MATH60046/MATH70046

Time Series Analysis

Ed Cohen

Room: 536 Huxley

email: e.cohen@imperial.ac.uk

Use Blackboard to obtain all course resources

Department of Mathematics

Imperial College London

180 Queen's Gate, London SW7 2BZ

Video 1

Chapter 1

Introduction

1.1 Module admin and structure

Pre-requisites

Probability for Statistics

Statistical Modelling I (preferable but not essential).

Course materials

Lecture notes

Figures booklets

30 hours of teaching (approx 25 lectures/5 problems class)

Non-assessed quiz questions

5 non-assessed problem sheets

Assessment

90% Exam

10% Coursework

1.2 What are times series?

A time series is a series of data points indexed (or listed) in time order [wikipedia]. Any metric that is measured over time is a time series.

Times series analysis (TSA) could be described as a branch of applied stochastic processes. We start with an indexed family of real-valued random variables

$$\{X_t : \quad t \in T\}$$

where t is the index, here taken to be time (but it could be space). T is called the index set. We have a state space of values of X.

	Possibilities	
State	Time	
Continuous	Continuous	$X(t)$
Continuous	Discrete	X_t
Discrete	Continuous	
Discrete	Discrete	

In addition X could be univariate or multivariate. We shall concentrate on discrete time. Samples are taken at equal intervals.

We wish to use TSA to characterize time series and understand structure. Our job is to make inference on the underlying stochastic process from a single realisation - the observed time series.

Diagram: paths/trajectories

Examples: Figures 1–4, in all cases points are joined for clarity.

[1] wind speed in a certain direction at a location, measured every 0.025s.

[2] monthly average measurements of the flow of water in the Willamette River at Salem, Oregon.

[3] the daily record of the change in average daily frequency that tells us how well an atomic clock keeps time on a day to day basis.

[4] the change in the level of ambient noise in the ocean from one second to the next.

[5] part of the Epstein-Barr Virus DNA sequence (the entire sequence consists of approximately 172,000 base pairs).

[6] Daily US Dollar/Sterling exchange rate and the corresponding returns from 1981 to 1985.

The visual appearances of these datasets are quite different. For example, consider the wind speed and atomic clock data,

Wind speed	Atomic clock
Adjacent points are close in value	Positive values tend to be followed by negative values

For the numerical data, we can illustrate this using lag 1 scatter plots.

Diagram: lag 1 scatter plots

(See Figures 4a and 5). Realizations of the series denoted $x_1, \ldots, x_N$. So plot x_t versus x_{t+1} as t varies from 1 to $N-1$.

From these scatter plots we note the following:

[1] for the wind speed and US dollar series, the values are positively correlated.

[2] Willamette river data is similar, but points are more spread out.

[3] for the atomic clock data, the values are negatively correlated.

[4] for the ocean noise data and the US dollar returns series there is no clear clustering tendency.

We could similarly create lag τ scatter plots by plotting x_t versus $x_{t+\tau}$ for integer τ, but they would be unwieldy to deal with and interpret.

A better approach is to realize that the series $x_1, \ldots, x_N$ can be regarded as a realization of the corresponding random variables $X_1, \ldots, X_N$, and we will proceed by studying the covariance relationships between these random variables.

1.3 A brief aside - covariance and correlation

The concept of covariance and correlation will be crucial in this module, therefore we BRIEFLY recap some of the key ideas.

Covariance

Covariance is a measure of joint variability of two random variables, X and Y say. Defined as

$$\text{cov}(X, Y) \equiv E\left\{(X - E\{X\})(Y - E\{Y\})\right\} = E\{XY\} - E\{X\}E\{Y\}.$$

- Positive covariance $\Rightarrow$ when X is above its mean then Y also tends to be.
- Negative covariance $\Rightarrow$ when X is above its mean then Y tends to be below its mean.
- Zero covariance means there is no relationship of this type and $E\{XY\} = E\{X\}E\{Y\}$

Therefore, this gives a measure of linear dependency between 2 random variables.

NOTE: $\text{cov}(X, X) = \text{var}(X)$

All variance and covariance terms (known as the joint second moments), can be summarized in the *variance-covariance* matrix (also commonly known as just the *covariance matrix*).

Define vector $\mathbf{X} = (X, Y)^T$ with mean $E\{\mathbf{X}\} = \mu = (\mu_X, \mu_Y)^T$ and variances σ_X^2 and σ_Y^2, respectively.

$$\Sigma \equiv E\{(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T\} = \begin{pmatrix} \sigma_X^2 & \text{cov}(X, Y) \\ \text{cov}(X, Y) & \sigma_Y^2 \end{pmatrix}$$

This can be extended to higher dimensions. For a random vector $\mathbf{X} = (X_1, X_2, ..., X_m)$, we have a $m \times m$ covariance matrix $\Sigma = (\sigma_{ij})$ where $\sigma_{ij} = \text{cov}(X_i, X_j)$.

Σ is a symmetric positive semi-definite matrix.

Correlation

Correlation is a normalized measure of covariance. It is useful because covariance is proportional to variance so can be misleading.

Covariance has all the properties of an inner-product, $\text{cov}(\cdot,\cdot) \equiv \langle \cdot, \cdot \rangle$, namely

- Bilinear: $\text{cov}(aX + bY, Z) = a\text{cov}(X, Z) + b\text{cov}(Y, Z)$

- Symmetric: $\text{cov}(X, Y) = \text{cov}(Y, X)$

- Positive definite: $\text{cov}(X, X) \geq 0$ for all X.

This means we can invoke the Cauchy-Swartz inequality:

$$\begin{aligned} |\langle x, y \rangle| &\leq \sqrt{\|x\|^2 \|y\|^2} \\ |\text{cov}(X, Y)| &\leq \sqrt{\text{var}(X)\text{var}(Y)} \end{aligned}$$

We therefore define correlation ρ as

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

CS inequality means $-1 \leq \rho \leq 1$.

Equality holds when there's a perfect linear relationship, i.e. $Y = aX + b$,

$$\begin{aligned} \text{cov}(X, Y) &= \text{cov}(X, aX + b) = a\text{cov}(X, X) = a\sigma_X^2 \\ \text{var}(Y) &= \text{cov}(aX + b, aX + b) = \text{cov}(aX, aX) = a^2\text{var}(X), \end{aligned}$$

Therefore

$$\rho = \frac{a\sigma_X^2}{\sqrt{a^2\sigma_X^2\sigma_X^2}} = \frac{a\sigma_X^2}{|a|\sigma_X^2} = \begin{cases} 1, & \text{when } a > 0 \\ -1, & \text{when } a < 0. \end{cases}$$

In this circumstance we say the random variables are perfectly positively/negatively correlated.

When $\text{cov}(X, Y) = 0$ we say random variables are uncorrelated

Diagram: scatter plots

Formally,

$$\text{cov}(X, Y) = \int\int xyf(x, y)\text{d}x\text{d}y - \int xf(x)\text{d}x \int yf(y)\text{d}y$$

Therefore, if X and Y are independent, then $f(x, y) = f(x)f(y)$ which implies

$$\text{cov}(X, Y) = \int xf(x)\text{d}x \int yf(y)\text{d}y - \int xf(x)\text{d}x \int yf(y)\text{d}y = 0.$$

So,

independence $\Rightarrow$ uncorrelated

uncorrelated $\not\Rightarrow$ independence, in general

BUT, uncorrelated $\Rightarrow$ independence for joint normal distribution

Its context in time series analysis

In time series analysis we will typically be interested in looking at the covariance or correlation within the random process $\{X_t\}$ at two time points, i.e. the covariance or correlation between random variables X_{t_1} and X_{t_2}.

Chapter 2

Real-Valued discrete time stationary processes

Video 4

Video 5

2.1 Stationarity

2.1.1 Joint distribution

Denote the process by $\{X_t\}$. For fixed t, X_t is a random variable (r.v.), and hence there is an associated cumulative probability distribution function (cdf):

$$\mathrm{F}_t(a) = \mathrm{P}(X_t \leq a),$$

and

$$\begin{aligned} \mathrm{E}\{X_t\} &= \int_{-\infty}^{\infty} x \, \mathrm{dF}_t(x) \\ \mathrm{var}\{X_t\} &= \int_{-\infty}^{\infty} (x - \mu_t)^2 \, \mathrm{dF}_t(x). \end{aligned}$$

But we are interested in the relationships between the various r.v.s that form the process. For example, for any t_1 and $t_2 \in T$,

$$\mathrm{F}_{t_1,t_2}(a_1, a_2) = \mathrm{P}(X_{t_1} \leq a_1, X_{t_2} \leq a_2)$$

gives the bivariate cdf. More generally for any $t_1, t_2, \ldots, t_n \in T$,

$$\mathrm{F}_{t_1,t_2,\ldots,t_n}(a_1, a_2, \ldots, a_n) = \mathrm{P}(X_{t_1} \leq a_1, \ldots, X_{t_n} \leq a_n)$$

The class of all stochastic processes is too large to work with in practice. We consider only the subclass of stationary processes.

2.1.2 COMPLETE/STRONG/STRICT stationarity

$\{X_t\}$ is said to be completely stationary if, for all $n \geq 1$, for any $t_1, t_2, \ldots, t_n \in T$, and for any s such that $t_1+s, t_2+s, \ldots, t_n+s \in T$ are also contained in the index set, the joint cdf of $\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$ is the same as that of $\{X_{t_1+s}, X_{t_2+s}, \ldots, X_{t_n+s}\}$ i.e.,

$$\mathrm{F}_{t_1,t_2,\ldots,t_n}(a_1, a_2, \ldots, a_n) = \mathrm{F}_{t_1+s,t_2+s,\ldots,t_n+s}(a_1, a_2, \ldots, a_n),$$

so that the probabilistic structure of a completely stationary process is invariant under a shift in time.

Diagram: strict stationarity

Video 6

2.1.3 SECOND-ORDER/WEAK/COVARIANCE stationarity

Definition 1

$\{X_t\}$ is said to be second-order stationary if, for all $n \geq 1$, for any $t_1, t_2, \ldots, t_n \in T$, and for any s such that $t_1 + s, t_2 + s, \ldots, t_n + s \in T$ are also contained in the index set, all the joint moments of orders 1 and 2 of $\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$ exist, are finite, and equal to the corresponding joint moments of $\{X_{t_1+s}, X_{t_2+s}, \ldots, X_{t_n+s}\}$.

Diagram: second-order stationarity

It is immediate from this definition that,

$$\mathrm{E}\{X_t\} \equiv \mu \quad ; \quad \mathrm{var}\{X_t\} \equiv \sigma^2 \ \ (= \mathrm{E}\{X_t^2\} - \mu^2),$$

are constants independent of t. If we let $s = -t_1$,

$$\begin{aligned} \mathrm{E}\{X_{t_1} X_{t_2}\} &= \mathrm{E}\{X_{t_1+s} X_{t_2+s}\} \\ &= \mathrm{E}\{X_0 X_{t_2-t_1}\}, \end{aligned}$$

and with $s = -t_2$,

$$\begin{aligned} \mathrm{E}\{X_{t_1} X_{t_2}\} &= \mathrm{E}\{X_{t_1+s} X_{t_2+s}\} \\ &= \mathrm{E}\{X_{t_1-t_2} X_0\}. \end{aligned}$$

Hence, $\mathrm{E}\{X_{t_1} X_{t_2}\}$ is a function of the absolute difference $|t_2 - t_1|$ only, similarly, for the covariance between X_{t_1} & X_{t_2}:

$$\mathrm{cov}\{X_{t_1}, X_{t_2}\} = \mathrm{E}\{(X_{t_1} - \mu)(X_{t_2} - \mu)\} = \mathrm{E}\{X_{t_1} X_{t_2}\} - \mu^2.$$

For a discrete time second-order stationary process $\{X_t\}$ we define the autocovariance sequence (acvs) by

$$s_\tau \equiv \mathrm{cov}\{X_t, X_{t+\tau}\} = \mathrm{cov}\{X_0, X_\tau\}.$$

Furthermore, suppose a stochastic process $\{X_t\}$ is not stationary, then by the Definition 1 there must exist some index set $t_1, ..., t_m \in T$ and some s such that not all joint moments of order 1 and 2 are constant. In the case of order 1 moments, that implies $\{X_t\}$ does not have a constant mean. In the case of order 2 moments, there must exist $t_i, t_j \in \{t_1, ..., t_m\}$ such that $E\{X_{t_i} X_{t_j}\} \neq E\{X_{t_i+s} X_{t_j+s}\}$, and therefore $\mathrm{cov}\{X_{t_i}, X_{t_j}\}$ does not depend only on the absolute difference $|t_2 - t_1|$.
An equivalent definition of second-order stationary is therefore as follows.
Definition 2
$\{X_t\}$ is said to be second-order stationary if for all $t \in T$ and all τ such that $t+\tau \in T$, $E\{X_t\}$ is finite and constant for all t, and $s_\tau \equiv \mathrm{cov}\{X_t, X_{t+\tau}\}$ is finite and depends only on τ and not on t.

Diagram: autocovariance

Video 7

2.1.4 Properties and Notation

1. τ is called the lag.

2. $s_0 = \sigma^2$ and $s_{-\tau} = s_\tau$.

3. The autocorrelation sequence (acs) is given by

$$\rho_\tau = \frac{s_\tau}{s_0} = \frac{\text{cov}\{X_t, X_{t+\tau}\}}{\sigma^2}.$$

   The sample or estimated autocorrelation sequence (acs), $\{\hat{\rho}_\tau\}$, for each of our time series are given in Figs. 6 and 7. [We shall see how to compute these in Chapter 4.] Note e.g., that for the Willamette river data X_t and X_{t+6} seem to be negatively correlated, while X_t and X_{t+12} seem positively correlated (consistent with the river flow varying with a period of roughly 12 months).

4. Since ρ_τ is a correlation coefficient, $|s_\tau| \leq s_0$.

   Diagram: autocovariance sequences

Worked example: variogram

For the stationary process $\{X_t\}$ with mean μ, acvs $\{s_\tau\}$ and variance $s_0 = \sigma_0^2$

$$v_\tau = E\{(X_{t+\tau} - X_t)^2\}/2$$

(used in geostatistics) has a maximum of $2\sigma^2$.

5. The sequence $\{s_\tau\}$ is positive semidefinite, i.e., for all $n \geq 1$, for any $t_1, t_2, \ldots, t_n$ contained in the index set, and for any set of nonzero real numbers $a_1, a_2, \ldots, a_n$

$$\sum_{j=1}^{n} \sum_{k=1}^{n} s_{t_j - t_k} a_j a_k \geq 0.$$

Proof

Let

$$\boldsymbol{a} = (a_1, a_2, \ldots, a_n)^{\mathrm{T}}, \quad \boldsymbol{V} = (X_{t_1}, X_{t_2}, \ldots, X_{t_n})^{\mathrm{T}}$$

and let Σ be the variance-covariance matrix of $\boldsymbol{V}$. Its j,k-th element is given by $s_{t_j - t_k} = \mathrm{E}\{(X_{t_j} - \mu)(X_{t_k} - \mu)\}$. Define the r.v.

$$w = \sum_{j=1}^{n} a_j X_{t_j} = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{V}.$$

Then

$$0 \leq \text{var}\{w\} = \text{var}\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{V}\} = \boldsymbol{a}^{\mathrm{T}}\Sigma\boldsymbol{a} = \sum_{j=1}^{n}\sum_{k=1}^{n} s_{t_j - t_k} a_j a_k.$$

6. The variance-covariance matrix of equispaced X's, $(X_1, X_2, \ldots, X_N)^{\mathrm{T}}$ has the form

$$\begin{bmatrix} s_0 & s_1 & \ldots & s_{N-2} & s_{N-1} \\ s_1 & s_0 & \ldots & s_{N-3} & s_{N-2} \\ \vdots & & \ddots & & \\ s_{N-2} & s_{N-3} & \ldots & s_0 & s_1 \\ s_{N-1} & s_{N-2} & \ldots & s_1 & s_0 \end{bmatrix}$$

which is known as a symmetric Toeplitz matrix – all elements on a diagonal are the same. Note the matrix has only N unique elements, $s_0, s_1, \ldots, s_{N-1}$.

7. A stochastic process $\{X_t\}$ is called Gaussian if, for all $n \geq 1$ and for any $t_1, t_2, \ldots, t_n$ contained in the index set, the joint cdf of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ is multivariate Gaussian.

   Diagram: jointly Gaussian

   - 2nd-order stationary Gaussian ⇒ complete stationarity (since MVN completely characterized by 1st and 2nd moments). It is not true in general that 2nd-order stationary ⇒ complete stationarity.
   - Complete stationarity ⇒ 2nd-order stationary in general.

8. The simple term "stationary" will be taken to mean second-order stationary unless stated otherwise.

2.1.5 Examples of discrete stationary processes

Video 8

[1] White noise process

Also known as a purely random process. Let $\{X_t\}$ be a sequence of uncorrelated r.v.s such that

$$\mathrm{E}\{X_t\} = \mu, \qquad \mathrm{var}\{X_t\} = \sigma^2 \qquad \forall t$$

and

$$s_\tau = \begin{cases} \sigma^2 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases} \quad \text{or} \quad \rho_\tau = \begin{cases} 1 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases}$$

forms a basic building block in time series analysis. Very different realizations of white noise can be obtained for different distributions of $\{X_t\}$. Examples are given in Figures 8 and 9 for processes with (a) Gaussian, (b) exponential, (c) uniform and (d) truncated Cauchy distributions.

