Contents

Part I

MATH3802
Time Series

Chapter 1

Introduction

A *time series* $\{X_t\}$ consists of observations X_t made at times $t = t_1, \ldots, t_n$ (e.g. every day). These data are influenced by systematic effects and randomness. Typically observations at nearby times are "similar".

We wish to describe the data (e.g. plot X_t against t), suggest a model (estimate parameters and test validity), and predict future behaviour of the time series (forecasting).

1.1 Fitting a trend

The *trend* of a time series is a slow change in mean level. Consider the model

$$X_t = \mu(t) + \varepsilon_t, \tag{1.1}$$

where

$\mu(t)$ is a (deterministic) trend as a function of t;

ε_t is a (random) fluctuation about the trend at time t, $E(\varepsilon_t) = 0$.

For example, consider $\mu(t) = \alpha + \beta t, \ \alpha, \beta \in \mathbb{R}$. Suppose we have observed values $X_1, \ldots, X_n$ at times $t_1, \ldots, t_n$. We can estimate α, β by linear

regression, i.e. by minimising

$$r(\alpha, \beta) = \sum_{i=1}^{n} (X_i - \alpha - \beta t_i)^2.$$

This yields

$$\widehat{\alpha} = \overline{X} - \widehat{\beta}\overline{t} \qquad \text{and} \qquad \widehat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(t_i - \overline{t})}{\sum_{i=1}^{n} (t_i - \overline{t})^2}.$$

The estimated trend is thus

$$\widehat{\mu}(t) = \widehat{\alpha} + \widehat{\beta} t.$$

hat -> estimate

Remarks:

1. If a trend is present in the data, the usual summary statistics (mean, variance, etc) will not be very useful.
2. After a trend is fitted, the *residuals* $Y_{t_i} = X_{t_i} - \widehat{\mu}(t_i)$ can be analysed further, e.g. by estimating $\mathrm{var}(Y_{t_i})$.
3. We can predict future behaviour of X by considering $\widehat{\mu}(t)$ for $t > t_n$. Once we have analysed the residuals, we can improve the forecast by considering $\widehat{\mu}(t) + \widehat{Y}_t$ for $t > t_n$.

   Note that forecast error depends on the variance of our estimates. For example, if the ε_t are i.i.d. $N(0, \sigma^2)$, then

   $$\mathrm{var}(\widehat{\mu}(t)) = \sigma^2 \left(\frac{1}{n} + \frac{(t - \overline{t})^2}{\sum_{i=1}^{n} (t_i - \overline{t})^2} \right).$$

4. Note that trends do not have to be linear! More general trend models are possible, such as

   $$\mu(t) = \alpha + \beta t + \gamma t^2$$
   $$\mu(t) = \alpha \exp(\beta t)$$
   $$\vdots$$

1.2 Seasonal effects

Seasonal effects are periodic components of a time series which repeat with a fixed frequency. We might revise model (1.1) to be

$$X_t = \mu(t) + s(t) + \varepsilon_t, \tag{1.2}$$

where

$\mu(t)$ — trend;

$s(t)$ is a seasonal effect, with $s(t) = s(t+a) \ \ \forall t$ where a is the period length;

ε_t — residuals/fluctuations.

If we assume $t = 0, 1, 2, \ldots$ and $a = n \in \mathbb{N}$, then we have

$$s = (s_0 \ s_1 \ \ldots \ s_{n-1} \ s_0 \ s_1 \ \ldots \ s_{n-1} \ \ldots)$$

periodic seasonal

so $s(t) = s_{t \bmod n}$. To find s we have to estimate $s_0, s_1, \ldots, s_{n-1}$.

We can write

$$X_t = \mu(t) + s_0\delta_{t,0} + \cdots + s_{n-1}\delta_{t,n-1} + \varepsilon_t,$$

where

$$\delta_{t,i} = \begin{cases} 1 & \text{if } t \text{ is in season } i \\ 0 & \text{else.} \end{cases}$$

The parameters $s_0, \ldots, s_{n-1}$ can be estimated using linear regression.

For example, if t is seasonal in quarters of a year and if the period length is one year, then we have to estimate s_0 (spring), s_1 (summer), s_2 (autumn), and s_3 (winter). For the airline data, we can estimate s_0 (Jan) — s_{11} (Dec).

1.3 Summary

We start the analysis of a time series by identifying the trend (if any) in the model $X_t = \mu(t) + \varepsilon_t$. We then analyse the residuals $Y_t = X_t - \widehat{\mu}(t)$ further.

Finally, any results about the $\{Y_t\}$ need to be converted back by adding $\widehat{\mu}(t)$.

For example, a forecast for X_t is given by the forecast for $Y_t + \widehat{\mu}(t)$.

After the trend is removed, we remove the periodic components (if any) in the model $Y_t = s(t) + \varepsilon_t$. We then analyse $Z_t = Y(t) - \widehat{s}(t)$ further. Convert results about the $\{Z_t\}$ into results about $\{Y_t\}$ by adding $\widehat{s}$.

Chapter 2

Stationary processes

To get a model of a time series, we consider the X_t to be random variables. The collection $\{X_t\}_{t\in\mathbb{N}}$ is called a *stochastic process*.

Definition 2.1. *For a stochastic process $\{X_t\}$, we use the following definitions:*

$$\begin{aligned}
\mu(t) &= \mathbb{E}(X_t) && \textit{(mean)}\\
\sigma^2(t) &= \operatorname{var}(X_t) && \textit{(variance)}\\
\gamma(s,t) &= \operatorname{cov}(X_s, X_t) = \mathbb{E}[(X_s - \mu(s))(X_t - \mu(t))] && \textit{(auto-covariance)}\\
\rho(s,t) &= \operatorname{corr}(X_s, X_t) = \frac{\gamma(s,t)}{\sqrt{\sigma^2(s)\sigma^2(t)}} && \textit{(auto-correlation).}
\end{aligned}$$

Definition 2.2 (Stationarity)**.** *The stochastic process $\{X_t\}$ is* stationary *if $\{X_1, \ldots, X_n\}$ and $\{X_{1+k}, \ldots, X_{n+k}\}$ have the same distribution for all $k, n \in \mathbb{N}$.*

stationarity

Remarks:

1. If $\{X_t\}$ is stationary, then X_k has the same distribution as X_1 for all $k \in \mathbb{N}$. In particular, if the first two moments are finite then $\mu(t) = \mathbb{E}(X_t) = \mu$ and $\sigma^2(t) = \operatorname{var}(X_t) = \sigma^2$ for all t. Therefore, a process with trend or seasonal effects cannot be stationary.
2. For stationary processes, sometimes it is useful to consider $t \in \mathbb{Z}$ instead of $t \in \mathbb{N}$.

Definition 2.3 (Weak stationarity)**.** *A stochastic process $\{X_t\}$ is* weakly stationary *or* second order stationary *if*

1. *$\mu(t)$ is constant, i.e. $\mu(t) = \mu$ for all t;*
2. *$\gamma(s,t)$ only depends on the times t, s via the lag $t-s$, i.e. $\gamma(s,t) = \gamma_{t-s}$.*

Note that $\sigma^2(t) = \sigma^2 = \gamma_0$.

Example 2.1 (White noise)**.** *Let $X_t \sim N(0,1)$, i.i.d. Then $\mu(t) = 0$, $\sigma^2(t) = 1$, $\gamma(s,t) = 0$ for all $s, t \in \mathbb{N}$. Clearly $\{X_t\}$ is weakly stationary.*

Indeed, the collection $\{X_{1+k}, \ldots, X_{n+k}\}$ is an i.i.d. collection of $N(0,1)$ r.v.s for any shift $k \in \mathbb{N}$, so $\{X_t\}$ is strongly stationary.

Example 2.2 (Symmetric random walk)**.** *Let $X_0 = 0$ and $X_t = \sum_{i=1}^{t} Z_i$ where Z_i takes values $+1, -1$ with probability 1/2 each.*

Then $\mu(t) = \sum_{k=1}^{t} \mathbb{E}(Z_i) = 0$ but $\sigma^2(t) = \sum_{k=1}^{t} \mathrm{var}(Z_k) = t$, which is not constant so $\{X_t\}$ is neither weakly nor strongly stationary.

Exercise: Show that $\gamma(s,t) = s$ and $\rho(s,t) = \sqrt{s/t}$.

Remarks:

1. Every stationary process is weakly stationary since, for all t,

$$\mu(t) = \mathbb{E}(X_t) = \mathbb{E}(X_1)$$
$$\gamma(t, t+k) = \mathrm{cov}(X_t, X_{t+k}) = \mathrm{cov}(X_1, X_k).$$

   Therefore, both quantities are independent of t.
2. If $\{X_t\}$ is weakly stationary,

$$\gamma_0 = \mathrm{cov}(X_t, X_t) = \mathrm{var}(X_t) = \sigma^2 \ \forall t$$
$$\gamma_k = \mathrm{cov}(X_t, X_{t+k}) = \mathrm{cov}(X_{t+k}, X_t) = \gamma_{t-(t+k)} = \gamma_{-k},$$

   so it is enough to consider lags $k \geqslant 0$.

   Also note that

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\sigma^2(s)\sigma^2(t)}} = \frac{\gamma_{t-s}}{\sqrt{\gamma_0\gamma_0}} = \frac{\gamma_{t-s}}{\gamma_0} =: \rho_{t-s},$$

   and hence $\rho(t,t) = \rho_0 = 1$.

Definition 2.4 (Autocorrelation function)**.** *The quantity ρ_k is called the* lag k autocorrelation*; ρ_k as a function of k is the* autocorrelation function (acf)*. A plot of ρ_k against k is called a* correlogram

Note that, strictly speaking, it only makes sense to compute the acf for weakly stationary processes. However, we shall see that we can use the acf to diagnose whether a process is stationary or not.

How can we estimate μ, γ_k, ρ_k from data? Normally we would use

$$\operatorname{cov}(X, Y) \approx \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

where X_i, Y_i are i.i.d. copies of X, Y.

For a time series, we typically have only one *realisation* $X_1, X_2, \ldots, X_n$, so we can't estimate $\operatorname{cov}(X_1, X_2)$ as above. However, if $\{X_t\}$ is (weakly) stationary, we can use

$$\begin{aligned}
\widehat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} X_i \approx \mu(X_t) \\
\widehat{\gamma}_k &= \frac{1}{n-k} \sum_{i=1}^{n-k} (X_i - \widehat{\mu})(X_{i+k} - \widehat{\mu}) \approx \gamma_k \\
\widehat{\rho}_k &= \frac{\widehat{\gamma}_k}{\widehat{\gamma}_0}.
\end{aligned}$$

Chapter 3

Models for time series

3.1 White noise

Definition 3.1 (White noise)**.** *The time series* $\{X_t\}$ *is called* white noise *if the* X_t *are i.i.d. with* $\mathbb{E}(X_t) = 0$ *for all* t*.*

If $\{X_t\}$ is white noise, then

$$\mu(t) = \mathbb{E}(X_t) = 0;$$

$$\gamma_k = \operatorname{cov}(X_t, X_{t+k}) = \begin{cases} \sigma^2 & \text{if } k = 0 \\ 0 & \text{else;} \end{cases}$$

$$\rho_k = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{else.} \end{cases}$$

Remarks:

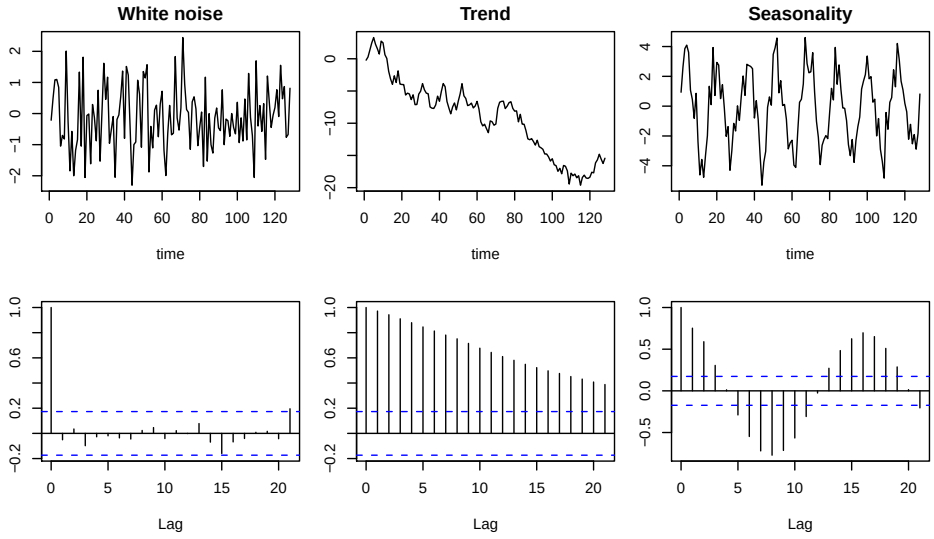
1. Often we assume $X_t \sim N(0, \sigma^2)$ for all t.
2. White noise is used to model the residuals of more complicated time series.
3. We will usually denote a white noise process as $\{\varepsilon_t\}$, with variance σ_ε^2.

We can use the correlogram to distinguish between white noise and processes

with dependence.

Theorem 3.1 (Bartlett, 1946)**.** *If $\{X_t\}$ is white noise, then for large n the distribution of $\widehat{\rho}_k, k \neq 0$, is approximately $N(0, 1/n)$.*

One considers values $|\widehat{\rho}_k| > 1.96/\sqrt{n}$ significant at the 5% level. But note that the $\widehat{\rho}_k$ are not independent of each other!



3.2 Moving average models

Definition 3.2 (Moving average processes)**.** *A stochastic process $\{X_t\}$ is called a* moving average process of order q *(or an $MA(q)$ process) if*

$$X_t = \sum_{k=0}^{q} \beta_k \varepsilon_{t-k} = \beta_0 \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q},$$

where $\beta_0, \ldots, \beta_q \in \mathbb{R}$ and $\{\varepsilon_t\}$ is white noise.

Remarks:

1. Without loss of generality, we can assume $\beta_0 = 1$ (since we can choose $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_t)$).

