend on θ, so that the roots of numerator and denominator are the same. And we know those roots, that are the $x_i \pm i$. It means that T is only equal when the values taken by the observations are the same up to ordering, which also implies that the increasing ordering of the observations — that is the order statistic — is equal on these samples.

Sufficiency is particularly interesting in the case of exponential families : having a minimal sufficient statistic of fixed dimension, independently of the sample size is equivalent to be in an exponential family. Proving this result is heavily complex.

This result however drastically reduces the interest of the minimal sufficiency. It also means there is no hope of finding a small dimensioned sufficient statistic for almost every model that exists.

Ancillary statistic The opposite of sufficiency.

Definition 203. A statistic S is ancillary with respect to parameter θ if its distribution does not depend on θ.

Example 204. *Let $\mathcal{L}(\theta)$ be a symmetric distribution. Then $sign(X)$ is ancillary.*

Exercise 205. *Show it.*

Exercise 206. *Let X be a rv with density with respect to Lebesgue : $f(x \mid \theta) = e^{-\theta}\theta^{\lfloor x \rfloor}/\lfloor x \rfloor !$. Show that $X - \lfloor X \rfloor$ is ancillary.*

Complete statistic

Definition 207. A statistic S is said complete if $\forall \theta, \mathbb{E}_\theta[f(S(X))] = 0 \ \Rightarrow f = 0, a.s.$

To show completeness you may need some results from analysis.

Example 208. *In a Poisson model, the statistic $S(X) = \sum X_i$ is complete. Indeed, $S(X)$ follows a Poisson $n\lambda$ distribution. So that, if $\mathbb{E}_\lambda[g(S(X)] = 0$, it means that $\mathrm{e}^{-n\lambda} \sum_{s=1}^{n} g(s)(n\lambda)^s/s! = 0$ for all λ.*

This sum is a power series. In particular it is constantly zero on an open set ($\mathbb{R}^+$), this implies by isolated zeros that $\forall s, g(s)n^s/s! = 0$ and necessarily, $g(s) = 0$.

Exercise 209. *Show that in a Binomial model with known number of draws k, $\sum X_i$ is complete.*

Example 210. *The natural statistic of an exponential statistic is always complete (complex proof, not in your course).*

Basu's theorem

Theorem 211. *If T is a complete and sufficient statistic, and S is ancillary, both with respect to a parameter θ, then T and S are independent for any θ.*

Exercise 212. *Let X_i be iid $\mathcal{E}(\theta)$. Show that:*

- *$g(X) = \frac{X_n}{X_1 + \cdots + X_n}$ is ancillary (you can introduce $Z_i = X_i/\theta$,*
- *$X_1 + \cdots + X_n$ is sufficient complete,*
- *deduce that $\mathbb{E}[g(X)] = 1/n$.*

D.2.6 Quality of estimators

When comparing two estimators, there are several ways to quantify their quality.

- The bias, an unbiased estimator is usually considered to be better than an asymptotically unbiased (*i.e.* $\mathbb{E}_\theta[\hat{\theta}_n] \to \theta$), which is better than a biased estimator.
- The variance of the estimator. Clearly, if an estimator has a smaller variance, it means that the associated confidence interval will be smaller.

Example 213. *In a $\mathcal{U}([0, \theta])$ model, $2\bar{X}_n/n$ is unbiased, but has a variance of order n^{-1}, beucause of the CLT.*

By comparison, $max(X_i)$ is a biased estimator, but we can compute its variance, that will be of order n^{-2}.

Best estimator This section is based upon Lehmann-Scheffé's theorem:

Theorem 214. *Let $f\hat{(\theta)}$ be an* unbiased *estimator of $f(\theta)$. Let S be a complete and sufficient statistic.*

Then $\mathbb{E}_\theta[f\hat{(\theta)} \mid S]$ is the unique best estimator in the sense of the variance among unbiased estimators *of $f(\theta)$.*

This results can be used more often under its particular form :

If $f\hat{(\theta)}$ depends on a the observations only through S, that is $f\hat{(\theta)} = g(S(X))$, then $\mathbb{E}_\theta[f\hat{(\theta)} \mid S] = f\hat{(\theta)}$, and it is the best estimator among etc.