[2] *q*-th order moving average process MA(q)

X_t can be expressed in the form

$$\begin{aligned} X_t &= \mu - \theta_{0,q}\epsilon_t - \theta_{1,q}\epsilon_{t-1} - \ldots - \theta_{q,q}\epsilon_{t-q} \\ &= \mu - \sum_{j=0}^{q} \theta_{j,q}\epsilon_{t-j}, \end{aligned}$$

where μ and $\theta_{j,q}$'s are constants ($\theta_{0,q} \equiv -1, \theta_{q,q} \neq 0$), and $\{\epsilon_t\}$ is a zero-mean white noise process with variance σ_ϵ^2.

Diagram: moving average process

W.l.o.g. assume $\mathrm{E}\{X_t\} = \mu = 0$.

Then $\mathrm{cov}\{X_t, X_{t+\tau}\} = \mathrm{E}\{X_t X_{t+\tau}\}$.

Recall: $\mathrm{cov}(X, Y) = \mathrm{E}\{(X - \mathrm{E}\{X\})(Y - \mathrm{E}\{Y\})\}$.

Since $\mathrm{E}\{\epsilon_t \epsilon_{t+\tau}\} = 0 \ \ \forall\, \tau \neq 0$ we have for $\tau \geq 0$.

$$\begin{aligned} \mathrm{cov}\{X_t, X_{t+\tau}\} &= \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_{j,q} \theta_{k,q} \mathrm{E}\{\epsilon_{t-j} \epsilon_{t+\tau-k}\} \\ &= \sigma_\epsilon^2 \sum_{j=0}^{q-\tau} \theta_{j,q} \theta_{j+\tau,q} \qquad (k = j + \tau) \\ &\equiv s_\tau, \end{aligned}$$

which does not depend on t. Since $s_\tau = s_{-\tau}$, $\{X_t\}$ is a stationary process with acvs given by

$$s_\tau = \begin{cases} \sigma_\epsilon^2 \sum_{j=0}^{q-|\tau|} \theta_{j,q} \theta_{j+|\tau|,q} & |\tau| \leq q \\ 0 & |\tau| > q \end{cases}$$

N.B. No restrictions were placed on the $\theta_{j,q}$'s to ensure stationarity, though obviously, $|\theta_{j,q}| < \infty, \ j = 1, \ldots, q$.

Examples (see Figures 10 and 11)

$$X_t = \epsilon_t - \theta_{1,1} \epsilon_{t-1} \qquad \mathrm{MA}(1)$$

acvs:

$$s_\tau = \sigma_\epsilon^2 \sum_{j=0}^{1-|\tau|} \theta_{j,1} \theta_{j+|\tau|,1} \qquad |\tau| \leq 1,$$

so,

$$\begin{aligned} s_0 &= \sigma_\epsilon^2 (\theta_{0,1} \theta_{0,1} + \theta_{1,1} \theta_{1,1}) \\ &= \sigma_\epsilon^2 (1 + \theta_{1,1}^2); \end{aligned}$$

and,

$$\begin{aligned} s_1 &= \sigma_\epsilon^2 \theta_{0,1} \theta_{1,1} \\ &= -\sigma_\epsilon^2 \theta_{1,1}. \end{aligned}$$

acs:

$$\rho_\tau = \frac{s_\tau}{s_0}.$$

$$\rho_0 = 1.0; \ \rho_1 = \frac{-\theta_{1,1}}{1 + \theta_{1,1}^2}; \ \rho_2 = \rho_3 = \cdots = 0.$$

(a) $\theta_{1,1} = 1.0, \sigma_\epsilon^2 = 1.0,$

we have,

$$s_0 = 2.0;\ s_1 = -1.0;\ s_2 = s_3 = \cdots = 0.0,$$

giving,

$$\rho_0 = 1.0;\ \rho_1 = -0.5;\ \rho_2 = \rho_3 = \cdots = 0.0.$$

(b) $\theta_{1,1} = -1.0, \sigma_\epsilon^2 = 1.0,$

we have,

$$s_0 = 2.0;\ s_1 = 1.0;\ s_2 = s_3 = \cdots = 0.0,$$

giving,

$$\rho_0 = 1.0;\ \rho_1 = 0.5;\ \rho_2 = \rho_3 = \cdots = 0.0.$$

Note: if we replace $\theta_{1,1}$ by $\theta_{1,1}^{-1}$ the model becomes

$$X_t = \epsilon_t - \frac{1}{\theta_{1,1}}\epsilon_{t-1}$$

and the autocorrelation becomes

$$\rho_1 = \frac{-\frac{1}{\theta_{1,1}}}{1 + \left(\frac{1}{\theta_{1,1}}\right)^2} = \frac{-\theta_{1,1}}{\theta_{1,1}^2 + 1},$$

i.e., is unchanged!!!

We cannot identify the MA(1) process uniquely from its autocorrelation!

[3] ***p*-th order autoregressive process** AR(p) Video 9

$\{X_t\}$ is expressed in the form

$$X_t = \phi_{1,p}X_{t-1} + \phi_{2,p}X_{t-2} + \ldots + \phi_{p,p}X_{t-p} + \epsilon_t,$$

where $\phi_{1,p}, \phi_{2,p}, \ldots, \phi_{p,p}$ are constants ($\phi_{p,p} \neq 0$) and $\{\epsilon_t\}$ is a zero mean white noise process with variance σ_ϵ^2. In contrast to the parameters of an MA(q) process, the $\{\phi_{k,p}\}$ must satisfy certain conditions for $\{X_t\}$ to be a stationary

process – i.e., not all AR(p) processes are stationary (more later).

Examples (Figures 12 and 13)

$$\begin{aligned} X_t &= \phi_{1,1}X_{t-1} + \epsilon_t \quad \text{AR(1) – Markov process} \\ &= \phi_{1,1}\{\phi_{1,1}X_{t-2} + \epsilon_{t-1}\} + \epsilon_t \\ &= \phi_{1,1}^2 X_{t-2} + \phi_{1,1}\epsilon_{t-1} + \epsilon_t \\ &= \phi_{1,1}^3 X_{t-3} + \phi_{1,1}^2\epsilon_{t-2} + \phi_{1,1}\epsilon_{t-1} + \epsilon_t \\ &\vdots \\ &= \sum_{k=0}^{\infty} \phi_{1,1}^k \epsilon_{t-k}. \end{aligned} \tag{2.1}$$

Here we take the initial condition $X_{-N} = 0$ and let $N \to \infty$.

Note $\mathrm{E}\{X_t\} = 0$.

$$\mathrm{var}\{X_t\} = E\left\{\left(\sum_{k=0}^{\infty}\phi^k\epsilon_{t-k}\right)^2\right\} = \sum_{k=0}^{\infty}\sum_{k'=0}^{\infty}\phi^k\phi^{k'}E\{\epsilon_{t-k}\epsilon_{t-k'}\} = \sum_{k=0}^{\infty}\phi^{2k}\sigma_\epsilon^2.$$

For $\mathrm{var}\{X_t\} < \infty$ we must have $|\phi_{1,1}| < 1$, in which case

$$\mathrm{var}\{X_t\} = \frac{\sigma_\epsilon^2}{1-\phi_{1,1}^2}.$$

To find the form of the acvs, we notice that for $\tau > 0$, $X_{t-\tau}$ is a linear function of $\epsilon_{t-\tau}, \epsilon_{t-\tau-1}, \ldots$ and is therefore uncorrelated with ϵ_t. Hence

$$\mathrm{E}\{\epsilon_t X_{t-\tau}\} = 0,$$

so, assuming stationarity and multiplying the defining equation (2.1) by $X_{t-\tau}$:

$$\begin{aligned} X_tX_{t-\tau} &= \phi_{1,1}X_{t-1}X_{t-\tau} + \epsilon_tX_{t-\tau} \\ \Rightarrow \mathrm{E}\{X_tX_{t-\tau}\} &= \phi_{1,1}\mathrm{E}\{X_{t-1}X_{t-\tau}\} \\ \text{i.e., } s_\tau &= \phi_{1,1}s_{\tau-1} = \phi_{1,1}^2 s_{\tau-2} = \ldots = \phi_{1,1}^\tau s_0 \\ \Rightarrow \rho_\tau &= \frac{s_\tau}{s_0} = \phi_{1,1}^\tau. \end{aligned}$$

But ρ_τ is an even function of τ, so we obtain an exponentially decaying sequence

$$\rho_\tau = \phi_{1,1}^{|\tau|} \qquad \tau = 0, \pm1, \pm2, \ldots.$$

Diagram: AR(1) autocovariance sequence

[4] **(p, q)'th order autoregressive-moving average process** ARMA(p, q)

Here $\{X_t\}$ is expressed as

$$X_t = \phi_{1,p} X_{t-1} + \ldots + \phi_{p,p} X_{t-p} + \epsilon_t - \theta_{1,q} \epsilon_{t-1} - \ldots - \theta_{q,q} \epsilon_{t-q},$$

where the $\phi_{j,p}$'s and the $\theta_{j,q}$'s are all constants ($\phi_{p,p} \neq 0; \theta_{q,q} \neq 0$) and again $\{\epsilon_t\}$ is a zero mean white noise process with variance σ_ϵ^2.

Diagram: ARMA(2,1) dependencies

The ARMA class is important as many data sets may be approximated in a more parsimonious way (meaning fewer parameters are needed) by a mixed ARMA model than by a pure AR or MA process.

Worked example: ARMA(p, q)

Suppose that $\{Y_t\}$ is a zero mean stationary $AR(p)$ process:

$$Y_t = \sum_{j=1}^{p} \phi_{j,p} Y_{t-j} + \epsilon_t.$$

Show that the process $\{X_t\}$ given by

$$X_t = \sum_{k=0}^{q} \beta_k Y_{t-k}; \qquad \beta_0 = 1$$

is an ARMA(p, q) process (i.e. a moving average of an AR(p) process is ARMA).

[5] *p*'th order autoregressive conditionally heteroscedastic model ARCH(p)

Video 10

Standard linear models were found to be inappropriate for describing the dependence of financial log-return series of stock indices, share prices, exchange rates etc. New multiplicative noise models were developed. One such is the ARCH(p) model.

Assume we have a time series $\{X_t\}$ that has a variance (volatility) that changes through time,

$$X_t = \sigma_t \varepsilon_t \tag{2.2}$$

where here $\{\varepsilon_t\}$ is a sequence of independent and identically distributed (iid) r.v.s with zero mean and unit variance. (This is stronger than simply uncorrelated). Here, σ_t represents the local conditional standard deviation of the process.

$\{X_t\}$ is ARCH(p) if it satisfies equation (2.2) and

$$\sigma_t^2 = \alpha + \beta_{1,p} X_{t-1}^2 + \ldots + \beta_{p,p} X_{t-p}^2, \tag{2.3}$$

where $\alpha > 0$ and $\beta_{j,p} \geq 0, j = 1, \ldots, p$ (to ensure σ_t^2 is positive).

Example: ARCH(1)

$$\sigma_t^2 = \alpha + \beta_{1,1} X_{t-1}^2$$

Define,

$$v_t = X_t^2 - \sigma_t^2, \quad \Rightarrow \quad \sigma_t^2 = X_t^2 - v_t.$$

So $X_t^2 = \sigma_t^2 + v_t$ and the model can be written as

$$X_t^2 = \alpha + \beta_{1,1} X_{t-1}^2 + v_t,$$

i.e., as an AR(1) model for $\{X_t^2\}$. The errors, $\{v_t\}$, have zero mean, but as $v_t = \sigma_t^2(\epsilon_t^2 - 1)$ the errors are heteroscedastic.

[6] Harmonic with random amplitude (see Figures 14 and 14a) Video 11

Here $\{X_t\}$ is expressed as

$$X_t = \epsilon_t \cos(2\pi f_0 t + \phi)$$

f_0 is a fixed frequency and $\{\epsilon_t\}$ is zero mean white noise with variance σ_ϵ^2.

Case (a) ϕ is constant.

$$\begin{aligned} \mathrm{E}\{X_t\} &= \mathrm{E}\{\epsilon_t \cos(2\pi f_0 t + \phi)\} \\ &= \mathrm{E}\{\epsilon_t\} \cos(2\pi f_0 t + \phi) = 0. \\ \mathrm{var}\{X_t\} &= \mathrm{E}\{X_t^2\} \\ &= \mathrm{E}\{\epsilon_t^2\} \cos^2(2\pi f_0 t + \phi) \\ &= \sigma_\epsilon^2 \cos^2(2\pi f_0 t + \phi). \end{aligned}$$

So the variance depends on t and the process is nonstationary.

Case (b) $\phi \sim U[-\pi, \pi]$ and indep. of $\{\epsilon_t\}$.

$$\mathrm{E}\{X_t\} = \mathrm{E}\{\epsilon_t \cos(2\pi f_0 t + \phi)\} = \mathrm{E}\{\epsilon_t\}\mathrm{E}\{\cos(2\pi f_0 t + \phi)\} = 0.$$

$$\begin{aligned} \mathrm{cov}\{X_t, X_{t+\tau}\} &= \mathrm{E}\{X_t X_{t+\tau}\} \\ &= \mathrm{E}\{\epsilon_t \epsilon_{t+\tau}\}\mathrm{E}\left\{\cos(2\pi f_0 t + \phi)\cos(2\pi f_0(t+\tau) + \phi)\right\} \end{aligned}$$

Since $\{\epsilon_t\}$ is white noise we have,

$$\mathrm{E}\{\epsilon_t \epsilon_{t+\tau}\} = \begin{cases} \sigma_\epsilon^2 & \text{if } \tau = 0, \\ 0 & \text{if } \tau \neq 0, \end{cases}$$

So, for $\tau \neq 0$, $\mathrm{cov}\{X_t, X_{t+\tau}\} = 0$, while for $\tau = 0$,

$$\mathrm{cov}\{X_t, X_t\} = s_0 = \sigma_\epsilon^2 \mathrm{E}\{\cos^2(2\pi f_0 t + \phi)\}.$$

Now,

$$\begin{aligned} \mathrm{E}\{\cos^2(2\pi f_0 t + \phi)\} &= \int_{-\pi}^{\pi} \cos^2(2\pi f_0 t + \phi) \frac{1}{2\pi}\, \mathrm{d}\phi \\ &= \frac{1}{2}\int_{-\pi}^{\pi} [1 + \cos(4\pi f_0 t + 2\phi)]\, \frac{1}{2\pi}\, \mathrm{d}\phi \\ &= \frac{1}{2}. \end{aligned}$$

So,

$$s_0 = \sigma_\epsilon^2/2,$$

and the process is stationary.

The random phase idea is illustrated in Figure 14a: the random point at which data collection started corresponds to breaking-in to the 'sinusoidal-like' behaviour at a random point, which equates to a random phase.

2.2 Trend removal and seasonal adjustment

There are certain, quite common, situations where the observations exhibit a trend – a tendency to increase or decrease slowly steadily over time – or may fluctuate in a periodic/seasonal manner. The model is modified to

$$X_t = \mu_t + Y_t$$

μ_t = time dependent mean.

Y_t = zero mean stationary process.

Diagram: trend and seasonal model

Example CO_2 data

$X_t =$ monthly atmospheric CO_2 concentrations expressed in parts per million (ppm) derived from in situ air samples collected at Mauna Loa observatory, Hawaii. Monthly data from May 1988 – December 1998, giving $N = 128$.

The data is plotted in Figure 15. We can see both a trend and periodic/seasonal effects.

2.2.1 Trend adjustment

Represent a simple linear trend by $\alpha + \beta t$. So take $X_t = \alpha + \beta t + Y_t$. At least two possible approaches:

(a) Estimate α and β by least squares, and work with the residuals

$$\hat{Y}_t = X_t - \hat{\alpha} - \hat{\beta} t.$$

For the CO_2 data these are shown in the middle plot of figure 15.

(b) Take first differences:

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

For the CO_2 data these are shown in the bottom plot of figure 15.

Note: if $\{Y_t\}$ is stationary so is $\{Y_t^{(1)}\}$

In the case of linear trend, if we difference again:

$$\begin{aligned} X_t^{(2)} &= X_t^{(1)} - X_{t-1}^{(1)} = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\ &= (\beta + Y_t - Y_{t-1}) - (\beta + Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2}, \quad (\equiv Y_t^{(1)} - Y_{t-1}^{(1)} = Y_t^{(2)}), \end{aligned}$$

so that the effect of $\mu_t (= \alpha + \beta t)$ has been completely removed.

If μ_t is a polynomial of degree $(d-1)$ in t, then dth differences of μ_t will be zero ($d = 2$ for linear trend). Further,

$$\begin{aligned} X_t^{(d)} &= \sum_{k=0}^{d} \binom{d}{k} (-1)^k X_{t-k} \\ &= \sum_{k=0}^{d} \binom{d}{k} (-1)^k Y_{t-k}. \end{aligned}$$

There are other ways of writing this. Define the difference operator

$$\Delta = (1 - B)$$

where $BX_t = X_{t-1}$ is the *backward shift operator* (sometimes known as the *lag operator* L – especially in econometrics). Then,

$$X_t^{(d)} = \Delta^d X_t = \Delta^d Y_t.$$

For example, for $d = 2$:

$$\begin{aligned} X_t^{(2)} &= (1-B)^2 X_t = (1-B)(X_t - X_{t-1}) \\ &= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\ &= (\beta + Y_t - Y_{t-1}) - (\beta + Y_{t-1} - Y_{t-2}) \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= (1-B)^2 Y_t = \Delta^2 Y_t. \end{aligned}$$

This notation can be incorporated into the ARMA set up. Recall if $\{X_t\}$ is ARMA(p, q),

$$X_t = \phi_{1,p} X_{t-1} + \ldots + \phi_{p,p} X_{t-p} + \epsilon_t - \theta_{1,q}\epsilon_{t-1} - \ldots - \theta_{q,q}\epsilon_{t-q},$$

$$\begin{aligned} X_t - \phi_{1,p} X_{t-1} - \ldots - \phi_{p,p} X_{t-p} &= \epsilon_t - \theta_{1,q}\epsilon_{t-1} - \ldots - \theta_{q,q}\epsilon_{t-q} \\ (1 - \phi_{1,p}B - \phi_{2,p}B^2 - \ldots - \phi_{p,p}B^p) X_t &= (1 - \theta_{1,q}B - \theta_{2,q}B^2 - \ldots - \theta_{q,q}B^q)\epsilon_t \end{aligned}$$

$$\Phi(B) X_t = \Theta(B)\epsilon_t.$$

Here

$$\begin{aligned} \Phi(B) &= 1 - \phi_{1,p}B - \phi_{2,p}B^2 - \ldots - \phi_{p,p}B^p \\ \text{and} \quad \Theta(B) &= 1 - \theta_{1,q}B - \theta_{2,q}B^2 - \ldots - \theta_{q,q}B^q \end{aligned}$$

are known as the *associated* or *characteristic polynomials*.