2. Since $\{\varepsilon_t\}$ is stationary, $\{X_t\}$ is stationary.

For an MA(q) process, we have

$$\mu = \mathbb{E}(X_t) = \mathbb{E}\left(\sum_{k=0}^{q} \beta_k \varepsilon_{t-k}\right) = \sum_{k=0}^{q} \beta_q \mathbb{E}(\varepsilon_{t-k}) = 0,$$

$$\begin{aligned}\gamma_0 = \operatorname{var}(X_t) &= \operatorname{var}\left(\sum_{k=0}^{q} \beta_k \varepsilon_{t-k}\right) = \sum_{k=0}^{q} \beta_k^2 \operatorname{var}(\varepsilon_{t-k}) \\ &= (\beta_0^2 + \cdots + \beta_q^2)\sigma_\varepsilon^2,\end{aligned}$$

$$\begin{aligned}\gamma_k = \operatorname{cov}(X_t, X_{t+k}) &= \operatorname{cov}\left(\sum_{i=0}^{q} \beta_i \varepsilon_{t-i}, \sum_{j=0}^{q} \beta_j \varepsilon_{t+k-j}\right) \\ &= \sum_{i,j=0}^{q} \beta_i \beta_j \operatorname{cov}(\varepsilon_{t-i}, \varepsilon_{t+k-j}).\end{aligned}$$

Now, $\operatorname{cov}(\varepsilon_{t-i}, \varepsilon_{t+k-j})$ is only non-zero (and is σ_ε^2) if $t-i = t+k-j \Leftrightarrow j = k+i$, i.e. there are non-zero terms for

$$(i, j) = (0, k), (1, k+1), \ldots, (q-k, q).$$

Therefore,

$$\gamma_k = \begin{cases} \sum_{i=0}^{q-k} \beta_i \beta_{i+k} \sigma_\varepsilon^2 & \text{if } k \leqslant q \\ 0 & \text{else.} \end{cases}$$

So the acf 'cuts off' at lag q. This is a special feature of MA processes and can be used to recognise data sets for which an MA process is suitable. More on this in chapter 4

Example 3.1 (MA(0), MA(1))**.**

1. *MA(0) is white noise:* $X_t = \beta_0 \varepsilon_t$*, iid* $N(0, \beta_0^2 \sigma_\varepsilon^2)$*.*

2. *MA(1) has $X_t = \varepsilon_t + \beta\varepsilon_{t-1}$ (assuming $\beta_0 = 1$). Therefore,*

$$\begin{aligned} \gamma_0 &= (1+\beta^2)\sigma_\varepsilon^2 & \rho_0 &= 1 \\ \gamma_1 &= \beta\sigma_\varepsilon^2 & \rho_1 &= \frac{\beta}{1+\beta^2} \\ \gamma_k &= 0 \ \ \forall k > 1 & \rho_k &= 0 \ \ \forall k > 1. \end{aligned}$$

We have seen that MA(1) has

$$\rho_0 = 1 \qquad \rho_1 = \frac{\beta}{1+\beta^2} \qquad \rho_2 = \rho_3 = \cdots = 0.$$

Assume we have observed ρ_1 and want to determine β:

$$\begin{aligned} &\rho_1 = \frac{\beta}{1+\beta^2} \\ \Leftrightarrow &\rho_1\beta^2 - \beta + \rho_1 = 0 \\ \Leftrightarrow &\beta_{1,2} = \frac{1 \pm \sqrt{1-4\rho_1^2}}{2\rho_1}. \end{aligned}$$

There are two MA(1) processes with the same autocorrelation function! This problem is referred to as “invertibility”; we shall discuss this in more detail later in section 4.3

3.3 Autoregressive models

Definition 3.3. *A stochastic process $\{X_t\}$ is an AR(p) process if*

$$X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t \qquad \forall t,$$

where $\alpha_1, \ldots, \alpha_p \in \mathbb{R}$, $\{\varepsilon_t\}$ is white noise, and ε_t is independent of X_s for all $s < t$.

Remarks:

1. When constructing the process $\{X_t\}$, the first p values need to be specified as an initial condition.

2. We shall see that whether $\{X_t\}$ is stationary depends on $\alpha_1, \ldots, \alpha_p$ and on the initial conditions.

Example 3.2. *The process $\{X_t\}$ given by $X_0 = 0, X_t = X_{t-1} + \varepsilon_t \; \forall t \in \mathbb{N}$, with $\{\varepsilon_t\}$ a white noise process, is an AR(1) process ($\alpha_1 = 1$). Such a process is called a* random walk. *We have $X_t = \sum_{k=1}^{t} \varepsilon_k$, so*

$$\begin{aligned}\mathbb{E}(X_t) &= \sum_{k=1}^{t} \mathbb{E}(\varepsilon_k) = 0\\ \operatorname{var}(X_t) &= \sum_{k=1}^{t} \operatorname{var}(\varepsilon_k) = t\sigma_\varepsilon^2.\end{aligned}$$

Therefore $\{X_t\}$ is not stationary.

A general AR(1) process has $X_t = \alpha X_{t-1} + \varepsilon_t$, so

$$\begin{aligned}\mathbb{E}(X_t) &= \alpha\mathbb{E}(X_{t-1}) + 0 = \cdots = \alpha^t\mathbb{E}(X_0)\\ \operatorname{var}(X_t) &= \alpha^2 \operatorname{var}(X_{t-1}) + \sigma_\varepsilon^2.\end{aligned}$$

Assume that $\{X_t\}$ is (weakly) stationary and $\alpha \neq 0$. Then

$$\mu = \mathbb{E}(X_t) = \alpha^t\mathbb{E}(X_0) \forall t \Leftrightarrow \mu = 0,$$

so $\mathbb{E}(X_0) = 0$. Also,

$$\begin{aligned}\sigma^2 &= \operatorname{var}(X_t) = \alpha^2\sigma^2 + \sigma_\varepsilon^2\\ \Leftrightarrow &(1-\alpha^2)\sigma^2 = \sigma_\varepsilon^2\\ \Leftrightarrow &\sigma^2 = \frac{\sigma_\varepsilon^2}{1-\alpha^2},\end{aligned}$$

which is only possible for $|\alpha| < 1$.

Thus an AR(1) process is weakly stationary if and only if $|\alpha| < 1$, $\mathbb{E}(X_t) = 0 \,\forall\, t$, and $\operatorname{var}(X_t) = \sigma_\varepsilon^2/(1-\alpha^2) \;\forall\; t$ including $t = 0$.

Example 3.3. *Consider the AR(1) process with $\alpha_1 = -0.8$. Note that $|\alpha_1| < 1$; necessary but not sufficient for weak stationarity.*

- *Let $X_t = -0.8X_{t-1} + \varepsilon_t$ with $\varepsilon_t iid N(0,1)$ and fix $X_0 = 0$. This is not stationary since $\operatorname{var}(X_0) = 0 \neq 1/0.36 = \operatorname{var}(X_t)$ for $t \gg 0$.*

- *Now let* $X_0 \sim N(0, 1/0.36)$. *Now* $|\alpha_1| < 1$, $\mathbb{E}(X_t) = 0$ *and* $\mathrm{var}(X_t) = 1/0.36$ *for all* t; $\{X_t\}$ *is weakly stationary.*

Proposition 3.1. *An AR(p) process is stationary if and only if all the roots* $y_1, \ldots, y_p$ *of the equation*

$$\alpha(y) = 1 - \alpha_1 y - \cdots - \alpha_p y^p$$

are such that $|y_i| > 1$.

Remark: Note that for the AR(1) process

$$\alpha(y) = 1 - \alpha_1 y = 0 \Leftrightarrow y = 1/\alpha$$

and

$$|y_1| > 1 \Leftrightarrow |\alpha| < 1$$

as we saw above.

Proof. We prove proposition 3.1 in three steps:

1. Write X_t in terms of $\alpha(\cdot)^{-1}$.
2. Write X_t in terms of $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$ — infinite recursion.
3. Examine coefficients $\{c_k\}$ of $\{\varepsilon_{t-k}\}$ to see when stationarity is possible.

step 1/3 Define the *backshift operator* B by

$$(BX)_t = X_{t-1} \qquad \forall t$$

(shift the process by one time step). Then

$$(B^2 X)_t = (BBX)_t = (BX)_{t-1} = X_{t-2}$$
$$(B^3 X)_t = X_{t-3}$$

etc. Thus

$$\begin{aligned} X_t &= \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t \\ &= \alpha_1 B X_t + \alpha_2 B^2 X_t + \cdots + \alpha_p B^p X_t + \varepsilon_t \\ \Rightarrow \quad & (1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p) X_t = \varepsilon_t \\ \Rightarrow X_t &= (1 - \alpha_1 B - \cdots - \alpha_p B^p)^{-1} \varepsilon_t \\ &= \alpha(B)^{-1} \varepsilon_t. \end{aligned}$$

step 2/3 To get rid of the *inverse* in $\alpha(B)^{-1}$, we shall find numbers c_k such that

$$\frac{1}{\alpha(y)} = \sum_{k=0}^{\infty} c_k y^k \qquad \forall y \in \mathbb{R}.$$

Then replace y by B to get

$$X_t = \alpha(B)^{-1} \varepsilon_t = \sum_{k=0}^{\infty} c_k B^k \varepsilon_t.$$

To see how to do this, we consider specific cases $p = 1, 2$ before the general case.

- For AR(1), $\alpha(B) = 1 - \alpha_1 B$, so

$$\frac{1}{1 - \alpha_1 y} = \sum_{k=0}^{\infty} (\alpha_1 y)^k = \sum_{k=0}^{\infty} \alpha_1^k y^k.$$

- For AR(2), we use a "partial fraction decomposition". Note that

$$\alpha(y) = 1 - \alpha_1 y - \alpha_2 y^2 = 0$$

has (up to) two (complex) roots $y_1, y_2 \neq 0$. Write $\mu_i = 1/y_i, i = 1, 2$. Then

$$\alpha(y) = (1 - \mu_1 y)(1 - \mu_2 y).$$

Now we assume

$$\frac{1}{(1 - \mu_1 y)(1 - \mu_2 y)} = \frac{A_1}{1 - \mu_1 y} + \frac{A_2}{1 - \mu_2 y}$$

and try to find A_1, A_2:

$$\frac{1}{(1-\mu_1 y)(1-\mu_2 y)} = \frac{A_1(1-\mu_2 y) + A_2(1-\mu_1 y)}{(1-\mu_1 y)(1-\mu_2 y)}$$
$$\Rightarrow \begin{cases} A_1 + A_2 &= 1 \\ \mu_2 A_1 + \mu_1 A_2 &= 0; \end{cases}$$

two equations with two unknowns — so can be solved.

The same procedure works for AR(p); $\alpha(y)$ is a polynomial of degree p with roots $y_1, y_2, \ldots, y_p$ and there are numbers $A_1, \ldots, A_p \in \mathbb{C}$ with

$$\frac{1}{\alpha(y)} = \frac{A_1}{1 - y/y_1} + \cdots + \frac{A_p}{1 - y/y_p}.$$

By the geometric series formula

$$\begin{aligned} \frac{1}{\alpha(y)} &= A_1 \sum_{k=0}^{\infty} \frac{1}{y_1^k} y^k + \cdots + A_p \sum_{k=0}^{\infty} \frac{1}{y_p^k} y^k \\ &= \sum_{k=0}^{\infty} \underbrace{(A_1 y_1^{-k} + \cdots + A_p y_p^{-k})}_{c_k} y^k \\ \Rightarrow X_t = \alpha(B)^{-1}\varepsilon_t &= \sum_{k=0}^{\infty} c_k B^k \varepsilon_t = \sum_{k=0}^{\infty} c_k \varepsilon_{t-k}. \end{aligned}$$

step 3/3

Assume $\{X_t\}$ is weakly stationary (and defined for all times $t \in \mathbb{Z}$). Then

$$\operatorname{var}(X_t) = \operatorname{var}\left(\sum_{k=0}^{\infty} c_k \varepsilon_{t-k}\right) = \left(\sum_{k=0}^{\infty} c_k^2\right) \sigma_\varepsilon^2.$$

For the variance to exist, we need $\sum_{k=0}^{\infty} c_k^2 < \infty$. We have seen

$$c_k = A_1 \frac{1}{y_1^k} + \cdots + A_p \frac{1}{y_p^k},$$

where $y_i, i = 1, \ldots, p$ are the roots of $\alpha(y)$. If any of the y_i has $|y_i| \leqslant 1$, then c_k^2 does not go to zero and hence $\sum c_k^2 = \infty$. If all $|y_i| > 1$ then $c_k \downarrow 0$ exponentially and then $\sum c_k^2 < \infty$.

In other words, an AR(p) process is stationary

$\Leftrightarrow$ $|y_i| > 1$ for $i = 1, \ldots, p$,

$\Leftrightarrow$ all roots of $\alpha(y) = 1 - \alpha_1 y - \cdots - \alpha_p y^p$ lie outside the complex unit circle.

□

Example 3.4 (AR(2))**.** *Consider*

$$X_t = \frac{1}{2}X_{t-1} + \frac{1}{2}X_{t-2} + \varepsilon_t.$$

The roots are $1, -2$ *(exercise: check), so this process is* not *stationary.*

Remarks:

1. In case of stationarity, we have

$$\mathbb{E}(X_t) = \mathbb{E}\left(\alpha(B)^{-1}\varepsilon_t\right) = \mathbb{E}\left(\sum_{k=0}^{\infty} c_k B^k \varepsilon_t\right) = \sum_k c_k \mathbb{E}(\varepsilon_{t-k}) = 0.$$

2. The roots of $\alpha()$ can be complex. For example, consider $X_t = -X_{t-2} + \varepsilon_t$. This has $\alpha(y) = 1 - (-1)y^2 = 1 + y^2$, with roots $y_1 = i, y_2 = -i$. (Hence $\{X_t\}$ is not stationary — Exercise: check this by working out $\mathrm{var}(X_t)$).

For AR(2) we have

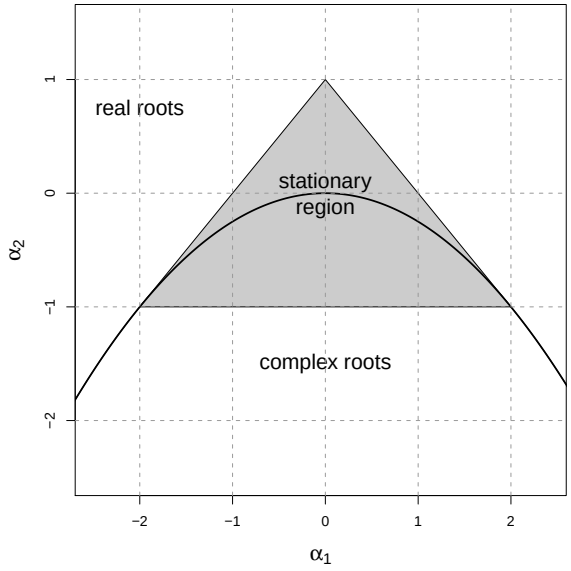
$$\begin{aligned} \alpha(y) &= 1 - \alpha_1 y - \alpha_2 y^2 \\ \Rightarrow \text{roots} \quad y_{1,2} &= \frac{\alpha_1 \pm \sqrt{\alpha_1^2 + 4\alpha_2}}{-2\alpha_2}. \end{aligned}$$

If $\alpha_1^2 + 4\alpha_2 > 0$ we have two real roots.

If $\alpha_1^2 + 4\alpha_2 < 0$ we have two complex roots. In this case,

$$\begin{aligned} y_{1,2} &= \frac{\alpha_1 \pm i\sqrt{-4\alpha_2 - \alpha_1^2}}{-2\alpha_2} \\ \Rightarrow |y_{1,2}|^2 &= \underbrace{\frac{\alpha_1^2}{4\alpha_2^2}}_{\text{Re}^2} + \underbrace{\frac{-4\alpha_2 - \alpha_1^2}{4\alpha_2^2}}_{\text{Im}^2}. \end{aligned}$$

One can check that the process is stationary if $\alpha_2 > -1$, $\alpha_2 < 1 - \alpha_1$ and $\alpha_2 < 1 + \alpha_1$.