Example 215. *In an exponential family, we know that $\sum T(X_i)$ is sufficient and complete. The estimators we build will then be the best estimators etc. as they are all build on these quantities, provided they are unbiased.*

In other situations, we can have to compute the conditional distribution of $f\hat{(\theta)} \mid S$.

Exercise 216. *In a Poisson model $\mathcal{P}(\lambda)$ we want to estimate $g(\lambda) = \lambda^2$.*

Show that $T(X) = (\bar{X}_n)^2 - \bar{X}_n/n$ is the best unbiased estimator.

Beware, in the exponential family, unbiasedness depend on the parameter estimated, we can unbiasedly estimate natural parameter, but not necessarily the original parameters.

D.2.7 Best estimator

If the statistic is not complete, we still have a result fairly interesting:

Theorem 217 (Rao-Blackwell's theorem)**.** *Let $f\hat{(\theta)}$ be an estimator of $f(\theta)$. Let S be a sufficient statistic.*

Then $\mathbb{E}_\theta[f\hat{(\theta)} \mid S]$ has a lower variance *than $f\hat{(\theta)}$.*

Notice that there is no unbiasedness assumption. Notice also that computing the improved version of an estimator can be complex and often not possible.

Exercise 218 (more difficult)**.** *In a Poisson model with parameter λ, let $T(X) = \mathbf{1}_{X_1=0}$ be an estimator. Compute the Rao-Blackwell improved version of T, through an idoneous statistic.*

You can use the fact that T is binary so that $\mathbb{E}[T \mid S] = P(T = 1 \mid S)$. You can also use the fact that the sum of independent Poisson r.v.'s follows a Poisson distribution.

D.2.8 A bound on goodness

The quality of estimators are bounded by the Fisher information.

Theorem 219 (Cramér-Rao bound)**.** *If $\hat{\theta}$ is an estimator of θ, with bias $b(\theta) = \mathbb{E}_\theta[\hat{\theta}] - \theta$, then,*

$$\mathrm{Var}_\theta(\hat{\theta}) \geq (1 + b'(\theta))I_X(\theta)^{-1}.$$

In particular, if the estimator is unbiased, the bound writes :

$$\mathrm{Var}_\theta(\hat{\theta}) \geq I_X(\theta)^{-1}.$$

This bound is in practice very rare to reach. For an iid sample, the Carmér-Rao bound has order n^{-1} (the one in the CLT), meaning that more efficient estimators in the sense of the variance will have to be biased.

Exercise 220. *Compute the Cramér-Rao bound for T in the previous exercice, notice the change of parameter.*

Show that $\mathrm{Var}(T) = \frac{4n\lambda^3+2\lambda^2}{n^2}$, *and conclude.*

D.2.9 Again some more exercises

Exercise 221. *Let X_i be an iid sample whose law has a density with respect to Lebesgue measure of the form $f(x \mid \lambda) = \mathbf{1}_{x>0} 2\lambda x \, \mathrm{e}^{-\theta x^2}$.*

- *Compute the distribution of $Y_i = \theta X^2$ and deduce the distribution of $\sum Y_i$. Deduce that $\hat{\lambda} = \frac{n}{\sum X_i^2}$ is a biased estimator of λ. We remind that if $Z \sim \Gamma(a,b)$, $1/Z \sim I\Gamma(a,b)$ and $\mathbb{E}[1/Z] = b/(a-1)$.*
- *From previous question, find an unbiased estimator of λ.*
- *Compute the variance of this estimator, the variance of an $\Pi\Gamma(a,b)$ is $\frac{b^2}{(a-1)^2(a-2)}$.*
- *compare it with the Cramér-rao bound that you will have computed as λ^2/n.*

Exercise 222. *Let X_i iid $\mathcal{B}(p)$.*

Show that $\hat{p} = 2 - 2/(1+X_1)$ is an unbiased estimator of p.