Further, we can generalize the class of ARMA models to include differencing to account for certain types of non-stationarity, namely, X_t is called ARIMA(p, d, q) if

$$\begin{aligned} \Phi(B)(1-B)^d X_t &= \Theta(B)\epsilon_t, \\ \text{or} \quad \Phi(B)\Delta^d X_t &= \Theta(B)\epsilon_t. \end{aligned}$$

2.2.2 Seasonal adjustment

The model is

$$X_t = \nu_t + Y_t$$

where

$\nu_t =$ seasonal component,

$Y_t =$ zero mean stationary process.

Presuming that the seasonal component maintains a constant pattern over time with period s, there are again several approaches to removing ν_t.

Diagram: seasonal behaviour

A popular approach used by Box & Jenkins is to use the operator $(1 - B^s)$:

$$\begin{aligned} X_t^{(s)} &= (1 - B^s)X_t = X_t - X_{t-s} \\ &= (\nu_t + Y_t) - (\nu_{t-s} + Y_{t-s}) \\ &= Y_t - Y_{t-s} \end{aligned}$$

since ν_t has period s (and so $\nu_{t-s} = \nu_t$).

Figure 16 shows this technique applied to the CO_2 data – most of the seasonal structure and trend has been removed by applying the seasonal operator after the differencing operator:

$$(1 - B^{12})(1 - B)X_t.$$

Worked example: trend and seasonality

Consider the model

$$X_t = (\beta_0 + \beta_1 t)\nu_t + Y_t$$

where Y_t is a zero-mean stationary process. Show that

$$W_t = (1 - B^{12})^2 X_t$$

only involves $\{Y_t\}$.

2.3 The General Linear Process and the stationarity and invertability of AR/MA/ARMA processes

Video 14

2.3.1 General Linear Process

Consider a process of the form

$$X_t = \sum_{k=-\infty}^{\infty} g_k \epsilon_{t-k},$$

where $\{\epsilon_t\}$ is a purely random white noise process, and $\{g_k\}$ is a given sequence of real-valued constants satisfying $\sum_{k=-\infty}^{\infty} g_k^2 < \infty$, which ensures that $\{X_t\}$ has finite variance.

Note: this processes breaks causality with the $\{X_t\}$ depending on future values of $\{\epsilon_t\}$. We therefore restrict ourselves to

$$g_{-1}, g_{-2}, \ldots = 0,$$

and we obtain what is called the General Linear Process (GLP)

$$X_t = \sum_{k=0}^{\infty} g_k \epsilon_{t-k},$$

where X_t depends only on present and past values $\epsilon_t, \epsilon_{t-2}, \epsilon_{t-2}, \ldots$ of the purely random process.

The GLP is stationary because

$$\begin{aligned}
E\{X_t\} &= \sum_{k=0}^{\infty} g_k E\{\epsilon_{t-k}\} = 0 \\
\text{cov}\{X_t, X_{t+\tau}\} &= E\{X_t X_{t+\tau}\} \\
&= E\left\{\sum_{k=0}^{\infty} g_k \epsilon_{t-k} \sum_{k'=0}^{\infty} g_{k'} \epsilon_{t+\tau-k'}\right\} \\
&= \sum_{k=0}^{\infty}\sum_{k'=0}^{\infty} g_k g_{k'} E\{\epsilon_{t-k}\epsilon_{t+\tau-k'}\} \\
&= \sigma_\epsilon^2 \sum_{k=0}^{\infty} g_k g_{k+\tau}.
\end{aligned}$$

This clearly depends only on τ. Furthermore, we know $|\rho_\tau| \leq 1$, so

$$|s_\tau| = |\text{cov}\{X_t, X_{t+\tau}\}| \leq \sigma_X^2 = \sigma_\epsilon^2 \sum_k g_k^2 < \infty,$$

so the covariance is bounded also. Hence, a GLP is stationary.

If we can write an AR or ARMA process in GLP form then we know it is stationary. NOTE: we have already shown MA processes are stationary, moreover, a MA process is already in GLP form. Video 15

To do this, we introduce the "z-polynomial"

$$G(z) = \sum_{k=0}^{\infty} g_k z^k,$$

where $z \in \mathbb{C}$. Note $X_t = G(B)\epsilon_t$.

We will be dealing with z-polynomials of the form

$$G(z) = \frac{G_1(z)}{G_2(z)}, \quad \text{say.}$$

Call the roots of $G_2(z)$ (the "poles" of $G(z)$) in the complex plane $z_1, z_2, \ldots, z_p$, where the zeros are ordered so that $z_1, \ldots, z_k$ are inside and $z_{k+1}, \ldots, z_p$ are outside the unit circle $|z| = 1$. Then,

$$\begin{aligned} \frac{1}{G_2(z)} = \sum_{j=1}^{p} \frac{A_j}{z - z_j} &= \sum_{j=1}^{k} \frac{A_j}{z} \times \frac{1}{\left(1 - \frac{z_j}{z}\right)} + \sum_{j=k+1}^{p} \frac{A_j}{z_j} \times \frac{-1}{\left(1 - \frac{z}{z_j}\right)} \\ &= \sum_{j=1}^{k} \frac{A_j}{z} \sum_{l=0}^{\infty} \left(\frac{z_j}{z}\right)^l - \sum_{j=k+1}^{p} \frac{A_j}{z_j} \sum_{l=0}^{\infty} \left(\frac{z}{z_j}\right)^l \end{aligned}$$

which is uniformly convergent for $|z| = 1$. Replace z by the backshift operator B and apply to $\{\epsilon_t\}$:

$$\begin{aligned} \left\{\frac{1}{G_2(B)}\right\}\epsilon_t &= \left\{\sum_{j=1}^{k} A_j B^{-1} \sum_{l=0}^{\infty} z_j^l B^{-l} - \sum_{j=k+1}^{p} A_j z_j^{-1} \sum_{l=0}^{\infty} z_j^{-l} B^l\right\}\epsilon_t \\ &= \sum_{j=1}^{k} A_j \sum_{l=0}^{\infty} z_j^l \epsilon_{t+l+1} - \sum_{j=k+1}^{p} A_j \sum_{l=0}^{\infty} z_j^{-l-1} \epsilon_{t-l}. \end{aligned}$$

Hence, if all the roots of $G_2(z)$ are outside the unit circle (i.e. all the poles of $G(z)$ are outside the unit circle) only past and present values of $\{\epsilon_t\}$ are involved and the GLP exists.

Another way of stating this is that

$$G(z) < \infty \qquad |z| \leq 1$$

i.e., $G(z)$ is analytic inside and on the unit circle.

So, all the

$$\begin{cases} \text{poles of } G(z) \text{ lie outside the unit circle} \\ \text{roots (zeros) of } G^{-1}(z) \text{ lie outside the unit circle} \end{cases}$$

2.3.2 Stationarity

For the AR(p) process

$$\begin{aligned} \Phi(B)X_t &= \epsilon_t \\ \Rightarrow X_t &= \Phi^{-1}(B)\epsilon_t = G(B)\epsilon_t, \end{aligned}$$

so that $G(z) = \Phi^{-1}(z)$. Thus the model is stationary if

$$G(z) < \infty, \quad |z| \leq 1.$$

$\Rightarrow$ All the poles of $G(z)$ are outside the unit circle.

Hence the requirement for stationarity is that all the roots of $\Phi(z)$ must lie outside the unit circle.

For the ARMA(p, q) process

$$\begin{aligned} \Phi(B)X_t &= \Theta(B)\epsilon_t \\ \Rightarrow X_t &= \frac{\Theta(B)}{\Phi(B)}\epsilon_t = G(B)\epsilon_t, \end{aligned}$$

so that $G(z) = \frac{\Theta(z)}{\Phi(z)}$. Thus the model is stationary if

$$G(z) < \infty, \quad |z| \leq 1.$$

$\Rightarrow$ All the poles of $G(z)$ are outside the unit circle.

Hence the requirement for stationarity is that all the roots of $\Phi(z)$ must lie outside the unit circle.

For the MA(q) process

$$X_t = \Theta(B)\epsilon_t = G(B)\epsilon_t$$

and since $G(B) = \Theta(B)$ is a polynomial of finite order $G(z) < \infty$, $|z| \leq 1$, automatically.

Video 16

2.3.3 Invertibility

We say a process is invertible if it can be written in AR form (AR processes are therefore invertible by definition).
Consider inverting the GLP into autoregressive form

$$\begin{aligned} X_t &= \sum_{k=0}^{\infty} g_k \epsilon_{t-k} \\ &= \sum_{k=0}^{\infty} g_k B^k \epsilon_t \\ X_t &= G(B)\epsilon_t \\ \Rightarrow \quad G^{-1}(B)X_t &= \epsilon_t \end{aligned}$$

The expansion of $G^{-1}(B)$ in powers of B gives the required autoregressive form provided $G^{-1}(B)$ admits a causal power series expansion

$$G^{-1}(z) = \sum_{k=0}^{\infty} h_k z^k.$$

Our requirement now is that $G^{-1}(z)$ is analytic for $|z| \leq 1$. Thus the model is invertible if

$$G^{-1}(z) < \infty, \quad |z| \leq 1.$$

$\Rightarrow$ All the poles of $G^{-1}(z)$ are outside the unit circle. Equivalently, all roots of $G(z)$ are outside the unit circle.

Consider the MA(q) model

$$X_t = \Theta(B)\epsilon_t,$$

then,

$$\Theta^{-1}(B)X_t = \epsilon_t$$

and in general, the expansion of $\Theta^{-1}(B)$ is a polynomial of infinite order. Hence, $G^{-1}(z) = \Theta^{-1}(z)$, and so the invertibility condition is that $\Theta(z)$ has no roots inside or on the unit circle; i.e. all the roots of $\Theta(z)$ lie outside the unit circle.

For the ARMA(p, q) process

$$\begin{aligned} \Phi(B)X_t &= \Theta(B)\epsilon_t \\ \Rightarrow \frac{\Phi(B)}{\Theta(B)}X_t &= \epsilon_t = \epsilon_t, \end{aligned}$$

so that $G^{-1}(z) = \frac{\Phi(z)}{\Theta(z)}$. Thus the model is invertible if

$$G^{-1}(z) < \infty, \quad |z| \leq 1.$$

$\Rightarrow$ All the poles of $G^{-1}(z)$ are outside the unit circle.
Hence the requirement for invertibility is that all the roots of $\Theta(z)$ must lie outside the unit circle.

2.3.4 Summary and examples

	AR(p)	MA(q)	ARMA(p, q)
Stationarity	Roots of $\Phi(z)$ outside $\|z\| \leq 1$	Always stationary	Roots of $\Phi(z)$ outside $\|z\| \leq 1$
Invertibility	Always invertible	Roots of $\Theta(z)$ outside $\|z\| \leq 1$	Roots of $\Theta(z)$ outside $\|z\| \leq 1$

Diagram: moving between different processes

Video 17

Worked example: stationarity Determine whether the following AR process is stationary

$$X_t = \frac{5}{2}X_{t-1} - X_{t-2} + \epsilon_t.$$

Worked example: invertability

Determine whether the following MA process is invertible

$$X_t = \epsilon_t - 1.3\epsilon_{t-1} + 0.4\epsilon_{t-2}.$$

Worked example: stationarity and invertability

Determine whether the following model is stationary and/or invertible,

$$X_t = 1.3X_{t-1} - 0.4X_{t-2} + \epsilon_t - 1.5\epsilon_{t-1}.$$

Worked example: stationarity, invertability and autocovariance

Consider the ARMA(1,1) model

$$X_t = \phi X_{t-1} + \epsilon_t - \theta \epsilon_{t-1}$$

1. State conditions on ϕ and θ for $\{X_t\}$ to be both stationary and invertible.

2. In the case where it is stationary, express $\{X_t\}$ as a GLP.

3. Determine the autocovariance sequence $\{s_\tau\}$ of $\{X_t\}$.

Chapter 3

Video 18

Spectral analysis of discrete time stationary processes

Spectral analysis is a study of the frequency domain characteristics of a process, and describes the contribution of each frequency to the variance of the process.

Video 19

3.1 Spectral representation for discrete time stationary processes

3.1.1 The spectral representation theorem

Let us define a complex-valued "jump" process $\{Z(f)\}$ on $[-1/2, 1/2]$

$$dZ(f) \equiv \begin{cases} Z(f+df) - Z(f), & 0 \le f < 1/2; \\ 0, & f = 1/2; \\ dZ^*(-f), & -1/2 \le f < 0, \end{cases}$$

where df is a small positive increment. If the intervals $[f, f+df]$ and $[f', f'+df']$ are non-intersecting subintervals of $[-1/2, 1/2]$, then the r.v.'s $dZ(f)$ and $dZ(f')$ are uncorrelated. We say that the process has *orthogonal increments*, and the process itself is called an *orthogonal process* – this orthogonality result is very important.

Diagram: orthogonal increment process

Let $\{X_t\}$ be a real-valued discrete time stationary process, with zero mean. The *spectral representation theorem* states that there exists an orthogonal process $\{Z(f)\}$, defined on $[-1/2, 1/2]$, such that

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi ft}\, dZ(f)$$

for all integers t. The process $\{Z(f)\}$ has the following properties:

[1] $\mathrm{E}\{dZ(f)\} = 0, \ \ \forall \ \ |f| \leq 1/2.$

[2] $\mathrm{E}\{|dZ(f)|^2\} \equiv dS^{(I)}(f)$, say, $\forall \ \ |f| \leq 1/2$, where $S^{(I)}(f)$ is called the integrated spectrum of $\{X_t\}$.

[3] For any two distinct frequencies f and $f' \in (-1/2, 1/2]$

$$\mathrm{cov}\{dZ(f'), dZ(f)\} = \mathrm{E}\{dZ^*(f')dZ(f)\} = 0.$$

The spectral representation

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi ft}\, dZ(f) = \int_{-1/2}^{1/2} e^{i2\pi ft}\, |dZ(f)| e^{i \arg\{dZ(f)\}},$$

means that we can represent <u>any</u> discrete stationary process as an "infinite" sum of complex exponentials at frequencies f with associated random amplitudes $|dZ(f)|$ and random phases $\arg\{dZ(f)\}$.

Diagram: spectral representation

The orthogonal increments property can be used to define the relationship between the autocovariance sequence $\{s_\tau\}$ and the integrated spectrum $S^I(f)$:

$$\begin{aligned} s_\tau = \mathrm{E}\{X_t X_{t+\tau}\} &= \mathrm{E}\{X_t^* X_{t+\tau}\} = \mathrm{E}\int_{-1/2}^{1/2} e^{-i2\pi f' t}\, dZ^*(f') \int_{-1/2}^{1/2} e^{i2\pi f(t+\tau)}\, dZ(f) \\ &= \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} e^{i2\pi(f-f')t} e^{i2\pi f\tau} \mathrm{E}\{dZ^*(f') dZ(f)\}. \end{aligned}$$

Because of the orthogonal increments property,

$$\mathrm{E}\{dZ^*(f') dZ(f)\} = \begin{cases} dS^{(I)}(f) & f = f' \\ 0 & f \neq f' \end{cases}$$

so

$$s_\tau = \int_{-1/2}^{1/2} e^{i2\pi f\tau}\, dS^{(I)}(f),$$

which shows that the integrated spectrum determines the acvs for a stationary process. If in fact $S^{(I)}(f)$ is differentiable everywhere with a derivative denoted by $S(f)$ we have

$$\mathrm{E}\{|dZ(f)|^2\} = dS^{(I)}(f) = S(f)\, df.$$

The function $S(\cdot)$ is called the spectral density function (sdf). Hence

$$s_\tau = \int_{-1/2}^{1/2} S(f) e^{i2\pi f\tau}\, df.$$

But, from standard Fourier theory, a square summable deterministic sequence $\{g_t\}$ *say has the Fourier representation*

$$g_t = \int_{-1/2}^{1/2} G(f) e^{i2\pi ft}\, df,$$

where

$$G(f) = \sum_{t=-\infty}^{\infty} g_t e^{-i2\pi ft}.$$

If we assume that $S(f)$ is square integrable, then $S(f)$ is the Fourier transform (FT) of $\{s_\tau\}$,

$$S(f) = \sum_{\tau=-\infty}^{\infty} s_\tau e^{-i2\pi f\tau}.$$

Hence,

$$\{s_\tau\} \longleftrightarrow S(f),$$

i.e., $\{s_\tau\}$ and $S(f)$ are a FT pair.