For AR(p), $p > 2$, we use a computer to find the roots. Eg the R command `polyroot` which takes the argument $(1, -\alpha_1, \ldots, -\alpha_p)$. Using

```
> polyroot(c(1, -0.5, -0.5))
```

will verify example 3.4.

Auto-covariances Recall $\gamma_k = \mathrm{cov}(X_t, X_{t+k})$

Since

$$X_{t+k} = \alpha_1 X_{t+k-1} + \cdots + \alpha_p X_{t+k-p} + \varepsilon_{t+k},$$

we have

$$\begin{aligned}
\gamma_k &= \mathrm{cov}(X_t, X_{t+k}) \\
&= \mathrm{cov}(X_t, \alpha_1 X_{t+k-1} + \cdots + \alpha_p X_{t+k-p} + \varepsilon_{t+k}) \\
&= \alpha_1 \,\mathrm{cov}(X_t, X_{t+k-1}) + \cdots + \alpha_p \,\mathrm{cov}(X_t, X_{t+k-p}) + \mathrm{cov}(X_t, \varepsilon_{t+k}) \\
&= \alpha_1 \gamma_{k-1} + \cdots + \alpha_p \gamma_{k-p} + 0 \\
&= \sum_{j=1}^{p} \alpha_j \gamma_{k-j} \qquad\qquad \forall k \geqslant 1.
\end{aligned}$$

Hence the autocorrelations $\rho_k = \gamma_k/\gamma_0$ are

$$\rho_k = \sum_{j=1}^{p} \alpha_j \rho_{k-j}.$$

Example 3.5 (AR(2))**.** *We have*

$$\rho_k = \alpha_1 \rho_{k-1} + \alpha_2 \rho_{k-2} \qquad\qquad k = 1, 2, \ldots,$$

so

$$\begin{aligned}
\rho_0 &= 1 \\
\rho_1 &= \alpha_1 \rho_0 + \alpha_2 \rho_{-1} = \alpha_1 + \alpha_2 \rho_1 = \frac{\alpha_1}{1-\alpha_2} \\
\rho_2 &= \alpha_1 \rho_1 + \alpha_2 \rho_0 = \frac{\alpha_1^2}{1-\alpha_2} + \alpha_2 \\
&\vdots
\end{aligned}$$

In general, we can determine $\alpha_1, \ldots, \alpha_p$ from $\rho_1, \ldots, \rho_p$ using

$$\left.\begin{aligned}
\rho_1 &= 1 \cdot \alpha_1 + \rho_1 \alpha_2 + \cdots + \rho_{p-1}\alpha_p \\
\rho_2 &= \rho_1 \alpha_1 + 1 \cdot \alpha_2 + \cdots + \rho_{p-2}\alpha_p \\
&\vdots \\
\rho_p &= \rho_{p-1}\alpha_1 + \rho_{p-2}\alpha_2 + \cdots + 1 \cdot \alpha_p.
\end{aligned}\right\} \tag{3.1}$$

We can estimate $\rho_1, \ldots, \rho_p$ from data and then solve these p equations to get estimates $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_p$.

The equations (3.1) are called the *Yule-Walker* equations. In matrix notation they are

$$\begin{bmatrix}
1 & \rho_1 & \cdots & \cdots & \rho_{p-1} \\
\rho_1 & 1 & \ddots & & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & & \ddots & 1 & \rho_1 \\
\rho_{p-1} & \cdots & \cdots & \rho_1 & 1
\end{bmatrix}
\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix}
=
\begin{bmatrix} \rho_1 \\ \vdots \\ \rho_p \end{bmatrix}.$$

Example 3.6 (AR(2))**.** *Consider an AR(2) process with* $\rho_0 = 1, \rho_1 = 1/6, \rho_2 =$

$-11/24$. *We get the Yule-Walker equations*

$$\alpha_1 + \frac{1}{6}\alpha_2 = \frac{1}{6}$$
$$\frac{1}{6}\alpha_1 + \alpha_2 = -\frac{11}{24}.$$

Solving these equations for α_1 *and* α_2 *yields*

$$\alpha_1 = \frac{1}{4} \qquad\qquad \alpha_2 = -\frac{1}{2}$$

(exercise — check).

We can confirm this in R:

```
> acf2AR(acf=c(1, 1/6, -11/24))
              1    2
ar(1) 0.1666667  0.0
ar(2) 0.2500000 -0.5
```

To fit an AR(p) model to data $\{X_t\} = (X_1, \ldots, X_n)$, we use the following steps.

1. Subtract trend and seasonal effects from $\{X_t\}$ to obtain residuals $\{Y_t\}$.
2. Estimate the a.c.f. of Y to obtain $\widehat{\rho}_1, \ldots, \widehat{\rho}_p$.
3. Solve the Yule-Walker equations to obtain $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_p$.
4. Consider the residuals $Z_t = Y_t - \widehat{\alpha}_1 Y_{t-1} - \cdots - \widehat{\alpha}_p Y_{t-p}$. Use the Bartlett bands to check whether the $\{Z_t\}$ are (approximately) white noise (otherwise the model is not a good fit to the data).
5. Use the sample variance of the $\{Z_t\}$ to estimate σ_ε^2.
6. Add trend and seasonal effects back on to conclusions about $\{Y_t\}$ to get conclusions for $\{X_t\}$.

3.4 Mixed autoregressive moving average (ARMA) models

Definition 3.4 (ARMA(p, q) model)**.** *The ARMA(p, q) model satisfies*

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j}, \tag{3.2}$$

with ε_t independent of $X_{t-1}, X_{t-2}, \ldots$.

Remarks.

1. Equation (3.2) can be written as

   $$\alpha(B) X_t = \beta(B) \varepsilon_t,$$

   where

   $$\alpha(y) = 1 - \sum_{i=1}^{p} \alpha_i y^i, \qquad \beta(y) = 1 + \sum_{j=1}^{q} \beta_j y^j.$$

2. As for AR(p), we can write

   $$X_t = \alpha(B)^{-1} \beta(B) \varepsilon_t.$$

   The model is weakly stationary if

   $$X_t = \sum_{k=0}^{\infty} \lambda_k \beta(B) \varepsilon_{t-k} = \sum_{k=0}^{\infty} \tilde{\lambda}_k \varepsilon_{t-k}$$

   with $\sum_{k=0}^{\infty} \tilde{\lambda}_k^2 < \infty$. This is again equivalent to the roots of α lying outside the complex unit circle.
3. If stationary, we have $\mathbb{E}(X_t) = \sum_{k=0}^{\infty} \tilde{\lambda}_k \mathbb{E}(\varepsilon_{t-k}) = 0$.
4. We can reconstruct the noise as

   $$\varepsilon_t = \beta(B)^{-1} \alpha(B) X_t = \sum_{k=0}^{\infty} \delta_k X_{t-k}$$

   If $\sum_{k=0}^{\infty} \delta_k^2 < \infty$, the process is called *invertible* (more on invertibility later in section 4.3). In this case, the influence of X_{t_0} (t_0 fixed) on X_t

gets smaller as $t \to \infty$.

Result $\{X_t\}$ is invertible iff the roots of β lie outside the complex unit circle.

Example 3.7 (ARMA(1,1))**.** *Consider the ARMA(1,1) process*

$$X_t = \alpha X_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1}.$$

Stationary iff $|\alpha| < 1$.

Invertible iff roots of $1 + \beta y$ *lie outside the unit circle, iff* $|\beta| < 1$.

Auto-covariances of ARMA

$$\begin{aligned}
\gamma_k &= \operatorname{cov}(X_t, X_{t+k}) \\
&= \sum_{i=1}^{p} \alpha_i \underbrace{\operatorname{cov}(X_t, X_{t+k-i})}_{\gamma_{k-i}} + \underbrace{\operatorname{cov}(X_t, \varepsilon_{t+k})}_{=0 \text{ for } k>0} + \underbrace{\sum_{j=1}^{q} \beta_j \operatorname{cov}(X_t, \varepsilon_{t+k-j})}_{=0 \text{ for } k>q} \\
\Rightarrow \gamma_k &= \sum_{i=1}^{p} \alpha_i \gamma_{k-i} \qquad \forall k > q \\
\Rightarrow \rho_k &= \sum_{i=1}^{p} \alpha_i \rho_{k-i} \qquad \forall k > q.
\end{aligned}$$

Reminder: AR(p) satisfies the same relationship (but for all $k > 0$). Hence the a.c.f. for AR and ARMA show the same behaviour for $k > q$.

3.5 Integrated models

Let $\varepsilon_t \sim N(0,1)$ be i.i.d. and $X_t = \sum_{s=1}^{t} \varepsilon_s \ \ \forall t \in \mathbb{N}$. This process has no drift since $\mathbb{E}(X_t) = 0$, but since

$$\operatorname{var}(X_t) = \sum_{s=1}^{t} \operatorname{var}(\varepsilon_s) = t,$$

the process is *not* stationary. (Note that $\{X_t\}$ is AR(1) with $\alpha_1 = 1$; since $\{X_t\}$ has a unit root it is non-stationary.) Hence $\{X_t\}$ cannot easily be described by our models so far.

Definition 3.5 (Difference operator)**.** *Let*

$$\nabla X_t = X_t - X_{t-1} \quad \forall t.$$

The operator ∇ ("nabla") is called the difference operator*.*

Remarks.

1. We can write $\nabla = 1 - B$ where B is the backshift operator:

$$(1 - B)X_t = X_t - BX_t = X_t - X_{t-1} = \nabla X_t.$$

2. If $\{X_t\}$ has stationary increments, then ∇X is stationary.
3. ∇ removes a constant mean:

$$\nabla(X_t + \mu) = (X_t + \mu) - (X_{t-1} + \mu) = X_t - X_{t-1} = \nabla X_t.$$

   A linear trend is converted to a constant mean:

$$\begin{aligned} \nabla(X_t + \alpha + \beta t) &= (X_t + \alpha + \beta_t) - (X_{t-1} + \alpha + \beta(t-1)) \\ &= X_t - X_{t-1} + \beta \\ &= \nabla X_t + \beta. \end{aligned}$$

Definition 3.6 (ARIMA(p,d,q))**.** *$\{X_t\}$ is an ARIMA(p,d,q) process if $\nabla^d X$ is a stationary ARMA(p,q) process ("I" $\sim$ "integrated").*

Remark Let $Y = \nabla^d X = (1-B)^d X$ be ARMA(p, q). Then

$$\alpha(B)Y = \beta(B)\varepsilon \qquad \Rightarrow \qquad \underbrace{\alpha(B)(1-B)^d}_{= \alpha^*(B),\ \text{degree } p+d} \quad X = \beta(B)\varepsilon$$

Thus an ARIMA(p, d, q) process can be written as an ARMA($p + d, q$) process. Since α^* has 1 as a root, $\{X_t\}$ cannot be stationary for $d > 0$.

Example 3.8. *Let $\varepsilon_t \sim N(0, \sigma^2)$ and $X_t = \sum_{s=1}^{t} \varepsilon_s$ (random walk).*

Then $\nabla X_t = X_t - X_{t-1} = \varepsilon_t$ is white noise, i.e. $\{\varepsilon_t\}$ is an ARMA(0,0) process. Hence $\{X_t\}$ is ARIMA(0,1,0).

Chapter 4

Model fitting

Assume we have a stationary time series $X_1, X_2, \ldots, X_n$, with mean zero. How do we fit one of the models from the previous chapter to these data? This problem has two components: model selection and parameter estimation.

4.1 Model selection

How to choose between AR, MA, ARMA, and ARIMA?

step 1 Estimate ρ_k. We know

Model	ρ_k
MA(q)	$\rho_k = 0 \quad \forall k > q$
AR(p)	$\rho_k = \alpha_1 \rho_{k-1} + \cdots + \alpha_p \rho_{k-p}$
ARMA(p, q)	As for AR(p), except for the first q values

For AR(p) and ARMA(p, q) we have $\rho_k = \alpha_1 \rho_{k-1} + \cdots + \alpha_p \rho_{k-p}$. How does ρ_k behave? We can find a non-recursive formula for ρ_k.

Lemma 4.1. *Let $y_1, \ldots, y_p$ be the roots of $\alpha(y) = 1 - \alpha_1 y - \cdots - \alpha_p y^p$. Then*

for AR(p),

$$\rho_k = \sum_{j=1}^{p} \frac{c_j}{y_j^k},$$

for suitable constants $c_1, \ldots, c_p$.

Proof. Starting with this definition for $\rho_{k-1}, \ldots, \rho_{k-p}$, we have

$$\begin{aligned}
\alpha_1\rho_{k-1} + \cdots + \alpha_p\rho_{k-p} &= \sum_{j=1}^{p} \left(\alpha_1 \frac{c_j}{y_j^{k-1}} + \cdots + \alpha_p \frac{c_j}{y_j^{k-p}} \right) \\
&= \sum_{j=1}^{p} \frac{c_j}{y_j^k} \underbrace{\left(\alpha_1 y_j + \cdots + \alpha_p y_j^p \right)}_{=1} \\
&= \sum_{j=1}^{p} \frac{c_j}{y_j^k} \cdot 1 = \rho_k \qquad (4.1)
\end{aligned}$$

The coefficients c_j can be found from $\rho_1, \ldots, \rho_{p-1}$. □

How to recognise solutions to (4.1)?

AR(1) $\rho_k = \alpha^k$. Hence

exponential decay ($\alpha > 0$) or
exponential decay with alternating signs ($\alpha < 0$).

AR(p), $p > 1$ Damped oscillations with

$$\rho_k = c^k \cos(\phi k).$$

for some constants ϕ, c, $|c| < 1$.

ARIMA(p, d, q) Processes can be recognised by the fact that ρ_k decays very slowly. The order d is found by differencing (i.e. considering $\nabla X, \nabla^2 X, \ldots$) until ρ_k indicates AR, MA, or ARMA.

To distinguish between AR(p) and ARMA(p, q) we use the following quantity.

Definition 4.1 (partial autocorrelation)**.** *For* $k > 0$*, use the Yule-Walker equations to fit an AR(k) model (using* $\rho_1, \ldots, \rho_k$*), obtaining coefficients* $\alpha_{k1}, \ldots, \alpha_{kk}$*. Then* α_{kk} *is called the* lag-k partial autocorrelation *(pacf) of* $\{X_t\}$.

One can show that the pacf satisfies

Model	α_{kk}
AR(p)	cut off after lag p
MA(q), ARMA(p, q)	exponential decay / damped oscillations

The pacf α_{kk} can be approximated by the sample pacf $\widehat{\alpha}_{kk}$, computed from $\widehat{\rho}_1, \ldots, \widehat{\rho}_k$ instead of $\rho_1, \ldots, \rho_k$.

Where ρ_k and α_{kk} cut off, the sample quantities $\widehat{\rho}_k$ and $\widehat{\alpha}_{kk}$ only cut off approximately.