Show that $\sum X_i$ is a sufficient statistic.

Compute the distribution of $1/(1+X_1) \mid \sum X_i$, and deduce a Rao-Blackwellised version of this estimator.

Show that this estimator reaches the Cramér-Rao bound.

D.2.10 Statistical tests

A test problem writes for a statistical model $\{\mathcal{P}(\theta) \mid \theta \in \Theta\}$ as a set of two hypotheses: $H_0 : \theta \in \Theta_0$ that tends to be the most restrictive hypothesis, and H_1 that is usually of the form $H_1 : \theta \in \Theta_0^c$ (but we can have $H_1 : \theta \in \Theta_1$). Usually, H_0 is the most restrictive hypothesis.

There are two results possible for a test:

- We reject H_0, that is we accept H_1;
- we do not reject H_0, which does not mean H_0 is true.

We can already see the problems with this definition of tests. Sometimes we want to be able to say that H_0 is true, but the setting does not allow to reach this conclusion.

The idea of the test is usually to find a statistic T such that under H_0 its law is known and independent of θ (that is ancillary).

From this we need to see if the observation $T(x)$ is improbable or not, that is we check if $T(x)$ falls into a region of large probability. The complementary of this region is called *rejection area* for the test.

Example We work on a Gaussian model:

$$\{\{\mathcal{N}(m,\sigma^2)^{\otimes n}\} \mid m \in \mathbb{R}, \sigma \in \mathbb{R}^{+*}\}$$

We note $x_1, \ldots, x_n$ the observations, and let's write $P_{m,\sigma}$ the density of $\mathcal{N}(m,\sigma^2)^{\otimes n}$.

We want to build a test of the form $H_0 : m \geq m_1$, against $H_1 : m < m_1$.

Let's write $\bar{x}_k = k^{-1}\sum_{i=1}^{k} u_i$, in this case a natural rejection area is $\{x \in \mathbb{R}^n \mid \bar{x}_n \leq s\}$ as H_0 is rejected if the observations are abnormally small. Unfortunately, the distribution of the statistic used to build the rejection region $\bar{x}_n$ is not available, so we cannot use this region.

To adapt the region, let t be the quantile of order α of the Student t distribution $\mathcal{T}_{n-1}$, then under H_0, we have, with $S_n(x)$ the unbiased estimator of the variance:

$$P_{m,\sigma}\left(\bar{x}_n \leq m_1 + t\frac{S_n(X)}{\sqrt{n}}\right) \leq P_{m,\sigma}\left(\bar{x}_n \leq m + t\frac{S_n(X)}{\sqrt{n}}\right).$$

Thus, we can rewrite the second term to make appear an object that follows the Student's t distribution:

$$P_{m,\sigma}\left(\bar{x}_n \leq m_1 + t\frac{S_n(X)}{\sqrt{n}}\right) \leq P_{m,\sigma}\left(\sqrt{n}\frac{\bar{x}_n - m}{S_n(X)} \leq t\right) = \alpha,$$

by definition of the quantile.

This allows us to define a rejection area:

$$R_{reject} = \{x \in \mathbb{R}^n : \bar{x}_n \leq m_1 + t\frac{S_n(x)}{\sqrt{n}}.\}$$

Interesting tests For the Gaussian model, numerous tests exist. Most of them rely on known distribution (χ^2 for example) and are implemented in `R`.

More complex tests involve independence and adequation tests, that are *non-parametric*, they rely on more complex theoretical probabilities, but these tests are interesting. They are also implemented in `R`. The energy-distance test used in the workshop is also such an example.