Video 20

3.1.2 Spectral Density Function

Subject to its existence, $S(\cdot)$ has the following interpretation: $S(f)\,df$ is the average contribution (over all realizations) to the power from components with frequencies in a small interval about f. The power – or variance – is

$$\sigma^2 = \mathrm{var}\{X_t\} = \int_{-1/2}^{1/2} S(f)\, df.$$

Hence, $S(f)$ is often called the power spectral density function or just power spectrum.

Diagram: integrated spectrum and spectral density function

Properties: (assuming existence)

[1] $S^{(I)}(f) = \int_{-1/2}^{f} S(f')\,df'$.

[2] $0 \leq S^{(I)}(f) \leq \sigma^2; \quad S(f) \geq 0.$

[3] $S^{(I)}(-1/2) = 0; \quad S^{(I)}(1/2) = \sigma^2; \quad \int_{-1/2}^{1/2} S(f)\,df = \sigma^2.$

[4] $f < f' \Rightarrow S^{(I)}(f) \leq S^{(I)}(f'); \quad S(-f) = S(f).$

Except, basically, for the scaling factor σ^2, $S^{(I)}(f)$ has all the properties of a probability distribution function, and hence is sometimes called a spectral distribution function.

Diagram: interpreting a spectral density function

White noise spectrum

Recall that a white noise process $\{\epsilon_t\}$ has acvs:

$$s_\tau = \begin{cases} \sigma_\epsilon^2 & \tau = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the spectrum of a white noise process is given by:

$$S_\epsilon(f) = \sum_{\tau=-\infty}^{\infty} s_\tau e^{-i2\pi f\tau} = s_0 = \sigma_\epsilon^2,$$

i.e., white noise has a constant spectrum.

Video 21

3.1.3 Classification of Spectra

For most practical purposes any integrated spectrum, $S^{(I)}(f)$ can be written as

$$S^{(I)}(f) = S_1^{(I)}(f) + S_2^{(I)}(f)$$

where the $S_j^{(I)}(f)$'s are nonnegative, nondecreasing functions with $S_j^{(I)}(-1/2) = 0$ and are of the following types:

[1] $S_1^{(I)}(\cdot)$ is absolutely continuous, i.e., its derivative exists for almost all f and is equal almost everywhere to an sdf $S(\cdot)$ such that

$$S_1^{(I)}(f) = \int_{-1/2}^{f} S(f')df'.$$

[2] $S_2^{(I)}(\cdot)$ is a step function with jumps of size $\{p_l : \ l = 1, 2, \ldots\}$ at the points $\{f_l : \ l = 1, 2, \ldots\}$.

Diagram: continuous and step function spectra

We consider the integrated spectrum for two 'pure' forms :

case (a): $S_1^{(I)}(f) \geq 0; S_2^{(I)}(f) = 0.$

$\{X_t\}$ is said to have a purely continuous spectrum and $S(f)$ is absolutely integrable, with

$$\int_{-1/2}^{1/2} S(f)\cos(2\pi f\tau)\, df \quad \text{and} \quad \int_{-1/2}^{1/2} S(f)\sin(2\pi f\tau)\, df \to 0,$$

as $|\tau| \to \infty$. [This is known as the Riemann-Lebesgue theorem]. But,

$$s_\tau = \int_{-1/2}^{1/2} e^{i2\pi f\tau} S(f)\, df = \int_{-1/2}^{1/2} S(f)\cos(2\pi f\tau)\, df + i\int_{-1/2}^{1/2} S(f)\sin(2\pi f\tau)\, df.$$

Hence $s_\tau \to 0$ as $|\tau| \to \infty$. In other words, the acvs diminishes to zero (called "mixing condition").

case (b): $S_1^{(I)}(f) = 0; S_2^{(I)}(f) \geq 0.$

Here the integrated spectrum consists entirely of a step function, and the $\{X_t\}$ is said to have a purely discrete spectrum or a line spectrum. The acvs for a process with a line spectrum never damps down to 0.

Examples

case (a): white noise, ARMA process.

case (b): harmonic process.

Figs. 18 and 19 give examples of processes with purely continuous and purely discrete spectra. Note that other combinations are possible:

	$S_1^{(I)}(\cdot)$	$S_2^{(I)}(\cdot)$
Purely cts.	≥ 0	$= 0$
Purely disc.	$= 0$	≥ 0
Mixed	Non-white	≥ 0
Discrete	white	≥ 0

An example of a process with a discrete spectrum is a harmonic with additive white noise.

Diagram: example

3.1.4 Spectral density function vs. autocovariance function

The sdf and acvs contain the same amount of information in that if we know one of them, we can calculate the other. However, they are often not equally informative. The sdf usually proves to be the more sensitive and interpretable diagnostic or exploratory tool.

Figure 20 show the sdf and acvs for two different processes - one with two spectral peaks and one with three. The sdf is able to distinguish between the processes while the acvs's are not noticeably different. [NB: dB $= 10\log_{10}(\text{power})$].

3.2 Sampling and Aliasing

Video 22

Why does the spectral representation of a stationary process only include frequencies in $[-1/2, 1/2]$? This is because the values of $\exp(i2\pi ft)$ and $\exp(i2\pi(f \pm k)t)$, $k \in \mathbb{Z}$, are identical, hence it is only necessary to have frequencies in the range $[-1/2, 1/2]$; all other frequencies are redundant.

Diagram: equivalent frequencies

Note:

$$S(f+1) = \sum_{\tau=-\infty}^{\infty} s_\tau e^{i2\pi(f+1)\tau} = \sum_{\tau=-\infty}^{\infty} s_\tau e^{i2\pi f\tau} e^{i2\pi\tau} = S(f).$$

So far we have only been looking at unit-less discrete time. Suppose we had ob-

servations every second ($\Delta t = 1s$) then the interval is $[-1/2, 1/2]$Hz. Observations every minute ($\Delta t = 1$min) and it becomes $[-1/2, 1/2]\text{min}^{-1}$. But observing every minute is the same as observing every 60 seconds ($\Delta t = 60s$) and $[-1/2, 1/2]\text{min}^{-1}$ is the same as $[-1/120, 1/120]$Hz. The general version of the spectral representation is

$$X_t = \int_{-f_{\mathcal{N}}}^{f_{\mathcal{N}}} e^{i2\pi f t \Delta t} dZ(f)$$

where $f_{\mathcal{N}} \equiv 1/(2\Delta t)$ is called the Nyquist frequency.

Diagram: equivalent frequencies

We have also so far we have only looked at discrete time series $\{X_t\}$. However, such a process is usually obtained by sampling a continuous time process at equal intervals Δt, i.e., for a sampling interval $\Delta t > 0$ and an arbitrary time offset t_0, we can define a discrete time process through

$$X_t \equiv X(t_0 + t\Delta t), \qquad t = 0, \pm 1, \pm 2, \ldots.$$

There also exists a spectral representation theorem for continuous time processes

$$X_t = \int_{-\infty}^{\infty} e^{i2\pi f t \Delta t} dZ(f)$$

Diagram: aliasing

If $\{X(t)\}$ is a stationary process with, say, sdf $S_{X(t)}(\cdot)$ and acvf $s(\tau)$, then $\{X_t\}$ is also a stationary process with, say, sdf $S_{X_t}(\cdot)$ and acvs $\{s_\tau\}$. It can be shown that

$$S_{X_t}(f) = \sum_{k=-\infty}^{\infty} S_{X(t)}\left(f + \frac{k}{\Delta t}\right) \quad \text{for} \quad |f| \leq \frac{1}{2\Delta t}.$$

Thus, the discrete time sdf at f is the sum of the continuous time sdf at frequencies $f \pm \frac{k}{\Delta t}$, $k = 0, 1, 2, \ldots$.

This formula can be interpreted as "fold $S_{X(t)}(f)$ about the Nyquist frequency, and add" (see Figure 21a). One translation of the English term "aliasing" in German is "faltung" meaning folding.

If $S_{X(t)}$ is essentially zero for $|f| > 1/(2\Delta t)$ we can expect good correspondence between $S_{X_t}(f)$ and $S_{X(t)}(f)$ for $|f| \leq 1/(2\Delta t)$ (since $S_{X(t)}(f \pm k/(2\Delta t)) \approx 0$ for $k = 1, 2, \ldots$). If $S_{X(t)}$ is large for some $|f| > 1/(2\Delta t)$, the correspondence can be quite poor, and an estimate of S_{X_t} will not tell us much about $S_{X(t)}$.

Figure 21 illustrates how sampling at a particular rate can't differentiate between different frequency waves. However, if you sample at a rate commensurate with the highest frequency wave present (the bottom one in the plot) — i.e., take samples frequently enough that you are sampling within single oscillations of that highest frequency wave — then you can distinguish the cases.

Worked example: aliasing

A continuous time process $X(t)$, t in seconds, has sdf

$$S_{X(t)}(f) = \begin{cases} 1 - 2|f| & |f| \leq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

with f in Hz. It is sampled with a sample interval $\Delta t = 2$s to produce $\{X_t\}$. What is the sdf of $\{X_t\}$?

3.3 Linear filtering

Video 23

Diagram: motivating example - radio receiver

3.3.1 Defining a linear filter

A digital filter maps a sequence to another sequence. A digital filter L that transforms an input sequence $\{x_t\}$ into an output sequence $\{y_t\}$ is called a linear time-invariant (LTI) digital filter if it has the following three properties:

[**1**] Scale-preservation:

$$L\{\{\alpha x_t\}\} = \alpha L\{\{x_t\}\}.$$

[**2**] Superposition:

$$L\{\{x_{t,1} + x_{t,2}\}\} = L\{\{x_{t,1}\}\} + L\{\{x_{t,2}\}\}.$$

[**3**] Time invariance:

If

$$L\{\{x_t\}\} = \{y_t\}, \quad \text{then} \quad L\{\{x_{t+\tau}\}\} = \{y_{t+\tau}\}.$$

Here τ is integer-valued, and the notation $\{x_{t+\tau}\}$ refers to the sequence whose t-th element is $x_{t+\tau}$.

From now on we shall drop the double brackets and the input and output being sequences will be implicitly understood, i.e., we will use $L\{x_t\} = y_t$ as shorthand for $L\{\{x_t\}\} = \{y_t\}$.

Suppose we use a sequence with t-th element $\exp(i2\pi ft)$ as the input to a LTI digital filter: Let $\xi_{f,t} = e^{i2\pi ft}$, and let $y_{f,t}$ denote the output:

$$y_{f,t} = L\{\xi_{f,t}\}.$$

By properties [1] and [3]:

$$y_{f,t+\tau} = L\{\xi_{f,t+\tau}\} = L\{e^{i2\pi f\tau}\xi_{f,t}\} = e^{i2\pi f\tau}L\{\xi_{f,t}\} = e^{i2\pi f\tau}y_{f,t}.$$

In particular, for $t = 0$:

$$y_{f,\tau} = e^{i2\pi f\tau}y_{f,0}.$$

Now set $\tau = t$:

$$y_{f,t} = e^{i2\pi ft}y_{f,0}.$$

Thus, when $\xi_{f,t}$ is input to the LTI digital filter, the output is the same sequence multiplied by some constant, $y_{f,0}$, which is independent of time but will depend on f. Let $G(f) = y_{f,0}$. Then

$$L\{\xi_{f,t}\} = \xi_{f,t}G(f).$$

$G(f)$ is called the transfer function or frequency response function of L. We can write

$$G(f) = |G(f)|e^{i\theta(f)}$$

where,

$$\begin{aligned} &|G(f)| && \text{gain} \\ &\theta(f) = \arg\{G(f)\} && \text{phase} \end{aligned}$$

Any LTI digital filter can be expressed in the form:

$$L\{X_t\} = \sum_{u=-\infty}^{\infty} g_u X_{t-u} \equiv Y_t,$$

where $\{g_u\}$ is a real-valued deterministic sequence called the impulse response sequence. Note,

$$L\{e^{i2\pi ft}\} = \sum_{u=-\infty}^{\infty} g_u e^{i2\pi f(t-u)} = e^{i2\pi ft}G(f),$$

with

$$G(f) = \sum_{u=-\infty}^{\infty} g_u e^{-i2\pi fu} \quad \text{for} \quad |f| \leq \frac{1}{2},$$

where

$$\{g_u\} \longleftrightarrow G(f) \quad \text{(FT pair)}.$$

Now

$$Y_t = \sum_u g_u X_{t-u}$$

and we recall from the spectral representation theorem that

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi ft}\, dZ_X(f) \qquad Y_t = \int_{-1/2}^{1/2} e^{i2\pi ft}\, dZ_Y(f),$$

$$\begin{aligned} \Rightarrow \int e^{i2\pi ft}\, dZ_Y(f) &= \sum_u g_u \int_{-1/2}^{1/2} e^{i2\pi f(t-u)}\, dZ_X(f) \\ &= \int_{-1/2}^{1/2} e^{i2\pi ft} G(f)\, dZ_X(f) \end{aligned}$$

so that, by the 1:1 property of the FT, $dZ_Y(f) = G(f)\, dZ_X(f)$, and

$$\mathrm{E}\{|dZ_Y(f)|^2\} = |G(f)|^2 \mathrm{E}\{|dZ_X(f)|^2\},$$

and if the spectral densities exist

$$S_Y(f) = |G(f)|^2 S_X(f).$$

Diagram: low-pass band filter

This relationship can also be used to determine the sdf's of discrete parameter stationary processes.

3.3.2 Determination of sdf's by LTI digital filtering

MA processes

q-th order moving average: MA(q),

$$X_t = \epsilon_t - \theta_{1,q}\epsilon_{t-1} - \ldots - \theta_{q,q}\epsilon_{t-q},$$

with usual assumptions (mean zero). Define

$$L\{\epsilon_t\} = \epsilon_t - \theta_{1,q}\epsilon_{t-1} - \ldots - \theta_{q,q}\epsilon_{t-q},$$

so that $X_t = L\{\epsilon_t\}$. To determine $G(f)$, input $e^{i2\pi ft}$:

$$\begin{aligned} L\left\{e^{i2\pi ft}\right\} &= e^{i2\pi ft} - \theta_{1,q}e^{i2\pi f(t-1)} - \ldots \theta_{q,q}e^{i2\pi f(t-q)} \\ &= e^{i2\pi ft}\left[1 - \theta_{1,q}e^{-i2\pi f} - \ldots - \theta_{q,q}e^{-i2\pi fq}\right], \end{aligned}$$

so that

$$G_\theta(f) = 1 - \theta_{1,q}e^{-i2\pi f} - \ldots - \theta_{q,q}e^{-i2\pi fq}.$$

Since,

$$S_X(f) = |G_\theta(f)|^2 S_\epsilon(f) \quad \text{and} \quad S_\epsilon(f) = \sigma_\epsilon^2,$$

we have

$$S_X(f) = \sigma_\epsilon^2 |1 - \theta_{1,q}e^{-i2\pi f} - \ldots - \theta_{q,q}e^{-i2\pi fq}|^2.$$

Example sdfs for MA(1) processes can be found in Figure 22.

Worked example: MA(1) spectrum Derive the spectral density function of the MA(1) process $X_t = \epsilon_t - \theta\epsilon_{t-1}$, and from it compute s_1.

AR processes

p-th order autoregressive process: AR(p),

$$X_t - \phi_{1,p}X_{t-1} - \ldots - \phi_{p,p}X_{t-p} = \epsilon_t$$

Define

$$L\{X_t\} = X_t - \phi_{1,p}X_{t-1} - \ldots - \phi_{p,p}X_{t-p},$$

so that $L\{X_t\} = \epsilon_t$. By analogy to MA(q)

$$G_\phi(f) = 1 - \phi_{1,p}e^{-i2\pi f} - \ldots - \phi_{p,p}e^{-i2\pi fp}.$$

Since,

$$|G_\phi(f)|^2 S_X(f) = S_\epsilon(f) \quad \text{and} \quad S_\epsilon(f) = \sigma_\epsilon^2,$$

we have

$$S_X(f) = \frac{\sigma_\epsilon^2}{|1 - \phi_{1,p}e^{-i2\pi f} - \ldots - \phi_{p,p}e^{-i2\pi fp}|^2}.$$

Worked example: AR(1) spectrum

Derive the spectrum of the AR(1) process $X_t = \phi X_{t-1} + \epsilon_t$, ($|\phi| < 1$).

Example sdfs for AR(1) processes are given in Figure 23.

Interpretation of AR spectra Video 25

Recall that for an AR process we have characteristic equation

$$1 - \phi_{1,p}z - \phi_{2,p}z^2 - \ldots - \phi_{p,p}z^p$$

and the process is stationary if the roots of this equation lie outside the unit circle.