Results

- if $X_1, \ldots, X_n$ is MA(q), then for $k > q$

$$\widehat{\rho}_k \sim N\left(0, \frac{1}{n}\left[1 + 2\sum_{i=1}^{q} \widehat{\rho}_i^2\right]\right).$$

- If $X_1, \ldots, X_n$ is AR(p), then for $k > p$

$$\widehat{\alpha}_{kk} \sim N\left(0, \frac{1}{n}\right).$$

4.2 Parameter estimation

So far, we have assumed our stationary processes are zero mean. We can cope with a non-zero mean μ for $\{X_t\}$ by modelling $Y_t = X_t - \mu$. For example, our AR(1) model becomes

$$X_t - \mu = \alpha(X_t - \mu) + \varepsilon_t \qquad\qquad t = 1, 2, \ldots, n.$$

We estimate μ by $\widehat{\mu} = n^{-1}\sum_{t=1}^{n} X_t$, subtract $\widehat{\mu}$ from the observations to estimate $\{Y_t\}$ and proceed with fitting a model to the $\{Y_t\}$.

There are several methods to estimate parameters for AR, MA and ARMA models.

4.2.1 Method of moments

Choose parameters such that the means and correlations for model and sample coincide. For AR(p) models, this gives the Yule-Walker equations (3.1).

Example 4.1 (AR(1))**.** *From the YW equations,* $\widehat{\alpha}_1 = \widehat{\rho}_1$*. We can estimate* σ_ε^2 *via the method of moments:*

$$X_t = \alpha_1 X_{t-1} + \varepsilon_t \qquad \Rightarrow \qquad \operatorname{var}(X_t) = \alpha_1^2 \operatorname{var}(X_{t-1}) + \operatorname{var}(\varepsilon_t).$$

Assuming $\{X_t\}$ *is stationary,*

$$\sigma_X^2 = \alpha_1^2 \sigma_X^2 + \sigma_\varepsilon^2 \qquad \rightarrow \qquad \widehat{\sigma}_\varepsilon^2 = \widehat{\sigma}_X^2 (1 - \widehat{\alpha}_1^2).$$

If $\{X_t\}$ *is mean zero, we can estimate* σ_X^2 *by the sample variance* s_x^2*.*

4.2.2 Least squares method (LSE)

Choose parameters to minimise the residual sum of squares

Example 4.2 (AR(1))**.** *We estimate* α *by minimising*

$$S(\alpha) = \sum_{t=1}^{n} (X_t - \alpha X_{t-1})^2 .$$

Since X_0 *is usually unknown, we set it to zero. Hence*

$$\begin{aligned} 0 = \frac{\partial}{\partial \alpha} S(\widehat{\alpha}) &= 2 \sum_{t=1}^{n} (X_t - \widehat{\alpha} X_{t-1}) (-X_{t-1}) \\ &= -2 \sum_{t=2}^{n} X_t X_{t-1} + 2\widehat{\alpha} \sum_{t=2}^{n} X_{t-1}^2 \\ \Leftrightarrow \qquad \widehat{\alpha} &= \frac{\frac{1}{n-1} \sum_{t=2}^{n} X_t X_{t-1}}{\frac{1}{n-1} \sum_{t=2}^{n} X_{t-1}^2} \\ &\approx \widehat{\rho}_1. \end{aligned}$$

Given $\widehat{\alpha}$, we can estimate σ_ε^2 by

$$\begin{aligned}\widehat{\sigma}_\varepsilon^2 &= \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \widehat{\alpha}X_{t-1})^2 \\ &= \underbrace{\frac{1}{n-1}\sum_{t=1}^{n}X_t^2}_{\approx s_X^2} - 2\widehat{\alpha}\underbrace{\frac{1}{n-1}\sum_{t=1}^{n}X_tX_{t-1}}_{\approx\widehat{\gamma}_1\approx s_X^2\widehat{\rho}_1\approx s_X^2\widehat{\alpha}} + \widehat{\alpha}^2\underbrace{\frac{1}{n-1}\sum_{t=1}^{n}X_{t-1}^2}_{\approx s_X^2} \\ &\approx s_X^2(1-\widehat{\alpha}^2)\end{aligned}$$

— the same as MOM.

4.2.3 Maximum likelihood estimation (MLE)

To use this method, we need to assume a distribution for the ε_t.

Example 4.3 (AR(1))**.** *Assume $X_0 = 0, X_t = \alpha X_{t-1} + \varepsilon_t, \varepsilon_t \sim N(0, \sigma^2)$.*

The density of $X_t|X_{t-1}$ is

$$f(X_t|X_{t-1}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(X_t - \alpha X_{t-1})^2}{2\sigma^2}\right\},$$

so the joint density of $X_1, X_2, \ldots, X_n$ is

$$\begin{aligned}f(X_1,\ldots,X_n) &= f(X_1)f(X_2|X_1)\cdots f(X_n|X_{n-1}) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\sum_{t=1}^{n}\frac{(X_t - \alpha X_{t-1})^2}{2\sigma^2}\right\}.\end{aligned}$$

To find the MLE of α, we need to maximise the log-likelihood

$$\begin{aligned}l(\alpha) = \log L(\alpha) &= \log f(X_1,\ldots,X_n) \\ &= \textit{const} - \frac{1}{2\sigma^2}\underbrace{\sum_{t=1}^{n}(X_t - \alpha X_{t-1})^2}_{S(\alpha)}.\end{aligned}$$

Thus, we need to minimise $S(\alpha)$ as above. Therefore, for the AR(1) process with normally distributed innovations, the MLE and LSE are the same.

4.3 MOM and invertibility for MA(q)

4.3.1 MA(1)

Recall that the MA(1) process $X_t = \varepsilon_t + \beta\varepsilon_{t-1}$ where ε is a white noise process with variance σ_ε^2 has acf

$$\rho_1 = \frac{\beta}{1+\beta^2} \qquad \rho_k = 0 \;\; k = 2, 3, \ldots.$$

Estimating β by MOM is not possible since solving $\widehat{\rho}_1 = \widehat{\beta}/(1+\widehat{\beta}^2)$ for $\widehat{\beta}$ yields

$$\widehat{\beta} = \frac{1 \pm \sqrt{1 - 4\widehat{\rho}_1^2}}{2\widehat{\rho}_1},$$

which has a corresponding effect on estimating σ_ε^2 since the obvious estimator is

$$\widehat{\sigma}_\varepsilon^2 = (1+\widehat{\beta}^2)^{-1}\widehat{\sigma}_X^2 = (1+\widehat{\beta}^2)^{-1}s_X^2.$$

How can we distinguish between these solutions? To motivate a choice, recall that we can write X_t as $X_t = (1+\beta B)\varepsilon_t$ where B is the backshift operator. Therefore,

$$\begin{aligned}
\varepsilon_t &= (1+\beta B)^{-1}X_t \\
&= (1 - \beta B + \beta^2 B - \beta^3 B + \cdots)X_t \\
&= X_t - \beta X_{t-1} + \beta^2 X_{t-2} - \beta^3 X_{t-3} + \cdots \\
\Rightarrow X_t &= \varepsilon_t + \beta X_{t-1} - \beta^2 X_{t-2} + \beta^3 X_{t-3} - \cdots,
\end{aligned}$$

an AR(∞) representation. When does this make sense?

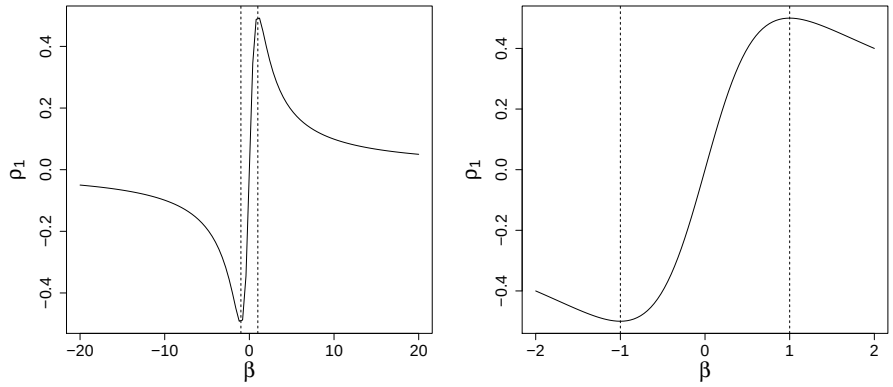
1. $|\beta| > 1$ — distant (infinite) past has more influence than the recent past.
2. $|\beta| < 1$ — past influence decays exponentially.

Case 1 is unrealistic so requiring $|\beta| < 1$ seems sensible. This condition is

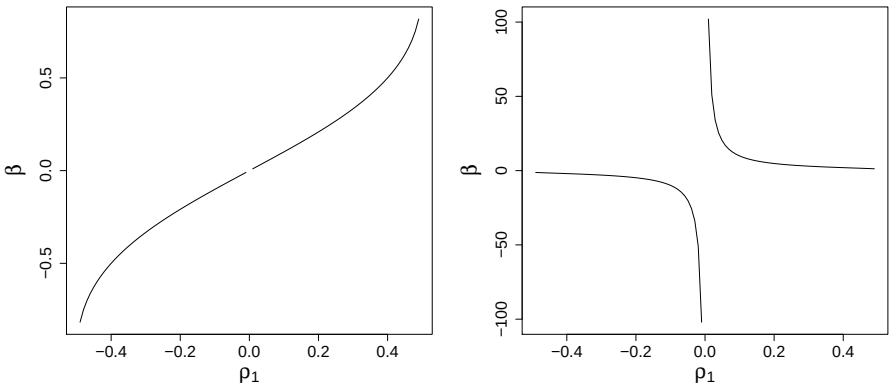
called *invertibility*; requiring it means we use

$$\widehat{\beta} = \frac{1 - \sqrt{1 - 4\widehat{\rho}_1^2}}{2\widehat{\rho}_1}$$

since then $|\widehat{\beta}| < 1$.



Invertible Non-invertible



4.3.2 MA(q)

For a general MA(q) process $X_t = \beta(B)\varepsilon$ where

$$\beta(y) = 1 + \beta_1 y + \beta_2 y^2 + \cdots + \beta_q y^q,$$

suppose $\beta(y) = 0$ has roots $y_1^{-1}, \ldots, y_q^{-1}$. Then

$$\beta(B) = (1 - By_1)(1 - By_2) \cdots (1 - By_q),$$

so (as in section 3.3)

$$\begin{aligned}
\varepsilon_t &= \beta(B)^{-1} X_t \\
&= \prod_{k=1}^{q} (1 - By_k)^{-1} X_t \\
&= \sum_{k=1}^{q} \frac{c_k}{1 - By_k} X_t && \text{by partial fractions} \\
&= \sum_{k=1}^{q} c_k \left(\sum_{i=0}^{\infty} y_k^i B^i \right) X_t \\
&= \sum_{i=0}^{\infty} h_i X_{t-i}
\end{aligned}$$

for $h_i = \sum_{k=1}^{q} c_k y_k^i$.

If any $|y_k| > 1$, then h_i is increasing and X_t depends on the infinite past. We say that $\{X_t\}$ is invertible if all the roots of $\beta(y)$ lie outside the unit circle; imposing this condition means that the parameters $\beta_1, \ldots, \beta_q$ are uniquely estimable from the ρ_k.

There are 2^q processes with the same acf, only one of which is invertible.

Chapter 5

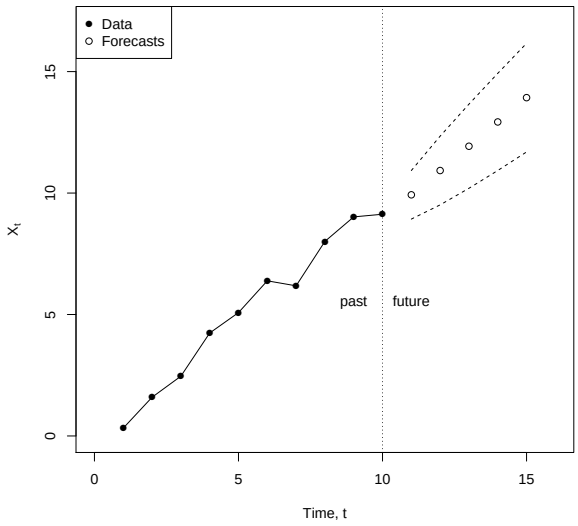
Forecasting

Earlier we spoke about predicting future values (*forecasting*) in terms of a fitted trend or seasonal effect. Now we consider the problem of forecasting future values of a stochastic process.

5.1 Introduction

Given data values $X_1, X_2, \ldots, X_n$, we want

- To predict a future value X_{n+l} which is l steps ahead, by a forecast $X_n(l)$ say, where l is the *lead time* of the forecast.
- To have confidence limits for $X_n(l)$.
- To ensure the *forecast error* $e_n(l) = X_{n+l} - X_n(l)$ is as small as possible.



5.2 Minimum mean square

5.2.1 Forecasting using minimum mean square error

We want the *forecast error* $e_n(l) = X_{n+l} - X_n(l)$ to be small, so we might choose $X_n(l)$ to minimise

$$S = \mathbb{E}\{[X_{n+l} - X_n(l)]^2\},$$

the mean square error between the forecast $X_n(l)$ and the true future value X_{n+l}.

Suppose X_t can be written using a (possibly infinite) MA model

$$\begin{aligned} X_t &= \beta_0\varepsilon_t + \beta_1\varepsilon_{t-1} + \beta_2\varepsilon_{t-2} + \beta_3\varepsilon_{t-3} + \cdots \\ &= \sum_{k=0}^{\infty} \beta_k\varepsilon_{t-k}, \end{aligned}$$

where $\{\varepsilon_t\}$ is a white noise process with variance σ_ε^2. Thus

$$X_{n+l} = \beta_0\varepsilon_{n+l} + \beta_1\varepsilon_{n+l-1} + \beta_2\varepsilon_{n+l-2} + \cdots + \beta_{l-1}\varepsilon_{n+1} + \beta_l\varepsilon_n + \beta_{l+1}\varepsilon_{n-1} + \beta_{l+2}\varepsilon_{n-2} + \cdots . \tag{5.1}$$

Suppose we also want to write $X_n(l)$ as a linear combination of the innovations $\{\varepsilon_t\}$ up to time n; let

$$\begin{aligned} X_n(l) &:= \beta'_0\varepsilon_n + \beta'_1\varepsilon_{n-1} + \beta'_2\varepsilon_{n-2} + \beta'_3\varepsilon_{n-3} + \cdots \\ &= \sum_{s=0}^{\infty} \beta'_s\varepsilon_{n-s}. \end{aligned}$$

Then the error in our forecast is

$$\begin{aligned} e_n(l) &= X_{n+l} - X_n(l) \\ &= \beta_0\varepsilon_{n+l} + \beta_1\varepsilon_{n+l-1} + \cdots + \beta_{l-1}\varepsilon_{n+1} \\ &\quad + (\beta_l - \beta'_0)\varepsilon_n + (\beta_{l+1} - \beta'_1)\varepsilon_{n-1} + (\beta_{l+2} - \beta'_2)\varepsilon_{n-2} + \cdots . \end{aligned} \tag{5.2}$$

Recall that $\mathbb{E}(\varepsilon_t\varepsilon_s) = 0$ for $t \neq s$ and σ_ε^2 for $t = s$. Then squaring (5.2) and taking expectations gives

$$\mathbb{E}[\{X_{n+l} - X_n(l)\}^2] = \sigma_\varepsilon^2 \left(\sum_{k=0}^{l-1} \beta_k^2 + \sum_{s=0}^{\infty} (\beta_{l+s} - \beta'_s)^2 \right).$$

This is minimised by choosing $\beta'_s = \beta_{l+s}$. The minimum mean square error forecast is thus

$$\begin{aligned} X_n(l) &= \beta_l\varepsilon_n + \beta_{l+1}\varepsilon_{n-1} + \beta_{l+2}\varepsilon_{n-2} + \cdots \\ &= \sum_{j=l}^{\infty} \beta_j\varepsilon_{n+l-j}. \end{aligned} \tag{5.3}$$

Comparing (5.1) with (5.3), it can be seen that $X_n(l)$ equals X_{n+l} but with the future unknown values $\varepsilon_{n+1}, \varepsilon_{n+2}, \ldots, \varepsilon_{n+l}$ set to zero, their expected value.

Note that if $\{X_t\}$ is an MA(q) process, then $X_n(l) = 0$ for $l > q$.

In many situations an *updating forecast* is appropriate. Given data up to time n, we would predict X_{n+l} by $X_n(l)$. One time step later, we now know X_{n+1}

and would predict X_{n+l} by $X_{n+1}(l-1)$. Then it is easy to show, with $\beta_0 = 1$, that

$$X_{n+1}(l-1) = X_n(l) + \beta_{l-1}[X_{n+1} - X_n(1)].$$

5.2.2 Residuals of common time series models

We need to be able to estimate the $\{\varepsilon_t\}$; to do this, we calculate the residuals of our fitted models.

For an AR(p) process $X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t$. The residuals satisfy

$$e_t = X_t - \sum_{k=1}^{p} \widehat{\alpha}_k X_{t-k} \qquad t \geqslant p.$$

For example, for the AR(1) process $\varepsilon_t = X_t - \alpha X_{t-1}$, we have

$$\begin{aligned} e_1 &= X_1 \text{ (if we assume } X_0 = 0\text{; otherwise missing)} \\ e_2 &= X_2 - \widehat{\alpha} X_1 \\ e_3 &= X_3 - \widehat{\alpha} X_2 \end{aligned}$$

etc.

For an MA(q) process $X_t = \varepsilon_t + \sum_{k=1}^{q} \beta_k \varepsilon_{t-k}$, assume $\varepsilon_0 = \cdots = \varepsilon_{1-q} = 0$ and compute the e_t recursively.

For example, for the MA(2) process $X_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2}$, we have $\varepsilon_t = X_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2}$. Therefore, the residuals are

$$\begin{aligned} e_1 &= X_1 \\ e_2 &= X_2 - \widehat{\beta}_1 e_1 \\ e_t &= X_t - \widehat{\beta}_1 e_{t-1} - \widehat{\beta}_2 e_{t-2} \quad t = 3, 4, \ldots. \end{aligned}$$

5.2.3 Forecast error

The forecast error is $e_n(l) = X_{n+l} - X_n(l)$. With $\beta'_s = \beta_{l+s}$, it follows that

$$e_n(l) = \beta_0\varepsilon_{n+l} + \beta_1\varepsilon_{n+l-1} + \cdots + \beta_{l-1}\varepsilon_{n+1}. \tag{5.4}$$

Since $\mathbb{E}(\varepsilon_t) = 0$ for future values $t > n$, it follows that

$$\mathbb{E}[e_n(l)] = \mathbb{E}[\beta_0\varepsilon_{n+l} + \beta_1\varepsilon_{n+l-1} + \cdots + \beta_{l-1}\varepsilon_{n+1}] = 0.$$

Similarly, $\operatorname{var}[\varepsilon_t] = \sigma_\varepsilon^2$ for future values so

$$\begin{aligned}\operatorname{var}[e_n(l)] &= \beta_0^2 \operatorname{var}[\varepsilon_{n+l}] + \beta_1^2 \operatorname{var}[\varepsilon_{n+l-1}] + \cdots + \beta_{l-1}^2 \operatorname{var}[\varepsilon_{n+1}] \\ &= \sigma_\varepsilon^2 \sum_{k=0}^{l-1} \beta_k^2.\end{aligned}$$

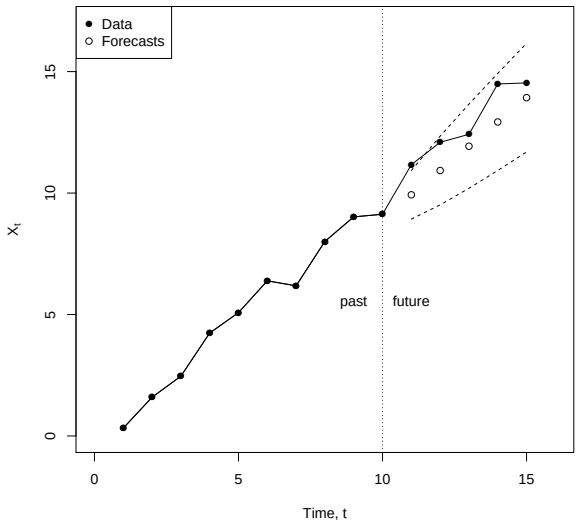
The variance increases as $l \to \infty$; the more we extrapolate into the future, the more uncertainty in our forecasts.

Equation (5.4) shows $e_n(l)$ forms an MA$(l-1)$ process. Hence the $e_n(l)$ are correlated. However notice that $\operatorname{cov}(e_n(l), \varepsilon_{n-k}) = 0$ for all $k \geqslant 0$. In particular, the forecast error is uncorrelated with the forecast itself.

It can be shown that, for $m > 0$,

$$\operatorname{cov}[e_n(l), e_n(l+m)] = \sigma_\varepsilon^2 \sum_{k=0}^{l-1} \beta_k\beta_{k+m}. \tag{5.5}$$

In practice, forecast errors tend to be to one side of actual future values.



Example 5.1 (One-step ahead forecasts)**.** *For example, consider the case* $l = 1$*. Assuming* $\beta_0 = 1$*,*

$$X_n(1) = \beta_1\varepsilon_n + \beta_2\varepsilon_{n-1} + \beta_3\varepsilon_{n-2} + \cdots$$
$$e_n(1) = X_{n+1} - X_n(1) = \varepsilon_{n+1}.$$

Since the $\{\varepsilon_t\}$ *are independent, it follows that* $e_n(1)$ *and* $X_n(1)$ *are independent.*

Example 5.2 (AR(1))**.** *Let* $X_t = \alpha X_{t-1} + \varepsilon_t$ *with* $|\alpha| < 1$ *and* $\{\varepsilon_t\}$ *being a white noise process with variance* σ_ε^2*. Hence*

$$X_t = \varepsilon_t + \alpha\varepsilon_{t-1} + \alpha^2\varepsilon_{t-2} + \cdots,$$

so $\beta_0 = 1, \beta_1 = \alpha, \beta_2 = \alpha^2$ *etc, giving*

$$\begin{aligned} X_n(1) &= \beta_1\varepsilon_n + \beta_2\varepsilon_{n-1} + \beta_3\varepsilon_{n-2} + \cdots \\ &= \alpha\varepsilon_n + \alpha^2\varepsilon_{n-1} + \alpha^3\varepsilon_{n-2} + \cdots \\ &= \alpha(\varepsilon_n + \alpha\varepsilon_{n-1} + \alpha^2\varepsilon_{n-2} + \cdots) \\ &= \alpha X_n. \end{aligned}$$

Similarly,

$$\begin{aligned} X_n(l) &= \beta_l \varepsilon_n + \beta_{l+1}\varepsilon_{n-1} + \beta_{l+2}\varepsilon_{n-2} + \cdots \\ &= \alpha^l(\varepsilon_n + \alpha\varepsilon_{n-1} + \alpha^2\varepsilon_{n-2} + \cdots) \\ &= \alpha^l X_n. \end{aligned}$$

Since $|\alpha| < 1$*, if follows that* $X_n(l) \xrightarrow[l\to\infty]{} 0$*. Therefore*

$$\begin{aligned} e_n(l) &= \varepsilon_{n+l} + \alpha\varepsilon_{n+l-1} + \cdots + \alpha^{l-2}\varepsilon_{n+2} + \alpha^{l-1}\varepsilon_{n+1} \\ \operatorname{var}(e_n(l)) &= (1 + \alpha^2 + \alpha^4 + \cdots + \alpha^{2l-2})\sigma_\varepsilon^2 = \frac{1-\alpha^{2l}}{1-\alpha^2}\sigma_\varepsilon^2 \\ &\xrightarrow[l\to\infty]{} \frac{\sigma_\varepsilon^2}{1-\alpha^2} = \gamma_0. \end{aligned}$$

Replacing α *by its estimate* $\widehat{\alpha}$ *will increase the variance of forecasts and their errors.*

Similarly

$$\operatorname{cov}(e_n(l), e_n(l+m)) = \sigma_\varepsilon^2 \sum_{j=0}^{l-1} \beta_j\beta_{j+m} = \sigma_\varepsilon^2 \sum_{j=0}^{l-1} \alpha^{2j+m} = \frac{(1-\alpha^{2l})\alpha^m}{(1-\alpha^2)}\sigma_\varepsilon^2$$

giving

$$\operatorname{corr}(e_n(l), e_n(l+m)) = \alpha^m \sqrt{\frac{1-\alpha^{2l}}{1-\alpha^{2l+2m}}}.$$

5.2.4 Forecast as a conditional mean

We can also think of the forecast $X_n(l)$ as the mean of X_{n+l} given the known values $X_1, X_2, \ldots, X_n$. Given $\{\varepsilon_t : t \leqslant n\}$ we can in principle determine $\{X_t : t \leqslant n\}$ and *vice versa* so

$$\mathbb{E}[\varepsilon_t | X_1, \ldots, X_n] = \begin{cases} 0 & \text{if } t > n, \\ \varepsilon_t & \text{if } t \leqslant n, \end{cases} \tag{5.6}$$

as future values are unknown and past values are "known".

Now recall

$$X_{n+l} = \beta_0\varepsilon_{n+l} + \beta_1\varepsilon_{n+l-1} + \beta_2\varepsilon_{n+l-2} + \cdots + \beta_l\varepsilon_n + \beta_{l+1}\varepsilon_{n-1} + \beta_{l+2}\varepsilon_{n-2} + \cdots ,$$

so

$$\begin{aligned}\mathbb{E}[X_{n+l}|X_1,\ldots,X_n] &= \beta_0\mathbb{E}[\varepsilon_{n+l}|X_1,\ldots,X_n] + \beta_1\mathbb{E}[\varepsilon_{n+l-1}|X_1,\ldots,X_n] + \cdots \\ &\quad +\beta_l\mathbb{E}[\varepsilon_n|X_1,\ldots,X_n] + \cdots \\ &= \beta_l\varepsilon_n + \beta_{l+1}\varepsilon_{n-1} + \beta_{l+2}\varepsilon_{n-2} + \cdots = X_n(l). \qquad (5.7)\end{aligned}$$

Also

$$\begin{aligned}\mathrm{var}[X_{n+l}|X_1,X_2,\ldots,X_n] &= \mathbb{E}[(X_{n+l} - \mathbb{E}[X_{n+l}|X_1,X_2,\ldots,X_n])^2|X_1,X_2,\ldots,X_n] \\ &= \mathbb{E}[(X_{n+l} - X_n(l))^2|X_1,X_2,\ldots,X_n] \equiv \mathrm{var}[e_n(l)].\end{aligned}$$

Assuming the $\{\varepsilon_t\}$ are normally distributed,

$$X_{n+l}|X_1,X_2,\ldots,X_n \sim \mathbb{N}(X_n(l),(\beta_0^2+\beta_1^2+\cdots+\beta_{l-1}^2)\sigma_\varepsilon^2),$$

so a 95% prediction interval for the forecast is given by

$$X_{n+l}|X_1,X_2,\ldots,X_n \in X_n(l) \pm 1.96\sqrt{(\beta_0^2+\beta_1^2+\cdots+\beta_{l-1}^2)\sigma_\varepsilon^2}.$$

The parameters of the time series model have to be estimated so that in practice the true 95% interval is slightly wider than using the normality assumption.

Consider equation (5.6) and notice also that $\mathbb{E}[X_t|X_1,\ldots,X_n]$ equals the forecast if $t > n$ and is known if $t \leqslant n$, so that

$$\mathbb{E}[X_t|X_1,\ldots,X_n] = \begin{cases} X_n(t-n) & \text{if } t > n, \\ X_t & \text{if } t \leqslant n. \end{cases}$$

5.3 Model based forecasting

An alternative approach to forecasting is to consider the model for time t after the last observation.

For example, imagine we have data $X_1, \ldots, X_n$ for times $t = 1, \ldots, n$. Assume we have fitted an AR(2) model assuming normally distributed innovations and obtained $\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\sigma}^2_\varepsilon$.

For time $n+1$, the model is

$$X_{n+1} = \alpha_1 X_n + \alpha_2 X_{n-1} + \varepsilon_{n+1},$$

where

X_n, X_{n-1} — known
α_1, α_2 — estimated
ε_{n+1} — value unknown, but variance estimated.