Consider an AR(2) process with complex characteristic roots, the roots forming a complex conjugate pair:

$$z_1 = \frac{1}{r}e^{-i2\pi f'}, \qquad z_2 = \frac{1}{r}e^{i2\pi f'}.$$

Now

$$1 - \phi_{1,2}z - \phi_{2,2}z^2 \;=\; (1-az)(1-bz) = 1 - (a+b)z + abz^2$$

so the roots are $z_1 = 1/a$ and $z_2 = 1/b$ and $\phi_{1,2} = (a+b), \phi_{2,2} = -ab$. Then $a = re^{i2\pi f'}$ and $b = re^{-i2\pi f'}$ and $\phi_{1,2} = 2r\cos(2\pi f')$ and $\phi_{2,2} = -r^2$. The AR process can be written

$$X_t - 2r\cos(2\pi f')X_{t-1} + r^2 X_{t-2} = \epsilon_t.$$

The spectrum can be written in terms of the complex roots, by substituting $z = e^{-i2\pi f}$ in the characteristic equation.

$$\begin{aligned} S_X(f) &= \left.\frac{\sigma_\epsilon^2}{|1-az|^2|1-bz|^2}\right|_{z=e^{-i2\pi f}} \\ &= \frac{\sigma_\epsilon^2}{|1-re^{i2\pi f'}e^{-i2\pi f}|^2|1-re^{-i2\pi f'}e^{-i2\pi f}|^2} \end{aligned}$$

Now,

$$|1-re^{i2\pi f'}e^{-i2\pi f}|^2 \;=\; 1 - 2r\cos(2\pi(f'-f)) + r^2.$$

Similarly,

$$|1-re^{-i2\pi f'}e^{-i2\pi f}|^2 = 1 - 2r\cos(2\pi(f'+f)) + r^2.$$

So,

$$S_X(f) = \frac{\sigma_\epsilon^2}{(1-2r\cos(2\pi(f'+f)) + r^2)(1-2r\cos(2\pi(f'-f) + r^2)}.$$

The spectrum will be at its largest when the denominator is at its smallest - when r is close to 1 this occurs when $f \approx \pm f'$. Also notice that at $f = \pm f'$ as $r \to 1$ (from below as $0 < r < 1$ since the root is outside the unit circle) so the spectrum becomes larger.

Generally speaking complex roots will induce a peak in the spectrum, indicating a tendency towards a cycle at frequency f'. Also, the larger the value of r the more dominant the cycle. This may be termed *pseudo-cyclical* behaviour (recall that a deterministic cycle will show up at a sharp spike – i.e., a line spectrum).

Example AR(2) spectra for real and complex-valued roots are given in Figures 24 and 25, respectvely.

ARMA processes

$(p,q)-$th order autoregressive, moving average process: ARMA(p,q),

$$X_t - \phi_{1,p}X_{t-1} - \ldots - \phi_{p,p}X_{t-p} = \epsilon_t - \theta_{1,q}\epsilon_{t-1} - \ldots - \theta_{q,q}\epsilon_{t-q}$$

If we write this as

$$X_t - \phi_{1,p}X_{t-1} - \ldots - \phi_{p,p}X_{t-p} = Y_t;$$

$$Y_t = \epsilon_t - \theta_{1,q}\epsilon_{t-1} - \ldots - \theta_{q,q}\epsilon_{t-q},$$

then we have

$$|G_\phi(f)|^2 S_X(f) = S_Y(f),$$

and

$$S_Y(f) = |G_\theta(f)|^2 S_\epsilon(f),$$

so that

$$\begin{aligned} S_X(f) &= S_\epsilon(f)\frac{|G_\theta(f)|^2}{|G_\phi(f)|^2} \\ &= \sigma_\epsilon^2 \frac{|1 - \theta_{1,q}e^{-i2\pi f} - \ldots - \theta_{q,q}e^{-i2\pi fq}|^2}{|1 - \phi_{1,p}e^{-i2\pi f} - \ldots - \phi_{p,p}e^{-i2\pi fp}|^2} \end{aligned}$$

Differencing

Let $\{X_t\}$ be a stationary process with sdf $S_X(f)$. Let $Y_t = X_t - X_{t-1}$. Then

$$\begin{aligned} L\left\{e^{i2\pi ft}\right\} &= e^{i2\pi ft} - e^{i2\pi f(t-1)} \\ &= e^{i2\pi ft}(1 - e^{-i2\pi f}) \\ &= e^{i2\pi ft}G(f), \end{aligned}$$

so

$$\begin{aligned} |G(f)|^2 &= |1 - e^{-i2\pi f}|^2 = |e^{-i\pi f}(e^{i\pi f} - e^{-i\pi f}|^2 \\ &= |e^{-i\pi f}2i\sin(\pi f)|^2 = 4\sin^2(\pi f). \end{aligned}$$

Chapter 4

Video 27

Estimation

4.1 Estimation of mean and autocovariance function

Ergodic property

Methods we shall look at for estimating quantities such as the autocovariance function will use observations from a single realization. Such methods are based on the strategy of replacing ensemble averages by their corresponding time averages.

Diagram: ergodic property

Reminder: bias, variance and mean square error

4.1.1 Sample mean

Given a stationary time series $X_1, X_2, \ldots, X_N$. Let

$$\bar{X} = \frac{1}{N} \sum X_t.$$

Then,

$$\mathrm{E}\{\bar{X}\} = \frac{1}{N} \sum_{t=1}^{n} \mathrm{E}\{X_t\} = \frac{1}{N} N\mu = \mu$$

so $\bar{X}$ is an unbiased estimator of μ. Hence, $\bar{X}$ converges to μ in mean square if

$$\lim_{N \to \infty} \mathrm{var}\{\bar{X}\} = 0.$$

Now,

$$\begin{aligned}
\mathrm{var}\{\bar{X}\} &= \mathrm{E}\{(\bar{X} - \mu)^2\} \\
&= \mathrm{E}\left\{ \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \mu) \right)^2 \right\} \\
&= \frac{1}{N^2} \sum_{t=1}^{N} \sum_{u=1}^{N} \mathrm{E}\{(X_t - \mu)(X_u - \mu)\} \\
&= \frac{1}{N^2} \sum_{t=1}^{N} \sum_{u=1}^{N} s_{u-t} \\
&= \frac{1}{N^2} \sum_{\tau=-(N-1)}^{N-1} \sum_{k=1}^{N-|\tau|} s_\tau \\
&= \frac{1}{N^2} \sum_{\tau=-(N-1)}^{N-1} (N - |\tau|) s_\tau \\
&= \frac{1}{N} \sum_{\tau=-(N-1)}^{N-1} \left(1 - \frac{|\tau|}{N} \right) s_\tau.
\end{aligned}$$

The summation interchange merely swaps row sums for diagonal sums.

To make further progress we need the condition $\sum_{\tau=-\infty}^{\infty} |s_\tau| < \infty$. By the Cesàro summability theorem, if $\sum_{\tau=-(N-1)}^{N-1} s_\tau$ converges to a limit as $N \to \infty$,

$$\left[\text{it must since} \quad \left| \sum_{\tau=-(N-1)}^{N-1} s_\tau \right| \leq \sum_{\tau=-(N-1)}^{N-1} |s_\tau| < \infty \ \ \forall N \right]$$

then $\sum_{\tau=-(N-1)}^{N-1} \left(1 - \frac{|\tau|}{N}\right) s_\tau$ converges to the same limit.

We can thus conclude that,

$$\lim_{N \to \infty} N \mathrm{var}\{\bar{X}\} = \lim_{N \to \infty} \sum_{\tau=-(N-1)}^{N-1} \left(1 - \frac{|\tau|}{N} \right) s_\tau$$

$$= \lim_{N\to\infty} \sum_{\tau=-(N-1)}^{N-1} s_\tau = \sum_{\tau=-\infty}^{\infty} s_\tau.$$

The assumption of absolute summability of $\{s_\tau\}$ also implies that $\{X_t\}$ has a purely continuous spectrum with sdf

$$S(f) = \sum_{\tau=-\infty}^{\infty} s_\tau e^{-i2\pi f\tau},$$

so that

$$S(0) = \sum_{\tau=-\infty}^{\infty} s_\tau.$$

Thus

$$\lim_{N\to\infty} N\text{var}\{\bar{X}\} = S(0),$$

i.e.,

$$\text{var}\{\bar{X}\} \approx \frac{S(0)}{N} \quad \text{for large } N,$$

and therefore, $\text{var}\{\bar{X}\} \to 0$. Note (i) that the convergence of $\bar{X}$ depends only on the spectrum at $S(0)$, i.e. at $f = 0$, and (ii) $\bar{X}$ is a consistent estimator for μ.

Brief aside: long memory

4.1.2 Autocovariance Sequence

Now,

$$s_\tau = \mathrm{E}\{(X_t - \mu)(X_{t+\tau} - \mu)\}$$

so that a natural estimator for the acvs is

$$\hat{s}_\tau^{(u)} = \frac{1}{N - |\tau|} \sum_{t=1}^{N-|\tau|} (X_t - \bar{X})(X_{t+|\tau|} - \bar{X}) \quad \tau = 0, \pm 1, \ldots, \pm(N-1).$$

Note $\hat{s}_{-\tau}^{(u)} = \hat{s}_\tau^{(u)}$ as it should.

Diagram: autocovariance estimator

If we replace $\bar{X}$ by μ:

$$\begin{aligned} \mathrm{E}\{\hat{s}_\tau^{(u)}\} &= \frac{1}{N - |\tau|} \sum_{t=1}^{N-|\tau|} \mathrm{E}\{(X_t - \mu)(X_{t+|\tau|} - \mu)\} \\ &= \frac{1}{N - |\tau|} \sum_{t=1}^{N-|\tau|} s_\tau = s_\tau, \quad \tau = 0, \pm 1, \ldots, \pm(N-1). \end{aligned}$$

Thus, $\hat{s}_\tau^{(u)}$ is an unbiased estimator of s_τ <u>when μ is known</u>. (Hence the (u) – for unbiased). Most texts refer to $\hat{s}_\tau^{(u)}$ as unbiased – however, if μ is estimated by $\bar{X}$, $\hat{s}_\tau^{(u)}$ is typically a biased estimator of s_τ!!!

A second estimator of s_τ is typically preferred:

$$\hat{s}_\tau^{(p)} = \frac{1}{N} \sum_{t=1}^{N-|\tau|} (X_t - \bar{X})(X_{t+|\tau|} - \bar{X}) \quad \tau = 0, \pm 1, \ldots, \pm(N-1).$$

With $\bar{X}$ replaced by μ:

$$\mathrm{E}\{\hat{s}_\tau^{(p)}\} = \frac{1}{N} \sum_{t=1}^{N-|\tau|} s_\tau = \left(1 - \frac{|\tau|}{N}\right) s_\tau,$$

so that $\hat{s}_\tau^{(p)}$ is a biased estimator, and the magnitude of its bias increases as $|\tau|$ increases. Most texts refer to $\hat{s}_\tau^{(p)}$ as biased.

Why should we prefer the "biased" estimator $\hat{s}_\tau^{(p)}$ to the "unbiased" estimator $\hat{s}_\tau^{(u)}$?

[1] For many stationary processes of practical interest

$$\text{mse}\{\hat{s}_\tau^{(p)}\} < \text{mse}\{\hat{s}_\tau^{(u)}\}.$$

[2] If $\{X_t\}$ has a purely continuous spectrum we know that $s_\tau \to 0$ as $|\tau| \to \infty$. It therefore makes sense to choose an estimator that decreases nicely as $|\tau| \to N-1$ (i.e. choose $\hat{s}_\tau^{(p)}$).

[3] We know that the acvs must be positive semidefinite, the sequence $\{\hat{s}_\tau^{(p)}\}$ has this property, whereas the sequence $\{\hat{s}_\tau^{(u)}\}$ may not.

4.2 Spectral estimation

4.2.1 A naïve non-parametric spectral estimator – the periodogram

Suppose the zero mean discrete stationary process $\{X_t\}$ has a purely continuous spectrum with sdf $S(f)$. We have,

$$S(f) = \sum_{\tau=-\infty}^{\infty} s_\tau e^{-i2\pi f\tau} \quad |f| \leq \frac{1}{2}.$$

With $\mu = 0$, we can use the biased estimator of s_τ:

$$\hat{s}_\tau^{(p)} = \frac{1}{N} \sum_{t=1}^{N-|\tau|} X_t X_{t+|\tau|}$$

for $|\tau| \leq N-1$, but not for $|\tau| \geq N$. Hence we could replace s_τ by $\hat{s}_\tau^{(p)}$ for $|\tau| \leq N-1$ and assume $s_\tau = 0$ for $|\tau| \geq N$. Then a spectrum estimate could be

$$\begin{aligned}
\hat{S}^{(p)}(f) = \sum_{\tau=-(N-1)}^{(N-1)} \hat{s}_\tau^{(p)} e^{-i2\pi f\tau} &= \frac{1}{N} \sum_{\tau=-(N-1)}^{(N-1)} \sum_{t=1}^{N-|\tau|} X_t X_{t+|\tau|} e^{-i2\pi f\tau} \\
&= \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{N} X_j X_k e^{-i2\pi f(k-j)} \\
&= \frac{1}{N} \left| \sum_{t=1}^{N} X_t e^{-i2\pi ft} \right|^2,
\end{aligned}$$

where the summation interchange has merely swapped diagonal sums for row sums.

Diagram: summation interchange

$\hat{S}^{(p)}(f)$ defined above is known as the periodogram, and is defined over $[-1/2, 1/2]$.

Note that $\{\hat{s}_\tau^{(p)}\}$ and $\hat{S}^{(p)}(f)$ are a FT pair:

$$\{\hat{s}_\tau^{(p)}\} \longleftrightarrow \hat{S}^{(p)}(f)$$

(hence the (p) for periodogram), just like the population quantities

$$\{s_\tau\} \longleftrightarrow S(f).$$

Hence, $\{s_\tau^{(p)}\}$ can be written as

$$s_\tau^{(p)} = \int_{-1/2}^{1/2} \hat{S}^{(p)}(f) e^{i2\pi f\tau}\, df \quad |\tau| \leq N-1.$$

If $\hat{S}^{(p)}(f)$ were an ideal estimator of $S(f)$ we would have

[1] $\mathrm{E}\{\hat{S}^{(p)}(f)\} \approx S(f) \quad \forall f.$

[2] $\mathrm{var}\{\hat{S}^{(p)}(f)\} \to 0$ as $N \to \infty$ and,

[3] $\mathrm{cov}\{\hat{S}^{(p)}(f), \hat{S}^{(p)}(f')\} \approx 0$ for $f \neq f'$.

We find that

[1] is a good approximation for some processes,

[2] is blatantly false (see Figure 26),

[3] holds if f and f' are certain distinct frequencies, namely, the Fourier frequencies $f_k = k/N \quad (\Delta t = 1)$.

We firstly look at the expectation in [1] (assuming $\mu = 0$). Video 31

$$\begin{aligned} \mathrm{E}\{\hat{S}^{(p)}(f)\} &= \sum_{\tau=-(N-1)}^{(N-1)} \mathrm{E}\{s_\tau^{(p)}\} e^{-i2\pi f\tau} \\ &= \sum_{\tau=-(N-1)}^{(N-1)} \left(1 - \frac{|\tau|}{N}\right) s_\tau e^{-i2\pi f\tau}. \end{aligned}$$

Hence, if we know the acvs $\{s_\tau\}$ we can work out from this what $\mathrm{E}\{\hat{S}^{(p)}(f)\}$ will be. We can obtain much more insight by considering:

$$\mathrm{E}\{|J(f)|^2\} \quad \text{where} \quad J(f) = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} X_t e^{-i2\pi ft}, \quad |f| \leq \frac{1}{2}.$$

We know from the spectral representation theorem that,

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi f' t}\, dZ(f'),$$

so that,

$$\begin{aligned} J(f) &= \sum_{t=1}^{N} \left(\int_{-1/2}^{1/2} \frac{1}{\sqrt{N}} e^{i2\pi f' t}\, dZ(f') \right) e^{-i2\pi f t} \\ &= \int_{-1/2}^{1/2} \sum_{t=1}^{N} \frac{1}{\sqrt{N}} e^{-i2\pi (f-f')t}\, dZ(f') \end{aligned}$$

Then

$$\begin{aligned} \mathrm{E}\{\hat{S}^{(p)}(f)\} &= \mathrm{E}\{|J(f)|^2\} = \mathrm{E}\{J^*(f)J(f)\} \\ &= \mathrm{E}\left\{ \int_{-1/2}^{1/2} \sum_{t=1}^{N} \frac{1}{\sqrt{N}} e^{i2\pi (f-f')t}\, dZ^*(f') \int_{-1/2}^{1/2} \sum_{s=1}^{N} \frac{1}{\sqrt{N}} e^{-i2\pi (f-f'')s}\, dZ(f'') \right\} \\ &= \int_{-1/2}^{1/2} \int_{-1/2}^{1/2} \sum_{t=1}^{N} \frac{1}{\sqrt{N}} e^{i2\pi (f-f')t} \sum_{s=1}^{N} \frac{1}{\sqrt{N}} e^{-i2\pi (f-f'')s} \mathrm{E}\{dZ^*(f')\, dZ(f'')\} \\ &= \int_{-1/2}^{1/2} \mathcal{F}(f-f') S(f')\, df', \end{aligned}$$

where $\mathcal{F}$ is Féjer's kernel defined by

$$\mathcal{F}(f) = \left| \sum_{t=1}^{N} \frac{1}{\sqrt{N}} e^{-i2\pi f t} \right|^2 = \frac{\sin^2(N\pi f)}{N \sin^2(\pi f)}.$$

This result tells us that the expected value of $\hat{S}^{(p)}(f)$ is the true spectrum convolved with Féjer's kernel.

Diagram: effects of convolution

To understand the implications of this we need to know the properties of Féjer's kernel:

[1] For all integers $N \geq 1, \mathcal{F}(f) \to N$ as $f \to 0$.

[2] For $N \geq 1$, $f \in [-1/2, 1/2]$ and $f \neq 0$, $\mathcal{F}(f) < \mathcal{F}(0)$.

[3] For $f \in [-1/2, 1/2]$, $f \neq 0$, $\mathcal{F}(f) \to 0$ as $N \to \infty$.

[4] For any integer $k \neq 0$ such that $f_k = k/N \in [-1/2, 1/2]$, $\mathcal{F}(f_k) = 0$.

[5] $\int_{-1/2}^{1/2} \mathcal{F}(f)\, df = 1$.

Figure 27 shows Féjer's kernel on a $10 \log_{10}$ scale (dBs) for $N = 8, 32$ and 128. $\mathcal{F}(f)$ is symmetric about the origin and consists of a broad central peak ("lobe") and $N - 2$ sidelobes which decrease as $|f|$ increases. From [1], [3] and [5] it follows that as $N \to \infty$, $\mathcal{F}(f)$ acts as a Dirac δ function with an infinite spike at $f = 0$.

So

$$\lim_{N \to \infty} \mathrm{E}\{\hat{S}^{(p)}(f)\} = \int_{-1/2}^{1/2} \delta(f - f') S(f')\, df' = S(f),$$

i.e., $\hat{S}^{(p)}(f)$ is *asymptotically* unbiased as an estimator of $S(f)$.

For a process with large dynamic range, defined as

$$10 \log_{10} \left(\frac{\max_f S(f)}{\min_f S(f)} \right),$$

since the expected value of the periodogram is a convolution of Féjer's kernel and the true spectrum, power from parts of the spectrum where $S(f)$ is large can "leak" via the sidelobes to other frequencies where $S(f)$ is small.

Diagram: side-lobe leakage

Figures 28 and 29 demonstrate this "sidelobe leakage" for two processes, the first with a dynamic range of 14 dB, the second with a dynamic range of 65 dB.

4.2.2 Bias reduction – Tapering

Much of the bias in the periodogram can be attributed to sidelobe leakage due to the presence of Féjer's kernel. Tapering is a technique which reduces the sidelobes associated with Féjer's kernel.

Let $X_1, X_2, \ldots, X_N$ be a portion of length N of a zero mean stationary process with sdf $S(f)$. We form the product $\{h_t X_t\}$ where $\{h_t\}$ is a sequence of real-valued constants called a data taper normalized so that

$$\sum_{t=1}^{N} h_t^2 = 1.$$

Diagram: data taper

Define

$$J(f) = \sum_{t=1}^{N} h_t X_t e^{-i2\pi ft} \qquad |f| \leq 1/2.$$

By the spectral representation theorem,

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi f't}\, dZ(f'),$$

so that,

$$\begin{aligned} J(f) &= \sum_{t=1}^{N} h_t \left(\int_{-1/2}^{1/2} e^{i2\pi f't}\, dZ(f') \right) e^{-i2\pi ft} \\ &= \int_{-1/2}^{1/2} \sum_{t=1}^{N} h_t e^{-i2\pi(f-f')t}\, dZ(f') \\ &= \int_{-1/2}^{1/2} H(f-f')\, dZ(f'), \end{aligned}$$

where,

$$H(f) = \sum_{t=1}^{N} h_t e^{-i2\pi ft} \quad \text{i.e.,} \quad \{h_t\} \longleftrightarrow H(f).$$

Let,

$$\hat{S}^{(d)}(f) = |J(f)|^2 = \left| \sum_{t=1}^{N} h_t X_t e^{-i2\pi ft} \right|^2.$$

Then,

$$|J(f)|^2 = J^*(f)J(f) = \int_{-1/2}^{1/2} H^*(f-f')\,dZ^*(f') \int_{-1/2}^{1/2} H(f-f'')\,dZ(f''),$$

and hence,

$$\begin{aligned} \mathrm{E}\{\hat{S}^{(d)}(f)\} &= \mathrm{E}\{|J(f)|^2\} \\ &= \int_{-1/2}^{1/2} |H(f-f')|^2 S(f')\,df' \\ &= \int_{-1/2}^{1/2} \mathcal{H}(f-f')S(f')\,df', \end{aligned}$$

where $\mathcal{H}(f) = |H(f)|^2$, i.e.,

$$\mathcal{H}(f) = \left| \sum_{t=1}^{N} h_t e^{-i2\pi ft} \right|^2.$$

A spectral estimator of the form of $\hat{S}^{(d)}(f)$ is called a direct spectral estimator (hence the (d)). Note, if $h_t = \frac{1}{\sqrt{N}}$ for $1 \le t \le N$, then

$$\hat{S}^{(d)}(f) = \hat{S}^{(p)}(f) \quad \text{and} \quad \mathcal{H}(f) = \mathcal{F}(f),$$

i.e., $\hat{S}^{(d)}(f)$ is the same as the periodogram,

$\mathcal{H}(f)$ is the same as Fejer's kernel.

The key idea behind tapering is to select $\{h_t\}$ so that $\mathcal{H}(f)$ has much lower sidelobes that $\mathcal{F}(f)$. Recall that $\mathcal{F}(f)$ corresponds to a rectangular taper

$$h_t = \begin{cases} \frac{1}{\sqrt{N}} & \text{for } 1 \le t \le N, \\ 0 & \text{otherwise.} \end{cases}$$

There is thus a sharp discontinuity between where the taper is "ON" $(1 \le t \le N)$ and where it is "OFF" $(N < t < 0)$. Tapering effectively creates a smooth transition at the ends of the data.

Diagram: aims of tapering

Figure 30 shows the effect of tapering on the shape of the spectral window $\mathcal{H}(f)$. The $p \times 100\%$ cosine taper is defined by

$$h_t = \begin{cases} \frac{C}{2}\left[1 - \cos\left(\frac{2\pi t}{\lfloor pN \rfloor + 1}\right)\right], & 1 \leq t \leq \frac{\lfloor pN \rfloor}{2}; \\ C, & \frac{\lfloor pN \rfloor}{2} < t < N + 1 - \frac{\lfloor pN \rfloor}{2}; \\ \frac{C}{2}\left[1 - \cos\left(\frac{2\pi(N+1-t)}{\lfloor pN \rfloor + 1}\right)\right], & N + 1 - \frac{\lfloor pN \rfloor}{2} \leq t \leq N, \end{cases}$$

where C is a normalizing constant that forces $\sum_{t=1}^{N} h_t^2 = 1$.

As we perform more tapering, the main lobe of $\mathcal{H}(f)$ gets wider, but the sidelobes get lower. This means that the more tapering we perform:

Resolution of the spectrum DECREASES (bad!)
Sidelobe leakage DECREASES (good!).

Figure 31 demonstrates how the modification of the spectral window inherent in tapering reduces the sidelobe leakage at the expense of widening the main lobe (this results in smoothing bias) for the AR(4) process with high dynamic range.

Chapter 5

Video 33

Parametric model fitting (for autoregressive processes)

Here we concentrate on zero-mean models of the form

$$X_t - \phi_{1,p} X_{t-1} - \ldots - \phi_{p,p} X_{t-p} = \epsilon_t.$$

As we have seen the corresponding sdf is

$$S(f) = \frac{\sigma_\epsilon^2}{|1 - \phi_{1,p} e^{-i2\pi f} - \ldots - \phi_{p,p} e^{-i2\pi fp}|^2}.$$

This class of models is appealing to use for time series analysis for several reasons:

[1] Any time series with a purely continuous sdf can be approximated well by an AR(p) model if p is large enough.

[2] There exist efficient algorithms for fitting AR(p) models to time series.

[3] Quite a few physical phenomena are reverberant and hence an AR model is naturally appropriate.

5.1 Yule-Walker method

Video 34

We start by multiplying the defining equation by X_{t-k}:

$$X_t X_{t-k} = \sum_{j=1}^{p} \phi_{j,p} X_{t-j} X_{t-k} + \epsilon_t X_{t-k}.$$

Taking expectations, for $k > 0$:

$$s_k = \sum_{j=1}^{p} \phi_{j,p} s_{k-j}.$$

Let $k = 1, 2, \ldots, p$ and recall that $s_{-\tau} = s_\tau$ to obtain

$$\begin{aligned} s_1 &= \phi_{1,p} s_0 + \phi_{2,p} s_1 + \ldots + \phi_{p,p} s_{p-1} \\ s_2 &= \phi_{1,p} s_1 + \phi_{2,p} s_0 + \ldots + \phi_{p,p} s_{p-2} \\ \vdots &\quad \vdots \\ s_p &= \phi_{1,p} s_{p-1} + \phi_{2,p} s_{p-2} + \ldots + \phi_{p,p} s_0 \end{aligned}$$

or in matrix notation,

$$\boldsymbol{\gamma}_p = \Gamma_p \boldsymbol{\phi}_p,$$

where $\boldsymbol{\gamma}_p = [s_1, s_2, \ldots, s_p]^T$; $\boldsymbol{\phi}_p = [\phi_{1,p}, \phi_{2,p}, \ldots, \phi_{p,p}]^T$ and

$$\Gamma_p = \begin{bmatrix} s_0 & s_1 & \ldots & s_{p-1} \\ s_1 & s_0 & \ldots & s_{p-2} \\ \vdots & \vdots & & \vdots \\ s_{p-1} & s_{p-2} & \ldots & s_0 \end{bmatrix}$$

Note: this is a symmetric Toeplitz matrix which we have met already. All elements on a diagonal are the same.

Suppose we don't know the $\{s_\tau\}$, but the mean is zero, then take

$$\hat{s}_\tau = \frac{1}{N} \sum_{t=1}^{N-|\tau|} X_t X_{t+|\tau|},$$

and substitute these for the s_τ's in $\boldsymbol{\gamma}_p$ and Γ_p to obtain $\hat{\boldsymbol{\gamma}}_p, \hat{\Gamma}_p$, from which we estimate $\boldsymbol{\phi}_p$ as $\hat{\boldsymbol{\phi}}_p$:

$$\hat{\boldsymbol{\phi}}_p = \hat{\Gamma}^{-1} \hat{\boldsymbol{\gamma}}_p.$$

Finally, we need to estimate σ_ϵ^2. To do so, we multiply the defining equation by X_t and take expectations to obtain

$$\begin{aligned} s_0 &= \sum_{j=1}^{p} \phi_{j,p} s_j + \mathrm{E}\{\epsilon_t X_t\} \\ &= \sum_{j=1}^{p} \phi_{j,p} s_j + \sigma_\epsilon^2, \end{aligned}$$

so that as an estimator for σ_ϵ^2 we take

$$\hat{\sigma}_\epsilon^2 = \hat{s}_o - \sum_{j=1}^{p} \hat{\phi}_{j,p} \hat{s}_j.$$

The estimators $\hat{\boldsymbol{\phi}}_p$ and $\hat{\sigma}_\epsilon^2$ are called the Yule-Walker estimators of the AR(p) parameters.

The estimate of the sdf resulting is

$$\hat{S}(f) = \frac{\hat{\sigma}_\epsilon^2}{\left|1 - \sum_{j=1}^{p} \hat{\phi}_{j,p} e^{-i2\pi f j}\right|^2}.$$

There are two important modifications which we can make to this approach:

[1] We could use for $\{\hat{s}_\tau\}$ a modified autocovariance incorporating tapering:

$$\hat{s}_\tau = \sum_{t=1}^{N-|\tau|} h_t X_t h_{t+|\tau|} X_{t+|\tau|}.$$

[2] To invert $\hat{\Gamma}_p$ by brute force matrix inversion requires $\mathrm{O}(p^3)$ operations. Fortunately, there is an algorithm due to Levinson and Durbin which takes advantage of the highly structured nature of the Toeplitz matrix, and carries out the estimation in $\mathrm{O}(p^2)$ or fewer operations.

Worked example: Yule-Walker

Examples: The AR(4) process again.

- Figure 32: Shows simulations from the AR(4) process defined by,

$$X_t = 2.7607X_{t-1} - 3.8106X_{t-2} + 2.6535X_{t-3} - 0.9258X_{t-3} + \epsilon_t$$

- Figure 33: Shows AR(4) processes fitted to the AR(4) data using Yule-Walker method and

$$\hat{s}_\tau = \frac{1}{N} \sum_{t=1}^{N-|\tau|} X_t X_{t+|\tau|}.$$

  Very poor, even for $N = 1024$.

- Figure 34: Shows AR(8) processes fitted to the AR(4) data using Yule-Walker method and

$$\hat{s}_\tau = \frac{1}{N} \sum_{t=1}^{N-|\tau|} X_t X_{t+|\tau|}.$$

  Although the process fitted is not the correct one, the extra parameters have improved the fit.

- Figure 35: Shows AR(4) process fitted to the AR(4) data, using Yule-Walker, but with the 50% split cosine bell taper used:

$$\hat{s}_\tau = \sum_{t=1}^{N-|\tau|} h_t X_t h_{t+|\tau|} X_{t+|\tau|}.$$

  The improvement over the other Yule-Walker estimates is dramatic.

The parameter estimates for the fitted AR(4) models when N=1024 are:

	true	Yule-Walker	tapered Y-W
$\phi_{1,4}$	2.7607	1.8459	2.7636
$\phi_{2,4}$	-3.8106	-1.7138	-3.8108
$\phi_{3,4}$	2.6535	0.6200	2.6502
$\phi_{4,4}$	-0.9258	-0.1437	-0.9211
σ_ϵ^2	1.0	14.9758	1.0841

5.2 Least squares

Video 35

Let $\{X_t\}$ be a zero-mean AR(p) process, i.e.,

$$X_t - \phi_{1,p}X_{t-1} - \phi_{2,p}X_{t-2} + \ldots - \phi_{p,p}X_{t-p} = \epsilon_t.$$

Diagram: Least squares formulation

We can formulate an appropriate least squares model in terms of data $X_1, X_2, \ldots, X_N$ as follows:

$$\boldsymbol{X} = F\boldsymbol{\phi} + \boldsymbol{\epsilon},$$

where,

$$F = \begin{bmatrix} X_p & X_{p-1} & \ldots & X_1 \\ X_{p+1} & X_p & \ldots & X_2 \\ \vdots & & & \vdots \\ X_{N-1} & X_{N-2} & \ldots & X_{N-p} \end{bmatrix}$$

and,

$$\boldsymbol{X} = \begin{bmatrix} X_{p+1} \\ X_{p+2} \\ \vdots \\ X_N \end{bmatrix}; \quad \boldsymbol{\phi} = \begin{bmatrix} \phi_{1,p} \\ \phi_{2,p} \\ \vdots \\ \phi_{p,p} \end{bmatrix}; \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{p+1} \\ \epsilon_{p+2} \\ \vdots \\ \epsilon_N \end{bmatrix}.$$

We can thus estimate $\boldsymbol{\phi}$ by finding that $\boldsymbol{\phi}$ such that

$$\begin{aligned} \mathrm{SS}(\boldsymbol{\phi}) &= \sum_{t=p+1}^{N} \left(X_t - \sum_{k=1}^{p} \phi_{k,p} X_{t-k} \right)^2 \quad \left[= \sum_{t=p+1}^{N} \epsilon_t^2 \right] \\ &= (\boldsymbol{X} - F\boldsymbol{\phi})^T (\boldsymbol{X} - F\boldsymbol{\phi}) \end{aligned}$$

is minimized. If we denote the vector that minimizes the above as $\hat{\boldsymbol{\phi}}$, standard least squares theory tells us that it is given by

$$\hat{\boldsymbol{\phi}} = (F^T F)^{-1} F^T \boldsymbol{X}.$$

We can estimate the innovations variance σ_ϵ^2 by the usual estimator of the residual variation, namely

$$\hat{\sigma}^2 = \frac{(\boldsymbol{X} - F\hat{\boldsymbol{\phi}})^T (\boldsymbol{X} - F\hat{\boldsymbol{\phi}})}{(N - 2p)}.$$

(Note: there are $N - p$ effective observations, and p parameters are estimated). The estimator $\hat{\boldsymbol{\phi}}$ is known as the forward least squares estimator of $\boldsymbol{\phi}$.

Figure 36 shows the AR(4) spectra corresponding to the least squares estimates of $\boldsymbol{\phi}$ and tapered Yule-Walker estimates for comparison.

5.3 Maximum likelihood

Video 36

For a portion $X_1, ..., X_N$ from an AR(p) process, the likelihood function for the parameters $\boldsymbol{\phi}$ and σ_ϵ^2 is

$$L(\boldsymbol{\phi}, \sigma_\epsilon^2) = f(X_1, ..., X_N | \boldsymbol{\phi}, \sigma_\epsilon^2)$$

where f is the joint density function of $X_1, ..., X_N$ for an AR(p) process. The maximum likelihood estimates are the values of $\boldsymbol{\phi}, \sigma_\epsilon^2$ that maximise L, namely

$$\hat{\boldsymbol{\phi}}, \hat{\sigma}_\epsilon^2 = \text{argmax } L(\boldsymbol{\phi}, \sigma_\epsilon^2).$$

Diagram: maximum likelihood

To define f, we need to make an assumption on the distribution of the data. Here, we will assume this is a Gaussian process.