Since we don't have information about ε_{n+1}, we cannot estimate the exact value of X_{n+1}. However, given X_n, X_{n-1}, the value X_{n+1} is normally distributed with mean

$$\begin{aligned}
\mathbb{E}(X_{n+1}|X_n, X_{n-1}) &= \mathbb{E}(\alpha_1 X_n + \alpha_2 X_{n-1} + \varepsilon_{n+1}|X_n, X_{n-1}) \\
&= \alpha_1 X_n + \alpha_2 X_{n-1} + \mathbb{E}(\varepsilon_{n+1}|X_n, X_{n-1}) \\
&\approx \widehat{\alpha}_1 X_n + \widehat{\alpha}_2 X_{n-1} + 0
\end{aligned}$$

and variance

$$\begin{aligned}
\operatorname{var}(X_{n+1}|X_n, X_{n-1}) &= \operatorname{var}(\underbrace{\alpha_1 X_n + \alpha_2 X_{n-1}}_{\text{const}} + \varepsilon_{n+1}|X_n, X_{n-1}) \\
&= \sigma^2_\varepsilon \\
&\approx \widehat{\sigma}^2_\varepsilon.
\end{aligned}$$

Therefore, an approximate 95% confidence interval for X_{n+1} is

$$\widehat{\alpha}_1 X_n + \widehat{\alpha}_2 X_{n-1} \pm 1.96 \widehat{\sigma}_\varepsilon.$$

Iterating this procedure gives forecasts for longer times. For example, consider the AR(1) process with parameter estimates $\widehat{\alpha}, \widehat{\sigma}_\varepsilon$. The estimates have means

$$\begin{aligned}
\mathbb{E}(X_{n+1}|X_n) &= \mathbb{E}(\alpha X_n + \varepsilon_{n+1}|X_n) = \alpha X_n \\
\mathbb{E}(X_{n+2}|X_n) &= \mathbb{E}(\alpha X_{n+1} + \varepsilon_{n+2}|X_n) \\
&= \alpha \mathbb{E}(X_{n+1}|X_n) + \mathbb{E}(\varepsilon_{n+2}|X_n) \\
&= \alpha^2 X_n \\
&\vdots \\
\mathbb{E}(X_{n+l}|X_n) &= \alpha^l X_n \xrightarrow[l\to\infty]{} 0
\end{aligned}$$

and variance

$$\begin{aligned}
\operatorname{var}(X_{n+1}|X_n) &= \operatorname{var}(\alpha X_n + \varepsilon_{n+1}|X_n) = \sigma_\varepsilon^2 \\
\operatorname{var}(X_{n+2}|X_n) &= \operatorname{var}(\alpha X_{n+1} + \varepsilon_{n+2}|X_n) \\
&= \alpha^2 \operatorname{var}(X_{n+1}|X_n) + \operatorname{var}(\varepsilon_{n+2}|X_n) \\
&= \alpha^2 \sigma_\varepsilon^2 + \sigma_\varepsilon^2 \\
&\vdots \\
\operatorname{var}(X_{n+l}|X_n) &= (1 + \alpha^2 + \cdots + \alpha^{2l})\sigma_\varepsilon^2 \\
&\xrightarrow[l\to\infty]{} \frac{1}{1-\alpha^2}\sigma_\varepsilon^2.
\end{aligned}$$

For long time forecasts, the mean and variance coincide with the values for the stationary process and the data do not provide any extra information.

Of course this looks rather familiar! For normally distributed errors, the model-based and minimum mean square error approaches will yield the same forecasts.

Part II

MATH5802M
Time Series and Spectral Analysis

Chapter 6

Fitting sine waves

6.1 Frequency

In *spectral analysis*, a time series is analysed by decomposing it into sine waves of different amplitude and frequency.

low frequency $\approx$ slowly changing trends

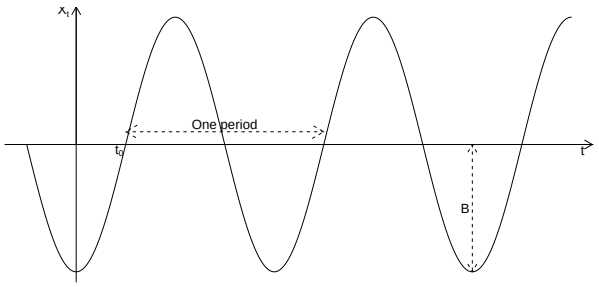
high frequency $\approx$ rapid changes (noise).

If the data have a periodic component of known frequency, we can fit the following model:

$$X_t = B\sin(2\pi f(t - t_0)) + \mu + \varepsilon_t,$$

where

μ — baseline or mean;
f = *frequency* (in cycles per unit time);
t_0 = start time of one period;
B = *amplitude*



Example 6.1. *If t is given in months since the start of measurements, then a period length of one year implies $f = 1/12$.*

Note that we can use the trigonometric identities

$$\sin(x \pm y) = \sin(x)\cos(y) \pm \cos(x)\sin(y)$$
$$\cos(x \pm y) = \cos(x)\cos(y) \mp \sin(x)\sin(y)$$

to get

$$\begin{aligned} X_t &= B\sin(2\pi ft - 2\pi ft_0)) + \mu + \varepsilon_t \\ &= \underbrace{B\cos(2\pi ft_0)}_{\tilde{B}}\sin(2\pi ft)\underbrace{-B\sin(2\pi ft_0)}_{\tilde{A}}\cos(2\pi ft) + \mu + \varepsilon_t \\ &= \tilde{A}\cos(2\pi ft) + \tilde{B}\sin(2\pi ft) + \mu + \varepsilon_t. \end{aligned}$$

Since any periodic component with frequency f can be written in this form, we can use the model

$$X_t = A\cos(2\pi ft) + B\sin(2\pi ft) + \mu + \varepsilon_t$$

for the sum of arbitrary many shifted sines/cosines with frequency f. Therefore we need just *two* parameters (A, B) per frequency.

6.2 Aliasing

We have seen that only two parameters per frequency are required; frequency f is described by

$$A\cos(2\pi ft) + B\sin(2\pi ft). \tag{6.1}$$

A model with more parameters for frequency f cannot be distinguished from (6.1) by using observations.

Since

$$A\cos(2\pi(-f)t) + B\sin(2\pi(-f)t) = A\cos(2\pi ft) - B\sin(2\pi ft),$$

the frequencies f and $-f$ are indistinguishable; they are *aliases* of each other. Hence we can assume $f \geqslant 0$.

Assume that observations are given at times $0, \Delta, 2\Delta, 3\Delta, \ldots$, i.e. we have sampling frequency $1/\Delta$. A component with frequency f will be observed as

$$X_k = A\cos(2\pi fk\Delta) + B\sin(2\pi fk\Delta).$$

The bigger f, the faster X_k oscillates, until $f = \frac{1}{2\Delta}$. Here, we have

$$\begin{aligned} X_k &= A\cos\left(2\pi\frac{1}{2\Delta}k\Delta\right) + B\sin\left(2\pi\frac{1}{2\Delta}k\Delta\right) \\ &= A\cos(k\pi) + \underbrace{B\sin(k\pi)}_{=0} \\ &= A(-1)^k. \end{aligned}$$

This is the fastest possible oscillation.

What happens for $f > \frac{1}{2\Delta}$? Let $f = \frac{1}{2\Delta} + \varepsilon$. Then

$$\begin{aligned}
&A\cos(2\pi f k\Delta) + B\sin(2\pi f k\Delta) \\
&= A\cos(\pi k + 2\pi\varepsilon k\Delta) + B\sin(\pi k + 2\pi\varepsilon k\Delta) \\
&= A\cos(-\pi k + 2\pi\varepsilon k\Delta) + B\sin(-\pi k + 2\pi\varepsilon k\Delta) \\
&= A\cos\left(-2\pi\left(\frac{1}{2\Delta} - \varepsilon\right)k\Delta\right) + B\sin\left(-2\pi\left(\frac{1}{2\Delta} - \varepsilon\right)k\Delta\right) \\
&= A\cos\left(2\pi\left(\frac{1}{2\Delta} - \varepsilon\right)k\Delta\right) + (-B)\sin\left(2\pi\left(\frac{1}{2\Delta} - \varepsilon\right)k\Delta\right).
\end{aligned}$$

Therefore, $\frac{1}{2\Delta} + \varepsilon$ and $\frac{1}{2\Delta} - \varepsilon$ are aliases, so we can assume $0 \leqslant f \leqslant \frac{1}{2\Delta}$.

Definition 6.1 (Nyquist frequency)**.** *The frequency $f = \frac{1}{2\Delta}$ is called the* Nyquist frequency *for the sampling frequency $1/\Delta$.*

Chapter 7

Computing the periodogram

We shall use a tool called the *discrete Fourier transform* (DFT) to draw *periodograms*. These graphs will help us explore which frequencies contribute to our time series.

The DFT is a method for decomposing a time series into periodic components. Assume we have observations X_t made at times $t = 0, 1, 2, \ldots, n-1$.

Definition 7.1 (Fourier frequencies)**.** *The frequencies* $f_j = j/n$ *are called* Fourier frequencies*.*

7.1 The DFT

Definition 7.2 (DFT)**.** *The DFT of* $X_0, \ldots, X_{n-1}$ *is given by*

$$\begin{aligned} \widehat{X}_j &= \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t e^{-2\pi i f_j t} \qquad \text{for } j = 0, \ldots, n-1 \\ &= \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t \left[\cos(2\pi f_j t) - i \sin(2\pi f_j t)\right]. \end{aligned}$$

Lemma 7.1.

$$\sum_{t=0}^{n-1} e^{2\pi i f_j t} e^{-2\pi i f_k t} = \begin{cases} n & \text{if } j = k \pmod n \\ 0 & \text{else.} \end{cases}$$

Proof. If $j = k$,

$$\sum_{t=0}^{n-1} \underbrace{e^{2\pi i f_j t} e^{-2\pi i f_k t}}_{=e^0=1} = \sum_{t=0}^{n-1} 1 = n.$$

If $j \neq k$, we use $\sum_{r=0}^{n-1} a^r = (1-a^n)/(1-a)$ to show

$$\begin{aligned} \sum_{t=0}^{n-1} e^{2\pi i f_j t} e^{-2\pi i f_k t} &= \sum_{t=0}^{n-1} e^{2\pi i \left(\frac{j}{n} - \frac{k}{n}\right) t} \\ &= \sum_{t=0}^{n-1} \left(e^{2\pi i \left(\frac{j-k}{n}\right)}\right)^t \\ &= \frac{1 - \left(e^{2\pi i \left(\frac{j-k}{n}\right)}\right)^n}{1 - e^{2\pi i \left(\frac{j-k}{n}\right)}} \\ &= \frac{1-1}{1 - e^{2\pi i \left(\frac{j-k}{n}\right)}} \\ &= 0. \end{aligned}$$

□

Definition 7.3 (Inverse DFT)**.** *Using lemma 7.1, we can recover* $\{X_t\}$ *from* $\{\widehat{X}_j\}$*:*

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} \widehat{X}_j e^{2\pi i f_j t} &= \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} \left(\frac{1}{\sqrt{n}} \sum_{s=0}^{n-1} X_s e^{-2\pi i f_j s} \right) e^{2\pi i f_j t} \\ &= \sum_{s=0}^{n-1} X_s \underbrace{\frac{1}{n} \sum_{j=0}^{n-1} e^{2\pi i \frac{t}{n} j} e^{-2\pi i \frac{s}{n} j}}_{= \begin{cases} 1 & \text{if } s=t \\ 0 & \text{else} \end{cases}} \\ &= X_t \qquad \text{for } t = 0, \ldots, n-1. \end{aligned}$$

This is called the inverse DFT*.*

Example 7.1. *Let $n = 4$, $X = 0, 1, 2, 3$. Then*

$$\widehat{X}_0 = \frac{1}{2}\sum_{t=0}^{3} X_t e^{-2\pi i 0t/4} = \frac{1}{2}\sum_{t=0}^{3} t = 3,$$

$$\begin{aligned}
\widehat{X}_1 &= \frac{1}{2}\sum_{t=0}^{4} te^{-2\pi it/4} \\
2\widehat{X}_1 &= 0 + e^{-2\pi i/4} + 2e^{-2\pi i/2} + 3e^{-2\pi i 3/4} \\
&= e^{-\pi i/2} + 2e^{-\pi i} + 3e^{-3\pi i/2} \\
&= (\cos(-\pi/2) + i\sin(-\pi/2)) + 2(\cos(-\pi) + i\sin(-\pi)) \\
&\qquad + 3(\cos(-3\pi/2) + i\sin(-3\pi/2)) \\
&= (0 - i) + 2(-1 + 0i) + 3(0 + i) \\
\widehat{X}_1 &= -1 + i,
\end{aligned}$$

$$\widehat{X}_2 = \cdots = -1 \qquad\qquad \widehat{X}_3 = \cdots = -1 - i.$$

So the DFT of $\{0, 1, 2, 3\}$ is $\{3, -1 + i, -1, -1 - i\}$. For the inverse,

$$\begin{aligned}
X_0 &= \frac{1}{2}\sum_{j=0}^{3} \widehat{X}_j e^{2\pi i f_j 0} = \frac{1}{2}\sum_{j=0}^{4} \widehat{X}_j \\
&= \frac{1}{2}(3 \quad -1+i \quad -1 \quad -1-i) = 0,
\end{aligned}$$

as it should.

Remark Some people define the DFT as

$$\tilde{X}_j = \sum_{t=0}^{n-1} X_t e^{-2\pi i f_j t}.$$

In this case,

$$X_t = \frac{1}{n}\sum_{j=0}^{n-1} \tilde{X}_j e^{2\pi i f_j t}.$$

Basic properties

1. The DFT is linear, i.e.

$$Y_t = cX_t \Rightarrow \widehat{Y}_j = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} cX_t e^{-2\pi i f_j t} = c\widehat{X}_j$$
$$Z_t = X_t + Y_t \Rightarrow \cdots \Rightarrow \widehat{Z}_j = \widehat{X}_j + \widehat{Y}_j.$$

2. We have "conservation of energy". Recall that if $X = a + ib$ then $\overline{X} = a - ib$ and so

$$|\widehat{X}_j|^2 = \widehat{X}_j \cdot \overline{\widehat{X}_j} = \frac{1}{\sqrt{n}} \sum_{s=0}^{n-1} X_s e^{-2\pi i f_j s} \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} \overline{X}_t e^{2\pi i f_j t}.$$

   Therefore,

$$\sum_{j=0}^{n-1} |\widehat{X}_j|^2 = \frac{1}{n} \sum_{s,t=0}^{n-1} X_s \overline{X}_t \underbrace{\sum_{j=0}^{n-1} e^{-2\pi i \frac{js}{n}} e^{2\pi i \frac{jt}{n}}}_{= \begin{cases} n & \text{if } s = t \\ 0 & \text{else} \end{cases}}$$
$$= \frac{1}{n} \sum_{t=0}^{n-1} X_t \overline{X}_t \cdot n$$
$$= \sum_{t=0}^{n-1} |X_t|^2,$$

   and hence $\|\widehat{X}\| = \|X\|$.

3. Time shifts. Let $Y = (X_{n-1}, X_0, X_1, \ldots, X_{n-2})$, i.e. $Y_t = X_{t-1 \bmod n}$. Then

$$\widehat{Y}_j = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} Y_t e^{-2\pi i f_j t}$$
$$= \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_{t-1 \bmod n} e^{-2\pi i f_j t}$$
$$= \frac{1}{\sqrt{n}} \sum_{s=0}^{n-1} X_s e^{-2\pi i f_j (s+1)}$$
$$= e^{-2\pi i f_j} \widehat{X}_j$$
$$\Rightarrow |\widehat{Y}_j|^2 = |\widehat{X}_j|^2;$$

only the "phase" is changed, the modulus is the same. Therefore, it doesn't matter what point in the cycle our data start from.

If instead of the cyclic shift Y we consider $Z = (X_{-1}, X_0, X_1, \ldots, X_{n-2})$ for some $X_{-1} \in \mathbb{R}$, we get

$$\|\widehat{Y} - \widehat{Z}\| = \|\widehat{Y - Z}\| = \|Y - Z\| = \sqrt{(X_{n-1} - X_{-1})^2} = |X_{n-1} - X_{-1}|.$$

For stationary time series this expression will stay bounded, while $\|\widehat{Y}\|, \|\widehat{Z}\| \to \infty$ as $n \to \infty$ (since the number of terms in $\|\widehat{Y}\|^2 = \sum_{j=0}^{n-1} \widehat{Y}_j^2$ grows). Therefore the relative difference between $\widehat{Y}$ and $\widehat{Z}$ goes to zero as $n \to \infty$.

Result For (long) stationary time series, we can use cyclic shifts instead of normal ones.

7.2 The spectral density and periodogram

Definition 7.4 (Spectral density and periodogram)**.** *The function $I : [0, 1] \to [0, \infty)$ with $I(f_j) = |\widehat{X}_j|^2$ is called the* spectral density *of X. A plot of I as a function of f is called a* periodogram*.*

Example 7.2. *Let $X = \{0, 1, 2, 3\}$ as above. Then we have $\widehat{X} = \{3, -1 + i, -1, -1 - i\}$, so*

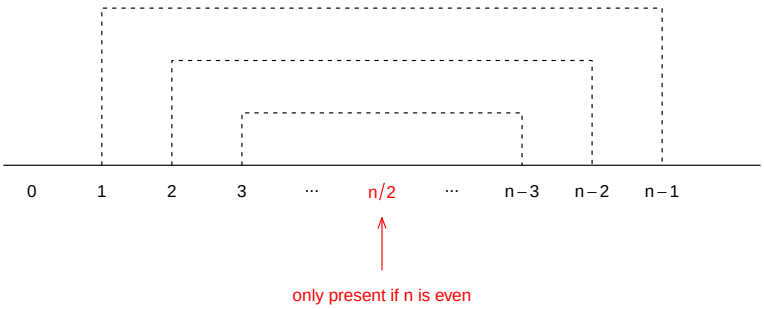
$$\begin{aligned} I(0) &= |\widehat{X}_0|^2 = 3^2 = 9, \\ I(1/4) &= |\widehat{X}_1|^2 = (-1 + i)(-1 - i) = 2, \\ I(2/4) &= |\widehat{X}_2|^2 = (-1)^2 = 1, \\ I(3/4) &= |\widehat{X}_3|^2 = 2. \end{aligned}$$

The spectral density $I(f_j)$ is a measurement of how strongly the frequency f_j

is represented in the data. For real data $\{X_t\}$, we have

$$\begin{aligned}
\widehat{X}_{n-j} &= \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t e^{-2\pi i \frac{n-j}{n} t} \\
&= \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} \underbrace{X_t}_{=\overline{X}_t} \underbrace{e^{-2\pi i t}}_{=1} \underbrace{e^{2\pi i \frac{j}{n} t}}_{=\overline{e^{-2\pi i f_j t}}} \\
&= \overline{\frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t e^{-2\pi i f_j t}} \\
&= \overline{\widehat{X}_j},
\end{aligned}$$

so only half the frequencies are needed. (This is expected, since f_j for $j > n/2$ is bigger than the Nyquist frequency $\frac{1}{2\Delta} = \frac{1}{2}$.)



Note that, defining $\tilde{X}$ to be the mean of $\{X_t\}$,

$$\widehat{X}_0 = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t \underbrace{e^{2\pi i f_0 t}}_{=1} = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t = \sqrt{n}\tilde{X},$$

so $\widehat{X}_0 \in \mathbb{R}$. Similarly, if n is even then $\widehat{X}_{n/2} = \overline{\widehat{X}}_{n-n/2} = \overline{\widehat{X}_{n/2}}$ and so $\widehat{X}_{n/2} \in \mathbb{R}$.

Let $\widehat{X}_j = a_j + ib_j \Rightarrow \widehat{X}_{n-j} = a_j - ib_j$. Now, since $a_0 = \widehat{X}_0$ and $a_{n/2} = \widehat{X}_{n/2}$,

we have

$$\begin{aligned} X_t &= \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} \widehat{X}_j e^{2\pi i f_j t} \\ &= \frac{1}{\sqrt{n}} a_0 + \frac{1}{\sqrt{n}} \sum_{1 \leqslant j < \frac{n}{2}} \left(\widehat{X}_j e^{2\pi i f_j t} + \widehat{X}_{n-j} e^{2\pi i f_{n-j} t} \right) \\ &\qquad + \frac{1}{\sqrt{n}} a_{n/2} e^{2\pi i f_{n/2} t} \qquad \text{only present if } n \text{ is even.} \end{aligned}$$

Note that

$$e^{2\pi i f_{n/2} t} = e^{2\pi i \frac{n/2}{n} t} = e^{\pi i t} = \cos(\pi t) + i \sin(\pi t) = (-1)^t.$$

Therefore

$$\begin{aligned} \widehat{X}_j e^{2\pi i f_j t} + \widehat{X}_{n-j} e^{2\pi i f_{n-j} t} &= (a_j + i b_j)[\cos(2\pi f_j t) + i \sin(2\pi f_j t)] \\ &\quad + (a_j - i b_j)[\cos(-2\pi f_j t) + i \sin(-2\pi f_j t)] \\ &= 2a_j \cos(2\pi f_j t) - 2b_j \sin(2\pi f_j t). \end{aligned}$$

(Exercise: What is the amplitude of $A \cos(2\pi f t) + B \sin(2\pi f t)$?) Therefore

$$X_t = \sum_{0 \leqslant j \leqslant \frac{n}{2}} [A_j \cos(2\pi f_j t) + B_j \sin(2\pi f_j t)],$$

where

$$\begin{aligned} A_0 &= \frac{a_0}{\sqrt{n}} & B_0 &= 0 & \\ A_j &= \frac{2a_j}{\sqrt{n}} & B_j &= -\frac{2b_j}{\sqrt{n}} & \text{for } 1 \leqslant j \leqslant \frac{n}{2} \\ A_{\frac{n}{2}} &= \frac{a_{n/2}}{\sqrt{n}} & B_{\frac{n}{2}} &= 0 & \text{if } n \text{ is even .} \end{aligned}$$

Chapter 8

DFT for time series

8.1 Spectral density and autocovariance

Let $X_0, X_1, \ldots, X_{n-1}$ be a time series. Then

1. $$\widehat{X}_0 = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t e^{-2\pi i 0 t} = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t = \sqrt{n}\tilde{X},$$

   where $\tilde{X}$ is the sample mean.

2. If $\tilde{X} = 0$, we have

   $$\mathrm{var}(X) = \frac{1}{n} \sum_{t=0}^{n-1} |X_t|^2 = \frac{1}{n} \sum_{j=0}^{n-1} |\widehat{X}_j|^2.$$

   Thus

   $$\mathrm{var}(X) = \frac{1}{n} \sum_{j=0}^{n-1} I(f_j) \approx \int_0^1 I(f) df. \tag{8.1}$$

   Similarly

   $$\frac{1}{n} \sum_{f_j \in [a,b]} I(f_j) \approx \int_a^b I(f_j) df$$

   quantifies the variance contributed by the frequency range $[a, b]$ (hence the term “spectral density”).

3. The spectral density is closely linked to the sample autocovariance since

$$
\begin{aligned}
I(f_j) &= \widehat{X}_j \cdot \overline{\widehat{X}}_j \\
&= \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t e^{-2\pi i f_j t} \; \frac{1}{\sqrt{n}} \sum_{s=0}^{n-1} X_s e^{+2\pi i f_j s} \\
&= \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} (X_t - \tilde{X}) e^{-2\pi i f_j t} \; \frac{1}{\sqrt{n}} \sum_{s=0}^{n-1} (X_s - \tilde{X}) e^{+2\pi i f_j s} \\
&\qquad \left[\text{since } \sum_{t=0}^{n-1} \tilde{X} e^{\pm 2\pi i f_j t} = \tilde{X} \frac{1 - e^{\pm 2\pi i f_j n}}{1 - e^{-2\pi i f_j}} = 0 \right] \\
&= \frac{1}{n} \sum_{s,t} (X_s - \tilde{X})(X_t - \tilde{X}) e^{-2\pi i f_j (t-s)} \\
&\overset{k=t-s}{=} \sum_k \left(\underbrace{\frac{1}{n} \sum_s (X_s - \tilde{X})(X_{s+k} - \tilde{X})}_{=g_k} \right) e^{-2\pi i f_j k} \\
&\qquad \text{(assuming } X \text{ to be periodic, so take subscripts modulo } n) \\
&= \sqrt{n}\, \widehat{g}_j,
\end{aligned}
$$

where g_k is the lag-k sample autocovariance of X.

Furthermore, since we assume X to be periodic, we have

$$
\begin{aligned}
g_{n-k} &= \frac{1}{n} \sum_s (X_s - \tilde{X})(X_{s+n-k} - \tilde{X}) \\
&\overset{t=s-k}{=} \frac{1}{n} \sum_t (X_{t+k} - \tilde{X})(X_t - \tilde{X}) \\
&= g_k
\end{aligned}
$$

and

$$
e^{-2\pi i f_j (n-k)} = e^{2\pi i f_j k}.
$$

Therefore,

$$
\begin{aligned}
I(f_j) &= g_0 + 2 \sum_{1 \leqslant k < \frac{n}{2}} g_k \frac{1}{2} \left(e^{-2\pi i f_j k} + e^{2\pi i f_j k} \right) \\
&\qquad + g_{\frac{n}{2}} (-1)^j \qquad \text{if } n \text{ is even} \\
&= g_0 + 2 \sum_{1 \leqslant k < \frac{n}{2}} g_k \cos(2\pi f_j k) + g_{\frac{n}{2}} (-1)^j.
\end{aligned}
$$

The spectral density I is an estimate for the *power spectrum*

$$\mathcal{I}(f) = \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k \cos(2\pi f k).$$

8.2 Spectral density of white noise

Let $X_t \sim N(0, \sigma^2)$, i.i.d. Then

$$\begin{aligned}\widehat{X}_j &= \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t e^{-2\pi i f_j t} \\ &= \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t \cos(2\pi f_j t) - i\frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t \sin(2\pi f_j t),\end{aligned}$$

so $\widehat{X}_j$ is (complex) normally distributed with $\mathbb{E}(\operatorname{Re} \widehat{X}_j) = \mathbb{E}(\operatorname{Im} \widehat{X}_j) = 0$.

For the variances we get

$$\begin{aligned}\operatorname{var}(\operatorname{Re} \widehat{X}_j) &= \frac{1}{n} \sum_{t=0}^{n-1} \cos(2\pi f_j t)^2 \sigma^2 \\ &= \begin{cases} \sigma^2 & \text{if } j \in \{0, \frac{n}{2}\} \\ \frac{\sigma^2}{2} & \text{if } j \notin \{0, \frac{n}{2}\}, \end{cases}\end{aligned}$$

$$\begin{aligned}\operatorname{var}(\operatorname{Im} \widehat{X}_j) &= \frac{1}{n} \sum_{t=0}^{n-1} \sin(2\pi f_t t)^2 \sigma^2 \\ &= \begin{cases} 0 & \text{if } j \in \{0, \frac{n}{2}\} \\ \frac{\sigma^2}{2} & \text{if } j \notin \{0, \frac{n}{2}\}, \end{cases}\end{aligned}$$

$$\begin{aligned}\operatorname{cov}(\operatorname{Re}\widehat{X}_j, \operatorname{Im}\widehat{X}_j) &= -\frac{1}{n}\sum_{s,t=0}^{n-1}\cos(2\pi f_j s)\sin(2\pi f_j t)\operatorname{cov}(X_s, X_t)\\ &= -\frac{\sigma^2}{n}\sum_{t=0}^{n-1}\cos(2\pi f_j t)\sin(2\pi f_j t)\\ &= 0.\end{aligned}$$

Also, $\widehat{X}_j, \widehat{X}_k$ are uncorrelated for $j \neq k$. In summary,

j	$\operatorname{Re}\widehat{X}_j$	$\operatorname{Im}\widehat{X}_j$	$\mathbb{E}\{I(f_j)\}$	$I(f_j)$	
0	$N(0,\sigma^2)$	0	σ^2	$\sigma^2\chi^2_1$	
1	$N(0,\frac{\sigma^2}{2})$	$N(0,\frac{\sigma^2}{2})$	$\frac{\sigma^2}{2}+\frac{\sigma^2}{2}=\sigma^2$	$\frac{\sigma^2}{2}\chi^2_2$	
⋮					
$\frac{n}{2}$	$N(0,\sigma^2)$	0	σ^2	$\sigma^2\chi^2_1$	only for even n
⋮					

All of these are independent. For $j > n/2, \widehat{X}_j = \overline{\widehat{X}}_{n-j}$.

On average, all frequencies contribute the same (hence the name “white noise”).

8.3 Spectral density of AR(p)

Let X be given by

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t,$$

where $\varepsilon_t \sim N(0,\sigma^2)$ i.i.d. What is the DFT of X?

We have seen (§7.1)

$$Y_t = X_{t-1} \quad \Rightarrow \quad \widehat{Y}_j = e^{-2\pi i f_j}\widehat{X}_j.$$

Similarly

$$\begin{aligned} Y_t = X_{t-2} \qquad &\Rightarrow \qquad \widehat{Y}_j = \underbrace{e^{-2\pi i f_j} e^{-2\pi i f_j}}_{e^{-2\pi i f_j \cdot 2}} \widehat{X}_j \\ &\vdots \\ Y_t = X_{t-p} \qquad &\Rightarrow \qquad \widehat{Y}_j = e^{-2\pi i f_j \cdot p} \widehat{X}_j \end{aligned}$$

We also saw that the DFT is linear; $\widehat{(aX + bY)}_j = a\widehat{X}_j + b\widehat{Y}_j$, so

$$\begin{aligned} &\widehat{X}_j = \alpha_1 e^{-2\pi i f_j} \widehat{X}_j + \alpha_2 e^{-2\pi i f_j \cdot 2} \widehat{X}_j + \cdots + \alpha_p e^{-2\pi i f_j \cdot p} \widehat{X}_j + \widehat{\varepsilon}_j \\ \Rightarrow &\left(1 - \sum_{k=1}^{p} \alpha_k e^{-2\pi i f_j \cdot k}\right) \widehat{X}_j = \widehat{\varepsilon}_j \\ \Rightarrow &\widehat{X}_j = \frac{\widehat{\varepsilon}_j}{1 - \sum_{k=1}^{p} \alpha_k e^{-2\pi i f_j \cdot k}}. \end{aligned}$$

The spectral density of X is given by

$$I(f_j) = |\widehat{X}_j|^2 = \frac{|\widehat{\varepsilon}_j|^2}{\left|1 - \sum_{k=1}^{p} \alpha_k e^{-2\pi i f_j \cdot k}\right|^2}.$$

We have seen $\mathbb{E}(|\widehat{\varepsilon}|^2) = \sigma^2 \ \ \forall j$, so

$$\mathbb{E}[I(f_j)] = \frac{\sigma^2}{\left|1 - \sum_{k=1}^{p} \alpha_k e^{-2\pi i f_j \cdot k}\right|^2}.$$

For example, if X_t is AR(1) we have (writing $\alpha = \alpha_1$)

$$\begin{aligned} \left|1 - \alpha e^{-2\pi i f_j}\right|^2 &= \left| \underbrace{1 - \alpha \cos(-2\pi f_j)}_{\text{Re}} - i \underbrace{\alpha \sin(-2\pi f_j)}_{\text{Im}} \right|^2 \\ &= [1 - \alpha \cos(2\pi f_j)]^2 + \alpha^2 \sin(2\pi f_j)^2 \\ &= 1 - 2\alpha \cos(2\pi f_j) + \underbrace{\alpha^2 \cos(2\pi f_j)^2 + \alpha^2 \sin(2\pi f_j)^2}_{\alpha^2} \\ &= 1 - 2\alpha \cos(2\pi f_j) + \alpha^2. \end{aligned}$$

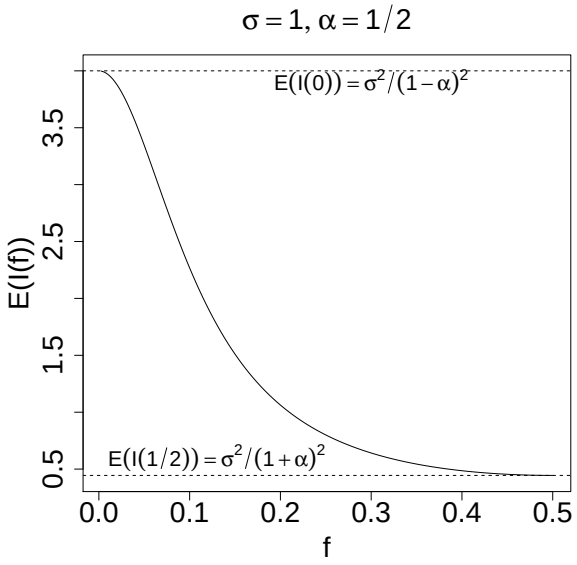


Figure 8.1: Expected spectral density function for AR(1).