We can write

$$\begin{aligned} f(X_1,...,X_N|\boldsymbol{\phi},\sigma_\epsilon^2) &= f(X_1,...,X_{N-1}|\boldsymbol{\phi},\sigma_\epsilon^2)f(X_N|X_1,...,X_{N-1},\boldsymbol{\phi},\sigma_\epsilon^2) \\ &= f(X_1,...,X_{N-1}|\boldsymbol{\phi},\sigma_\epsilon^2)f(X_N|X_{N-1},...,X_{N-p},\boldsymbol{\phi},\sigma_\epsilon^2). \end{aligned}$$

Applying the same argument gives

$$\begin{aligned} f(X_1,...,X_{N-1}|\boldsymbol{\phi},\sigma_\epsilon^2) &= f(X_1,...,X_{N-2}|\boldsymbol{\phi},\sigma_\epsilon^2)f(X_{N-1}|X_1,...,X_{N-2},\boldsymbol{\phi},\sigma_\epsilon^2) \\ &= f(X_1,...,X_{N-2}|\boldsymbol{\phi},\sigma_\epsilon^2)f(X_{N-1}|X_{N-2},...,X_{N-p-1},\boldsymbol{\phi},\sigma_\epsilon^2). \end{aligned}$$

Iterating gives

$$f(X_1,...,X_N|\boldsymbol{\phi},\sigma_\epsilon^2) = f(X_1,...,X_p)\prod_{t=p+1}^{N} f(X_t|X_{t-1},...,X_{t-p},\boldsymbol{\phi},\sigma_\epsilon^2).$$

With

$$X_t = \epsilon_t + \sum_{j=1}^{p}\phi_{j,p}X_{t-j},$$

under the Gaussian assumption we have $X_t|X_{t-1},...,X_{t-p} \sim N\left(\sum_{j=1}^{p}\phi_{j,p}X_{t-j},\sigma_\epsilon^2\right)$, and therefore

$$f(X_t|X_{t-1},...,X_{t-p},\boldsymbol{\phi},\sigma_\epsilon^2) = \frac{1}{\sqrt{2\pi}\sigma_\epsilon^2}\exp(-\frac{1}{2\sigma_\epsilon^2}(X_t - \sum_{j=1}^{p}\phi_{j,p}X_{t-j})^2).$$

Method 1: it is possible to obtain a closed form approximate solution by considering the $X_1,...,X_p$ to be deterministic. Then

$$L(\boldsymbol{\phi},\sigma_\epsilon^2) \propto \prod_{t=p+1}^{N} f(X_t|X_{t-1},...,X_{t-p},\boldsymbol{\phi},\sigma_\epsilon^2).$$

Maximising $L(\boldsymbol{\phi},\sigma_\epsilon^2)$ is equivalent to maximising the log-likelihood $\ell(\boldsymbol{\phi},\sigma_\epsilon^2) = \ln(L(\boldsymbol{\phi},\sigma_\epsilon^2))$. We therefore have

$$\begin{aligned} \ell(\boldsymbol{\phi},\sigma_\epsilon^2) &= C + \sum_{t=p+1}^{N}\ln f(X_t|X_{t-1},...,X_{t-p},\boldsymbol{\phi},\sigma_\epsilon^2) \\ &= C' - (N-p)\ln(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2}\sum_{t=p+1}^{N}(X_t - \sum_{j=1}^{p}\phi_{j,p}X_{t-j})^2 \\ &= C' - (N-p)\ln(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2}SS(\boldsymbol{\phi}) \end{aligned}$$

We notice that the $\boldsymbol{\phi}$ that maximises this expression is the Least Squares estimator, namely

$$\hat{\boldsymbol{\phi}} = (F^T F)^{-1} F^T \boldsymbol{X}.$$

This gives

$$\ell(\hat{\boldsymbol{\phi}}, \sigma_\epsilon^2) = C' - (N-p)\ln(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2}(\boldsymbol{X} - F\hat{\boldsymbol{\phi}})^T(\boldsymbol{X} - F\hat{\boldsymbol{\phi}}).$$

To find the maximum likelihood estimator of σ_ϵ^2, we now differentiate this wrt σ_ϵ^2 and set to zero. It is straight forward to show

$$\hat{\sigma}_\epsilon^2 = \frac{(\boldsymbol{X} - F\hat{\boldsymbol{\phi}})^T(\boldsymbol{X} - F\hat{\boldsymbol{\phi}})}{(N-p)}.$$

However, this is a biased estimator and therefore the unbiased estimator presented for Least Squares estimation is often preferred.

Method 2: we consider maximising the complete likelihood

$$L(\boldsymbol{\phi}, \sigma_\epsilon^2) = f(X_1, ..., X_p) \prod_{t=p+1}^{N} f(X_t | X_{t-1}, ..., X_{t-p}, \boldsymbol{\phi}, \sigma_\epsilon^2).$$

However, this is a complicated function of the model parameters because

$$f(X_1, ..., X_p) = (2\pi)^{-p/2} \det(\Sigma)^{-1/2} \exp(-\frac{1}{2}\boldsymbol{X_p}^T \Sigma^{-1} \boldsymbol{X_p}),$$

where $\boldsymbol{X_p} = (X_1, ..., X_p)^T$ and Σ is the covariance matrix of $\boldsymbol{X_p}$, namely

$$\Sigma = \begin{pmatrix} s_0 & s_1 & \cdots & s_{p-1} \\ s_1 & s_0 & & \\ \vdots & & \ddots & \\ s_{p-1} & & & s_0 \end{pmatrix}.$$

We have seen that s_τ can be a complicated function of the model parameters. There is no closed form solution to this but instead it can be optimised by numerical procedures (e.g. Nelder-Mead simplex).

5.4 Notes

Video 37

[1] Least squares and maximum likelihood methods produce estimated models which need not be stationary. This may be a concern for prediction, however, for spectral estimation, the parameter values will still produce a valid sdf (i.e., nonnegative everywhere, symmetric about the origin and integrates to a finite number).

[2] The Yule-Walker can be considered to be a matching of moments method. Furthermore, estimates can be formulated as a least squares problem. Consider adding zeros to our observations $X_1, X_2, \ldots, X_N$, both at the beginning and end of the data.

This gives

$$\boldsymbol{X}_{YW} = W\boldsymbol{\phi} + \boldsymbol{\epsilon}_{YW},$$

where,

$$W = \begin{bmatrix} 0 & 0 & 0 & \dots & \dots & 0 \\ X_1 & 0 & 0 & \dots & \dots & 0 \\ X_2 & X_1 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & & & & \vdots \\ X_{p-1} & \vdots & & & & 0 \\ X_p & X_{p-1} & \dots & \dots & \dots & X_1 \\ \vdots & \vdots & & & & \vdots \\ X_N & X_{N-1} & \dots & \dots & \dots & X_{N-p+1} \\ 0 & X_N & & & & X_{N-p+2} \\ \vdots & \vdots & & & & \vdots \\ 0 & 0 & & & & X_N \end{bmatrix}$$

and,

$$\boldsymbol{X}_{YW} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon}_{YW} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Note that,

$$\frac{1}{N}W^TW = \begin{bmatrix} \hat{s}_0^{(p)} & \hat{s}_1^{(p)} & \dots & \hat{s}_{p-1}^{(p)} \\ \hat{s}_1^{(p)} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \hat{s}_{p-1}^{(p)} & \dots & \dots & \hat{s}_0^{(p)} \end{bmatrix} = \hat{\Gamma}_p$$

and

$$\frac{1}{N}W^T\boldsymbol{X}_{YW} = \begin{bmatrix} \hat{s}_1^{(p)} \\ \vdots \\ \hat{s}_p^{(p)} \end{bmatrix} = \hat{\boldsymbol{\gamma}}_p,$$

so that

$$(W^TW)^{-1}W^T\boldsymbol{X}_{YW} = (\hat{\Gamma}_p)^{-1}\hat{\boldsymbol{\gamma}}_p.$$

which is identical to the Yule-Walker estimate.

[3] Maximum likelihood method is in some ways equivalent to Least Squares when assuming Gaussianity. Other distributions can be assumed, however, if the distribution is chosen incorrectly it will give poor estimates.

5.5 Model selection

Video 38

The above estimation methods for the parameters of an AR(p) process assume we know p when fitting the model. Often this is not the case so we wish to fit the model with several different values of p and selecting the "best" model.

Typically, this is achieved via the Akaike information criterion (AIC), which is defined as

$$\text{AIC} = 2k - 2\ell(\hat{\boldsymbol{\phi}}, \hat{\sigma}^2_\epsilon).$$

We choose the model which gives *lowest* AIC score. Here, k is the number of free parameters to be estimated when fitting the model, so for an AR(p) process, $k = p+1$ (the p coefficients plus the variance term). The AIC looks to reward goodness of fit but penalise model complexity.

The procedure is as follows

1. Choose a set of values for p that you wish to try, e.g. $p = 1, 2, 3, 4, 5, ..., 10$.

2. For each value of p, estimate the model parameters via maximum likelihood, compute the value of the likelihood function at these estimates, and compute the AIC.

3. Choose the value of p and its associated parameter estimates that gave the lowest AIC.

Video 45

Chapter 6

Forecasting

Video 46

6.1 Formulation

Suppose we wish to predict the value of X_{t+l} of a process, given $X_t, X_{t-1}, X_{t-2}, \ldots$. Let the appropriate model for $\{X_t\}$ be an ARMA(p, q) process:

$$\Phi(B)X_t = \Theta(B)\epsilon_t.$$

Consider a forecast $X_t(l)$ of X_{t+l} (an l-step ahead forecast) which is a linear combination of $X_t, X_{t-1}, X_{t-2}, \ldots$:

$$X_t(l) = \sum_{k=0}^{\infty} \pi_k X_{t-k}.$$

Note: this assumes a semi-infinite realization of $\{X_t\}$. Let us now assume that $\{X_t\}$ can be written as a one-sided linear process, so that

$$X_t = \sum_{k=0}^{\infty} \psi_k \epsilon_{t-k} = \Psi(B)\epsilon_t,$$

and

$$X_{t+l} = \sum_{k=0}^{\infty} \psi_k \epsilon_{t+l-k} = \Psi(B)\epsilon_{t+l}.$$

Hence,

$$\begin{aligned} X_t(l) = \sum_{k=0}^{\infty} \pi_k X_{t-k} &= \sum_{k=0}^{\infty} \pi_k \Psi(B)\epsilon_{t-k} \\ &= \Pi(B)\Psi(B)\epsilon_t. \end{aligned}$$

Let $\delta(B) = \Pi(B)\Psi(B)$ so that,

$$\begin{aligned} X_t(l) &= \delta(B)\epsilon_t \\ &= \sum_{k=0}^{\infty} \delta_k \epsilon_{t-k}. \end{aligned}$$

Now,

$$\begin{aligned} X_{t+l} &= \textstyle\sum_{k=0}^{\infty} \psi_k \epsilon_{t+l-k} \\ &= \textstyle\sum_{k=0}^{l-1} \psi_k \epsilon_{t+l-k} \quad + \sum_{k=l}^{\infty} \psi_k \epsilon_{t+l-k} \\ &\qquad \text{(A)} \qquad\qquad\qquad \text{(B)} \end{aligned}$$

(A) Involves future ϵ_ts, and so represents the "unpredictable" part of X_{t+l}.

(B) Depends only on past and present values of ϵ_t, thus representing the "predictable" part of X_{t+l}.

Diagram: predictable and unpredictable parts

Hence we would expect,

$$\begin{aligned} X_t(l) &= \sum_{k=l}^{\infty} \psi_k \epsilon_{t+l-k} \\ &= \sum_{j=0}^{\infty} \psi_{j+l} \epsilon_{t-j}, \end{aligned}$$

so that $\delta_k \equiv \psi_{k+l}$.

This can be readily proved. For linear least squares, we want to minimize,

$$\begin{aligned}
\mathrm{E}\{(X_{t+l} - X_t(l))^2\} &= \mathrm{E}\left\{\left(\sum_{k=0}^{l-1}\psi_k\epsilon_{t+l-k} + \sum_{k=0}^{\infty}\psi_{k+l}\epsilon_{t-k} - \sum_{k=0}^{\infty}\delta_k\epsilon_{t-k}\right)^2\right\} \\
&= \mathrm{E}\left\{\left(\sum_{k=0}^{l-1}\psi_k\epsilon_{t+l-k} + \sum_{k=0}^{\infty}[\psi_{k+l} - \delta_k]\epsilon_{t-k}\right)^2\right\} \\
&= \sigma_\epsilon^2\left\{\left(\sum_{k=0}^{l-1}\psi_k^2\right) + \sum_{k=0}^{\infty}(\psi_{k+l} - \delta_k)^2\right\}.
\end{aligned}$$

The first term is independent of the choice of the $\{\delta_k\}$ and the second term is clearly minimized by choosing $\delta_k = \psi_{k+l}, k = 0, 1, 2, \ldots$ as expected. With this choice of $\{\delta_k\}$ the second term vanishes, and we have,

$$\begin{aligned}
\sigma^2(l) &= \mathrm{E}\{(X_{t+l} - X_t(l))^2\} \\
&= \sigma_\epsilon^2\sum_{k=0}^{l-1}\psi_k^2,
\end{aligned}$$

which is known as the l-step prediction variance.

When $l = 1$, $\delta_k = \psi_{k+1}$,

$$\begin{aligned}
X_t(1) &= \delta_0\epsilon_t + \delta_1\epsilon_{t-1} + \delta_2\epsilon_{t-2} + \ldots \\
&= \psi_1\epsilon_t + \psi_2\epsilon_{t-1} + \psi_3\epsilon_{t-2} + \ldots \\
X_{t+1} &= \psi_0\epsilon_{t+1} + \psi_1\epsilon_t + \psi_2\epsilon_{t-1} + \ldots
\end{aligned}$$

so that,

$$X_{t+1} - X_t(1) = \psi_0\epsilon_{t+1} = \epsilon_{t+1} \quad \text{since} \quad \psi_0 = 1.$$

Hence ϵ_{t+1} can be thought of as the "one step prediction error". Also of course,

$$X_{t+1} = X_t(1) + \epsilon_{t+1}$$

so that ϵ_{t+1} is the essentially "new" part of X_{t+1} which is not linearly dependent on past observations. The sequence $\{\epsilon_t\}$ is often called the innovations process of $\{X_t\}$, and, as used here, σ_ϵ^2 is called the innovations variance.

If we wish to write $X_t(l)$ explicitly as a function of $X_t, X_{t-1}, \ldots$ then we note first that,

$$X_t(l) = \sum_{k=0}^{\infty} \delta_k \epsilon_{t-k} = \sum_{k=0}^{\infty} \psi_{k+l} \epsilon_{t-k},$$

so that,

$$X_t(l) = \Psi^{(l)}(B)\epsilon_t, \qquad \text{say}$$

where,

$$\Psi^{(l)}(z) = \sum_{k=0}^{\infty} \psi_{k+l} z^k.$$

Assuming that $\Psi(z)$ is analytic in and on the unit circle (stationary and invertible) then we can write

$$X_t = \Psi(B)\epsilon_t \quad \text{and} \quad \epsilon_t = \Psi^{-1}(B)X_t,$$

and thus

$$\begin{aligned} X_t(l) = \Psi^{(l)}(B)\epsilon_t &= \Psi^{(l)}(B)\Psi^{-1}(B)X_t \\ &= G^{(l)}(B)X_t, \qquad \text{say} \end{aligned}$$

with,

$$G^{(l)}(z) = \Psi^{(l)}(z)\Psi^{-1}(z).$$

If we consider the sequence of predictors $X_t(l)$ for different values of t (with l fixed) then this forms a new process, which since

$$X_t(l) = G^{(l)}(B)X_t,$$

may be regarded as the output of a linear filter acting on the $\{X_t\}$.

6.2 Examples

Video 47

6.2.1 AR(1)

$$X_t - \phi_{1,1} X_{t-1} = \epsilon_t \quad |\phi_{1,1}| < 1.$$

Then

$$X_t = (1 - \phi_{1,1} B)^{-1} \epsilon_t.$$

So,

$$\begin{aligned} \Psi(z) &= 1 + \phi_{1,1} z + \phi_{1,1}^2 z^2 + \ldots \\ &= \psi_0 + \psi_1 z + \psi_2 z^2 + \ldots \end{aligned}$$

i.e., $\psi_k = \phi_{1,1}^k$. Hence,

$$\begin{aligned} X_t(l) &= \sum_{k=0}^{\infty} \delta_k \epsilon_{t-k} = \sum_{k=0}^{\infty} \psi_{k+l} \epsilon_{t-k} \\ &= \sum_{k=0}^{\infty} \phi_{1,1}^{k+l} \epsilon_{t-k} = \phi_{1,1}^l \sum_{k=0}^{\infty} \phi_{1,1}^k \epsilon_{t-k} \\ &= \phi_{1,1}^l X_t. \end{aligned}$$

The l-step prediction variance is

$$\sigma^2(l) = \sigma_\epsilon^2 \left(\sum_{k=0}^{l-1} \psi_k^2 \right) = \sigma_\epsilon^2 \left(\sum_{k=0}^{l-1} \phi_{1,1}^{2k} \right) = \sigma_\epsilon^2 \frac{(1 - \phi_{1,1}^{2l})}{(1 - \phi_{1,1}^2)}.$$

Alternatively,

$$X_t(l) = G^{(l)}(B) X_t,$$

with $G^{(l)}(z) = \Psi^{(l)}(z) \Psi^{-1}(z)$. But,

$$\Psi^{(l)}(z) = \sum_{k=0}^{\infty} \psi_{k+l} z^k = \sum_{k=0}^{\infty} \phi_{1,1}^{k+l} z^k,$$

and,

$$\Psi^{-1}(z) = 1 - \phi_{1,1} z,$$

so that

$$\begin{aligned} G^{(l)}(z) &= (\phi_{1,1}^l + \phi_{1,1}^{l+1} z + \phi_{1,1}^{l+2} z^2 + \ldots)(1 - \phi_{1,1} z) \\ &= \phi_{1,1}^l, \end{aligned}$$

i.e., $X_t(l) = \phi_{1,1}^l X_t$ as before.