Therefore

$$I(F_j) = \frac{|\widehat{\varepsilon}_j|^2}{1 - 2\alpha\cos(2\pi f_j) + \alpha^2}$$
$$\mathbb{E}[I(f_j)] = \frac{\sigma^2}{1 - 2\alpha\cos(2\pi f_j) + \alpha^2}.$$

For example, figure 8.1 shows the expected spectral density function of an AR(1) process with $\sigma = 1, \alpha = 1/2$.

Chapter 9

Complications

9.1 Smoothing the periodogram

9.1.1 Inconsistency

Recall that if $\{\varepsilon_t\}$ is a white noise process then $\mathrm{Re}(\widehat{\varepsilon}_j) \sim N(0, \sigma_\varepsilon^2/2)$ and $\mathrm{Im}(\widehat{\varepsilon}_j) \sim N(0, \sigma_\varepsilon^2/2)$, with these components being independent. Therefore

$$I(f_j) = |\widehat{\varepsilon}_j|^2 = \mathrm{Re}(\widehat{\varepsilon}_j) + \mathrm{Im}(\widehat{\varepsilon}_j)^2$$

is the sum of the squares of two independent normal r.v.s and hence

$$\frac{2}{\sigma_\varepsilon^2} I(f_j) \sim \chi_2^2 \qquad\qquad 1 \leqslant j < n/2.$$

Since the variance of a χ_ν^2 r.v. is 2ν, we have $\mathrm{var}[I(f_j)] = \sigma_\varepsilon^4$, which is constant as $n \to \infty$. Similarly, if $\{X_t\}$ is an AR(1) process, then $\mathrm{var}[I_X(f_j)] = \kappa\sigma_\varepsilon^4$ for some constant κ.

In general, $\mathrm{var}[I(f)]$ is constant as n increases; $I(f)$ is an *inconsistent* estimate of the power spectrum $\mathcal{I}(f)$. As n increases, we still attempt to estimate $I(f)$ at n frequencies. Therefore, although we have more data, the amount of data per frequency remains constant.

We would prefer an estimate which gets more reliable as we gather more data. There are a number of ways of overcoming this problem. We shall consider truncating the autocovariance function and smoothing the raw periodogram.

9.1.2 Truncating the autocovariance function

We have seen that

$$I(f_j) = g_0 + 2\sum_{1\leqslant k<n/2} g_k \cos(2\pi f_j k) + g_{n/2}(-1)^j, \qquad n \text{ even},$$

where g_k is the lag k sample autocovariance. If we do not assume $\{X_t\}$ to be periodic, then the number of pairs (X_t, X_{t+k}) decreases as k increases, so g_k is less reliable for large lags. Hence we might use

$$\mathcal{I}_T(f_j) = \lambda_0 g_0 + 2\sum_{k=1}^{m} \lambda_k g_k \cos(2\pi f_j k), \tag{9.1}$$

with $m < n/2$, for some suitable weights $\{\lambda_k\}$ called the *lag window*. Two possible choices are

$$\text{Tukey window} \quad \lambda_k = \frac{1}{2}\left[1 + \cos\left(\frac{\pi k}{m}\right)\right] \qquad k = 0, 1, \ldots, m$$

$$\text{Parzen window} \quad \lambda_k = \begin{cases} 1 - 6\left(\frac{k}{m}\right)^2 + 6\left(\frac{k}{m}\right)^3 & k = 0, 1, \ldots, m/2 \\ 2\left(1 - \frac{k}{m}\right)^3 & m/2 + 1, \ldots, m. \end{cases}$$

These sets of weights are quite similar and, not surprisingly, give similar results.

We must also choose a value of m. Increasing m leads to the raw periodogram (too noisy), but choosing m too small over-smooths (low variance, but biased). One rule of thumb is to take $m = \sqrt{n}/2$.

9.1.3 Smoothing the raw periodogram

[This approach ("Daniel smoothing") is the default choice in R.]

Instead of using a lag window, we might smooth the raw periodogram. This is based on the assumption that the power spectrum $\mathcal{I}(f)$ will be "similar" for values of f "close" to each other. Let $w = 2w^* + 1$ be an odd positive integer and define

$$\mathcal{I}_S(f_j) = \frac{1}{w} \sum_{k=-w^*}^{w^*} I(f_{j+k}). \tag{9.2}$$

Recall that $I(f)$ is asymptotically unbiased but has constant variance; also $I(f_j)$ is asymptotically independent of $I(f_k)$ if $j \neq k$. Therefore the variance of $\mathcal{I}_S(f)$ is of order $1/w$, but $\mathcal{I}_S(f)$ is biased since

$$\mathbb{E}[\mathcal{I}_S(f)] = \frac{1}{w} \sum_{k=-w^*}^{w^*} \mathbb{E}[I(f_{j+k})] \approx \frac{1}{w} \sum_{k=-w^*}^{w^*} \mathcal{I}_S(f_{j+k}),$$

which only equals $\mathcal{I}_S(f_j)$ if the power spectrum is linear in $[f_{j-w^*}, f_{j+w^*}]$. Therefore, we need to balance large w (small variance) with small w (small bias). One rule of thumb is to use $w \approx \sqrt{n}$. Another suggestion is to use repeated application of the Daniel smoother with small values of w.

In fact, smoothing the periodogram in this way is equivalent to using (9.1) with

$$\lambda_k = \begin{cases} 1 & k = 0 \\ \frac{\sin(wk\pi/n)}{m \sin(k\pi/n)} & k = 1, 2, \ldots, n-1. \end{cases}$$

9.1.4 Spectral window

The procedures in sections 9.1.2 and 9.1.3 can be compared using a quantity called the *spectral window* $K(f)$, the DFT of the lag window. Note that

$$K(f_j) = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} \lambda_k e^{-2\pi i f_j k} \qquad \text{so} \qquad \lambda_k = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} K(f_j) e^{2\pi i f_j k}.$$

The truncated estimate of the power spectrum is

$$\begin{aligned}\mathcal{I}_T(f_j) &= \frac{1}{\sqrt{n}} \sum_{k=-(n-1)}^{n-1} \lambda_k g_k e^{-2\pi i f_j k} \\ &= \frac{1}{\sqrt{n}} \sum_{k=-(n-1)}^{n-1} \left[\frac{1}{\sqrt{n}} \sum_{j'=0}^{n-1} K(f_{j'}) e^{2\pi i f_{j'} k} \right] g_k e^{-2\pi i f_j k} \\ &= \frac{1}{n} \sum_{j'=0}^{n-1} K(f_{j'}) \sum_{k=-(n-1)}^{n-1} g_k e^{-2\pi i (f_j - f_{j'})k} \\ &= \frac{1}{n} \sum_{j'=0}^{n-1} K(f_{j'}) I(f_j - f_{j'}),\end{aligned}$$

so our estimation procedures are all equivalent to smoothing the periodogram using a kernel function K.

There can be a sharp cut-off in the frequency domain (Daniel smoothing) *or* in the time domain (truncation), but not both.

9.2 Spectral leakage

So far, we have assumed that all periodic components are sine waves existing at one of the Fourier frequencies $f_j = j/n$. But this need not be the case.

If $\widehat{X}_j = a_j + ib_j$ for $j \leqslant n/2$, we have seen

$$X_t = \sum_{j<\frac{n}{2}} \frac{2a_j}{\sqrt{n}} \cos(2\pi f_j t) + \sum_{j<\frac{n}{2}} \frac{2b_j}{\sqrt{n}} \sin(2\pi f_j t) + \frac{a_0}{\sqrt{n}} + \frac{a_{n/2}}{\sqrt{n}}(-1)^t \text{ if } n \text{ is even.}$$

If $X_t = \sin(2\pi f t)$ for $f = f_k$, we have

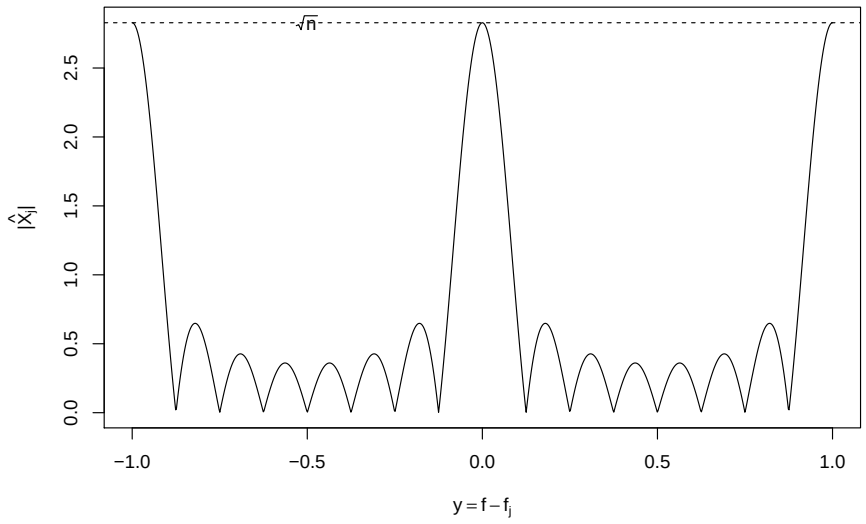
$$\widehat{X}_j = \begin{cases} \frac{\sqrt{n}}{2i} & \text{if } j = k \\ -\frac{\sqrt{n}}{2i} & \text{if } j = -k \\ 0 & \text{else.} \end{cases}$$

What happens if f is not a Fourier frequency? For simplicity, we consider $X_t = e^{2\pi i f t}$ instead of $\sin(2\pi f t)$. Then

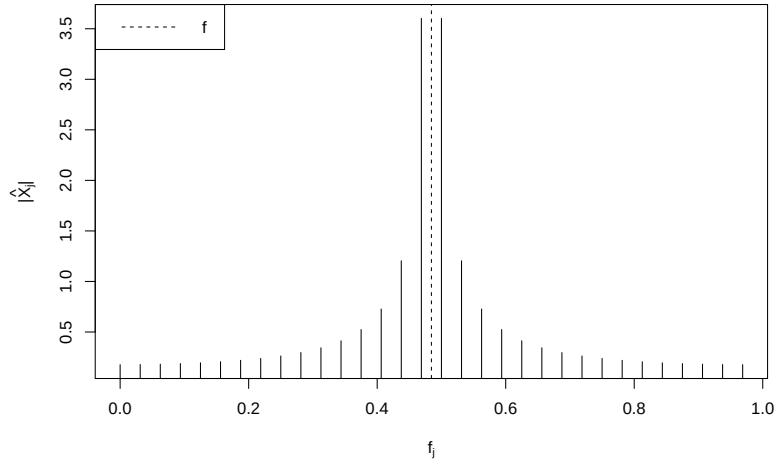
$$\begin{aligned}\widehat{X}_j &= \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} e^{2\pi i f t} e^{-2\pi i f_j t} \\ &= \begin{cases} \sqrt{n} & \text{if } f = f_j \\ 0 & \text{if } f = f_k, k \neq j \\ \frac{1}{\sqrt{n}} \frac{1-e^{2\pi(f-f_j)n}}{1-e^{2\pi i(f-f_j)}} & \text{else.} \end{cases}\end{aligned}$$

In the third case,

$$\begin{aligned}|\widehat{X}_j| &= \left| \frac{1}{\sqrt{n}} \frac{\frac{1}{2i}(e^{-\pi i(f-f_j)n} - e^{\pi i(f-f_j)n})}{\frac{1}{2i}(e^{-\pi i(f-f_j)} - e^{\pi i(f-f_j)})} \underbrace{e^{-\pi i(f-f_j)(n-1)}}_{1\cdot 1=1} \right| \\ &= \frac{1}{\sqrt{n}} \left| \frac{\sin[\pi n(f-f_j)]}{\sin[\pi(f-f_j)]} \right|.\end{aligned}$$



If f is not a Fourier frequency, we will get contributions for several $\widehat{X}_j$. The worst case is $f = (j \pm \frac{1}{2})/n$.



This can make it difficult to resolve contributions from frequencies which are close to each other. There are modified transforms which suffer less from this problem ("Hamming", "data windows").

9.3 Harmonics

What if there is a periodic component which is not a sine wave?

Consider the contribution to $\widehat{X}_j$ from time t:

$$\delta_{j,t} = \frac{1}{\sqrt{n}} X_t e^{-2\pi i f_j t} \qquad 0 \leqslant j \leqslant n/2,$$

and to $\widehat{X}_{2j}$; $f_{2j} = 2j/n = 2f_j$ so

$$\delta_{2j,t} = \frac{1}{\sqrt{n}} X_t e^{-2\pi i 2 f_j t} = \frac{1}{\sqrt{n}} X_t \left(e^{-2\pi i f_j t}\right)^2 = \delta_{j,t} e^{-2\pi i f_j t}$$

$$\delta_{3j,t} = \delta_{j,t} \left(e^{-2\pi i f_j t}\right)^2$$

$$\delta_{mj,t} = \delta_{j,t} \left(e^{-2\pi i f_j t}\right)^{(m-1)} \qquad \begin{cases} j < mj < n/2 \\ 1 < m < n/(2j) \end{cases}$$

So if there is a non-trigonometric periodic component at frequency f_j, its effects will also be felt at frequencies $f_{2j}, f_{3j}, \cdots$. These are called the *harmonics* of the *fundamental frequency* f_j.