We have demonstrated that for the AR(1) model the linear least squares predictor of X_{t+l} depends only on the most recent observation, X_t, and does not involve $X_{t-1}, X_{t-2}, \ldots$, which is what we would expect bearing in mind the Markov nature of the AR(1) model. As $l \to \infty$, $X_t(l) \to 0$, since $X_t(l) = \phi_{1,1}^l X_t$ and $|\phi_{1,1}| < 1$. Also, the l-step prediction variance,

$$\sigma^2(l) \to \frac{\sigma_\epsilon^2}{(1-\phi_{1,1}^2)} = \text{var}\{X_t\}.$$

In fact the solution to the forecasting problem for the AR(1) model can be derived directly from the difference equation,

$$X_t - \phi_{1,1} X_{t-1} = \epsilon_t.$$

by setting future innovations ϵ_t to be zero:

$$\begin{aligned} X_t(1) &= \phi_{1,1} X_t + 0 \\ X_t(2) &= \phi_{1,1} X_t(1) + 0 \\ &\vdots \\ X_t(l) &= \phi_{1,1} X_t(l-1) + 0 \end{aligned}$$

so that,

$$X_t(l) = \phi_{1,1}^l X_t.$$

For general AR(p) processes it turns out that $X_t(l)$ depends only on the last p observed values of $\{X_t\}$, and may be obtained by solving the AR(p) difference equation with the future $\{\epsilon_t\}$ set to zero. For example for an AR(p) process and $l = 1$,

$$X_t(1) = \phi_{1,p} X_t + \ldots + \phi_{p,p} X_{t-p+1}.$$

6.2.2 ARMA(1,1)

$$(1-\phi_{1,1}B)X_t = (1-\theta_{1,1}B)\epsilon_t.$$

Take $\phi_{1,1}=\phi$ and $\theta_{1,1}=\theta$,

$$X_t = \frac{(1-\theta B)}{(1-\phi B)}\epsilon_t = \Psi(B)\epsilon_t.$$

So,

$$\begin{aligned}\Psi(z) &= (1-\theta z)(1+\phi z+\phi^2 z^2+\phi^3 z^3+\ldots)\\ &= 1+(\phi-\theta)z+\phi(\phi-\theta)z^2+\ldots+\phi^{l-1}(\phi-\theta)z^l+\ldots\\ &= \psi_0+\psi_1 z+\psi_2 z^2+\ldots\end{aligned}$$

So,

$$\psi_l = \begin{cases} 1 & l=0\\ \phi^{l-1}(\phi-\theta) & l\geq 1\end{cases}$$

The l-step prediction variance is

$$\begin{aligned}\sigma^2(l) &= \sigma_\epsilon^2\left(\sum_{k=0}^{l-1}\psi_k^2\right)=\sigma_\epsilon^2\left(1+\sum_{k=1}^{l-1}\psi_k^2\right)\\ &= \sigma_\epsilon^2\left(1+(\phi-\theta)^2\sum_{k=1}^{l-1}\phi^{2k-2}\right)\\ &= \sigma_\epsilon^2\left(1+(\phi-\theta)^2\frac{(1-\phi^{2l-2})}{(1-\phi^2)}\right).\end{aligned}$$

Now,

$$\begin{aligned}\Psi^{(l)}(z) &= \sum_{k=0}^{\infty}\psi_{k+l}z^k\\ &= \phi^{l-1}(\phi-\theta)\sum_{k=0}^{\infty}\phi^k z^k\\ &= \phi^{l-1}(\phi-\theta)(1-\phi z)^{-1},\end{aligned}$$

and,

$$\Psi^{-1}(z) = \frac{(1-\phi z)}{(1-\theta z)}.$$

So,

$$\begin{aligned}G^{(l)}(z) &= \Psi^{(l)}(z)\Psi^{-1}(z)\\ &= \phi^{l-1}(\phi-\theta)(1-\theta z)^{-1},\end{aligned}$$

and,

$$\begin{aligned} X_t(l) &= G^{(l)}(B)X_t \\ &= \phi^{l-1}(\phi-\theta)(1-\theta B)^{-1}X_t. \end{aligned}$$

Consider $l = 1$,

$$\begin{aligned} X_t(1) &= (\phi-\theta)(1-\theta B)^{-1}X_t \\ &= (\phi-\theta)(1+\theta B+\theta^2B^2+\theta^3B^3+\ldots)X_t \\ &= (\phi-\theta)X_t+\theta(\phi-\theta)X_{t-1}+\theta^2(\phi-\theta)X_{t-2}+\ldots \\ &= \phi X_t-\theta\left[X_t-(\phi-\theta)X_{t-1}-\theta(\phi-\theta)X_{t-2}-\ldots-\theta^{k-1}(\phi-\theta)X_{t-k}-\ldots\right] \end{aligned}$$

But consider,

$$\begin{aligned} \epsilon_t &= \Psi^{-1}(B)X_t = (1-\phi B)(1-\theta B)^{-1}X_t \\ &= (1-\phi B)(1+\theta B+\theta^2B^2+\theta^3B^3+\ldots)X_t \\ &= X_t-(\phi-\theta)X_{t-1}-\theta(\phi-\theta)X_{t-2}-\ldots-\theta^{k-1}(\phi-\theta)X_{t-k}-\ldots. \end{aligned}$$

Therefore,

$$X_t(1) = \phi X_t - \theta\epsilon_t.$$

So can again be derived directly from the difference equation,

$$X_t = \phi X_{t-1} - \theta\epsilon_{t-1} + \epsilon_t,$$

by setting future innovations ϵ_t to zero.

Video 49

6.2.3 MA(1) (invertible)

$$X_t = \epsilon_t - \theta_{1,1}\epsilon_{t-1} \qquad |\theta_{1,1}| < 1.$$

So,

$$\begin{aligned} \Psi(z) &= \psi_0+\psi_1z+\psi_2z^2+\ldots \\ &= 1-\theta_{1,1}z \end{aligned}$$

Hence, $\psi_0 = 1; \quad \psi_1 = -\theta_{1,1}; \quad \psi_k = 0, \;\; k \geq 2.$

$$\begin{aligned} X_t(l) &= \sum_{k=0}^{\infty} \psi_{k+l}\epsilon_{t-k} = \Psi^{(l)}(B)\epsilon_t \\ &= \psi_l\epsilon_t + \psi_{l+1}\epsilon_{t-1} + \ldots \end{aligned}$$

So,

$$\begin{aligned} \Psi^{(l)}(z) &= \sum_{k=0}^{\infty} \psi_{k+l}z^k = \psi_l z^0 + \psi_{l+1}z^1 + \cdots \\ &= \begin{cases} -\theta_{1,1} & l = 1 \\ 0 & l \geq 2. \end{cases} \end{aligned}$$

Hence,

$$G^{(l)}(z) = \Psi^{(l)}(z)\Psi^{-1}(z) = \begin{cases} -\theta_{1,1}(1-\theta_{1,1}z)^{-1} & l = 1 \\ 0 & l \geq 2. \end{cases}$$

Thus, for $l = 1$,

$$G^{(1)}(z) = -\theta_{1,1}(1 + \theta_{1,1}z + \theta_{1,1}^2 z^2 + \ldots),$$

and hence,

$$\begin{aligned} X_t(1) &= G^{(1)}(B)X_t \\ &= -\sum_{k=0}^{\infty} \theta_{1,1}^{k+1} X_{t-k} \end{aligned}$$

6.3 Forecast errors and updating

Video 50

We have seen that when $\delta_k = \psi_{k+l}$ the forecast error is $\sum_{k=0}^{l-1} \psi_k \epsilon_{t+l-k}$.

Let,

$$\begin{aligned} e_t(l) &= X_{t+l} - X_t(l) \\ &= \sum_{k=0}^{l-1} \psi_k \epsilon_{t+l-k}. \end{aligned}$$

Then,

$$e_t(l+m) = \sum_{j=0}^{l+m-1} \psi_j \epsilon_{t+l+m-j}.$$

Clearly,

$$\mathrm{E}\{e_t(l)\} = \mathrm{E}\{e_t(l+m)\} = 0.$$

Hence,

$$\mathrm{cov}\{e_t(l), e_t(l+m)\} = \mathrm{E}\{e_t(l)e_t(l+m)\} = \sigma_\epsilon^2 \sum_{k=0}^{l-1} \psi_k \psi_{k+m} \qquad (j = k - m),$$

and

$$\mathrm{var}\{e_t(l)\} = \sigma_\epsilon^2 \sum_{k=0}^{l-1} \psi_k^2 = \sigma^2(l).$$

E.g.,

$$\mathrm{cov}\{e_t(1), e_t(2)\} = \sigma_\epsilon^2 \psi_1.$$

This could be quite large – should the forecast for a series wander off target, it is possible for it to remain there in the short run since forecast errors can be quite highly correlated. Hence, when X_{t+1} becomes available we should update the forecast.

$$\begin{aligned} X_{t+1}(l) &= \sum_{k=0}^{\infty} \psi_{k+l} \epsilon_{t+1-k} \\ &= \psi_l \epsilon_{t+1} + \psi_{l+1} \epsilon_t + \psi_{l+2} \epsilon_{t-1} + \ldots, \end{aligned}$$

but,

$$\begin{aligned} X_t(l+1) &= \sum_{k=0}^{\infty} \psi_{k+l+1} \epsilon_{t-k} \\ &= \psi_{l+1} \epsilon_t + \psi_{l+2} \epsilon_{t-1} + \psi_{l+3} \epsilon_{t-2} + \ldots, \end{aligned}$$

and,

$$\begin{aligned} X_{t+1}(l) &= X_t(l+1) + \psi_l \epsilon_{t+1} \\ &= X_t(l+1) + \psi_l (X_{t+1} - X_t(1)). \end{aligned}$$

Hence, to forecast X_{t+l+1} we can modify the $l+1$- step ahead forecast at time t by producing an l-step ahead forecast at time $t+1$ using X_{t+1} as it becomes available.

Video 39

Chapter 7

Bivariate Time Series

Diagram: origins of bivariate time series

7.1 Joint stationarity and cross-covariance

Video 40

7.1.1 Joint stationarity

The two real-valued discrete time stochastic processes $\{X_{1,t}\}$ and $\{X_{2,t}\}$ are said to be jointly stationary stochastic processes if $\{X_{1,t}\}$ and $\{X_{2,t}\}$ are each, separately, second-order stationary processes, and $\text{cov}\{X_{1,t}, X_{2,t+\tau}\}$ is a function of τ only. Then $\{X_{1,t}; X_{2,t}\}$ forms a stationary bivariate process.

7.1.2 Cross-covariance

The acvs are

$$s_{X_1,\tau} = \text{E}\{[X_{1,t} - \mu_{X_1}][X_{1,t+\tau} - \mu_{X_1}]\}$$

$$s_{X_2,\tau} = \mathrm{E}\{[X_{2,t} - \mu_{X_2}][X_{2,t+\tau} - \mu_{X_2}]\}$$

so that,

$$\begin{aligned} s_{X_1,0} &= \mathrm{var}\{X_{1,t}\} = \sigma^2_{X_1} \\ s_{X_2,0} &= \mathrm{var}\{X_{2,t}\} = \sigma^2_{X_2}. \end{aligned}$$

The cross-covariance sequence (ccvs) is given by

$$\begin{aligned} s_{X_1X_2,\tau} &= \mathrm{cov}\{X_{1,t}, X_{2,t+\tau}\} \\ &= \mathrm{E}\{[X_{1,t} - \mu_{X_1}][X_{2,t+\tau} - \mu_{X_2}]\}. \end{aligned}$$

The cross-correlation sequence (ccs) is

$$\rho_{X_1X_2,\tau} = \frac{s_{X_1X_2,\tau}}{\sqrt{s_{X_1,0}s_{X_2,0}}} = \frac{s_{X_1X_2,\tau}}{\sigma_{X_1}\sigma_{X_2}}.$$

Note that,

$$\begin{aligned} s_{X_2X_1,\tau} &= \mathrm{cov}\{X_{2,t}, X_{1,t+\tau}\} \\ &= \mathrm{E}\{[X_{2,t} - \mu_{X_2}][X_{1,t+\tau} - \mu_{X_1}]\}. \end{aligned}$$

Hence,

$$\begin{aligned} s_{X_1X_2,\tau} &= s_{X_2X_1,-\tau} \quad \text{but} \\ s_{X_1X_2,\tau} &\neq s_{X_1X_2,-\tau} \quad \text{(unlike acvs)} \end{aligned}$$

The ccvs is generally quite asymmetric.

Diagram: cross-covariance sequence

Estimation

Given

$$X_{1,1}, X_{1,2}, \ldots, X_{1,N}$$

$$X_{2,1}, X_{2,2}, \ldots, X_{2,N}$$

a natural estimator for the ccvs is

$$\hat{s}_{X_1X_2,\tau} = \begin{cases} \frac{1}{N}\sum_{t=1}^{N-\tau}(X_{1,t} - \bar{X}_1)(X_{2,t+\tau} - \bar{X}_2) & \tau = 0, 1, 2, \ldots, N-1 \\ \frac{1}{N}\sum_{t=1-\tau}^{N}(X_{1,t} - \bar{X}_1)(X_{2,t+\tau} - \bar{X}_2) & \tau = -1, -2, \ldots, -(N-1), \end{cases}$$

so that the estimated ccs is

$$\hat{\rho}_{X_1X_2,\tau} = \frac{\hat{s}_{X_1X_2,\tau}}{\hat{\sigma}_{X_1}\hat{\sigma}_{X_2}}.$$

Linear filtering with noise Video 41

$$X_{2,t} = \sum_{u=-k}^{k} g_u X_{1,t-u} + \eta_t$$

where $\{X_{1,t}\}$ and $\{X_{2,t}\}$ are zero mean stationary processes, $\{\eta_t\}$ is a zero mean (possible coloured) noise with variance σ_η^2, uncorrelated with $\{X_{1,t}\}$.

Diagram: linear filter with noise model

Then,

$$\begin{aligned} s_{X_1X_2,\tau} &= \mathrm{cov}\{X_{1,t}, X_{2,t+\tau}\} \\ &= \mathrm{E}\{X_{1,t}X_{2,t+\tau}\} \\ &= \mathrm{E}\left\{X_{1,t}\left[\sum_{u=-k}^{k} g_u X_{1,t+\tau-u} + \eta_{t+\tau}\right]\right\} \\ &= \sum_{u=-k}^{k} g_u \mathrm{E}\{X_{1,t}, X_{1,t+\tau-u}\} \\ &= \sum_{u=-k}^{k} g_u s_{X_1,\tau-u}. \end{aligned}$$

Since,

$$\begin{aligned}
\sigma^2_{X_2} &= \text{var}\{X_{2,t}\} = \text{E}\{X^2_{2,t}\} \\
&= \text{E}\left\{\left(\sum_{u=-k}^{k} g_u X_{1,t-u} + \eta_t\right)^2\right\} \\
&= \text{E}\left\{\left(\sum_{u=-k}^{k} g_u X_{1,t-u}\right)^2\right\} + \text{E}\{\eta_t^2\} \\
&= \text{E}\left\{\sum_{u=-k}^{k} g_u X_{1,t-u} \sum_{v=-k}^{k} g_v X_{1,t-v}\right\} + \sigma^2_\eta \\
&= \sum_{u=-k}^{k}\sum_{v=-k}^{k} g_u g_v \text{E}\{X_{1,t-u}X_{1,t-v}\} + \sigma^2_\eta \\
&= \sum_{u=-k}^{k}\sum_{v=-k}^{k} g_u g_v s_{X_1,u-v} + \sigma^2_\eta
\end{aligned}$$

the ccs is

$$\rho_{X_1X_2,\tau} = \frac{\sum_{u=-k}^{k} g_u s_{X_1,\tau-u}}{\sigma_{X_1}\sqrt{\sum_{u=-k}^{k}\sum_{v=-k}^{k} g_u g_v s_{X_1,u-v} + \sigma^2_\eta}}.$$

7.2 Cross-Spectra

Video 42

Consider frequency domain characterization of the real-valued bivariate process $\{X_{1,t}; X_{2,t}\}$. Assume that $\{X_{1,t}\}$ and $\{X_{2,t}\}$ are both zero mean processes with spectral density functions

$$S_{X_j}(f) = \sum_{\tau=-\infty}^{\infty} s_{X_j,\tau} e^{-i2\pi f\tau}; \quad |f| \le 1/2, \;\; j = 1, 2.$$

Then the cross spectra are

$$S_{X_jX_k}(f) = \sum_{\tau=-\infty}^{\infty} s_{X_jX_k,\tau} e^{-i2\pi f\tau}; \quad |f| \le 1/2, \;\; j \neq k = 1, 2,$$

assuming the ccvs is square summable.

Note that for real processes $S^*_{X_jX_k}(f) = S_{X_jX_k}(-f)$.

Inverse Fourier transformation gives

$$s_{X_jX_k,\tau} = \int_{-1/2}^{1/2} S_{X_jX_k}(f) e^{i2\pi f\tau}\, df.$$

Now write

$$X_{j,t} = \int_{-1/2}^{1/2} e^{i2\pi ft}\, dZ_{X_j}(f); \quad X_{k,t} = \int_{-1/2}^{1/2} e^{i2\pi f't}\, dZ_{X_k}(f'),$$

so that,

$$\begin{aligned} s_{X_jX_k,\tau} &= \mathrm{cov}\{X_{j,t}, X_{k,t+\tau}\} \\ &= \mathrm{E}\{X_{j,t}X_{k,t+\tau}\} \\ &= \mathrm{E}\{X^*_{j,t}X_{k,t+\tau}\} \\ &= \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} e^{-i2\pi ft} e^{i2\pi f'(t+\tau)} \mathrm{E}\{dZ^*_{X_j}(f) dZ_{X_k}(f')\}. \end{aligned}$$

But this must be a function of τ only, so that $\mathrm{E}\{dZ^*_{X_j}(f) dZ_{X_k}(f')\} = 0$ for $f \neq f'$, i.e., $dZ^*_{X_j}$ and dZ_{X_k} are cross-orthogonal as well as individually orthogonal.

Hence,

$$\begin{aligned} s_{X_jX_k,\tau} &= \int_{-1/2}^{1/2} e^{i2\pi f\tau} \mathrm{E}\{dZ^*_{X_j}(f) dZ_{X_k}(f)\} \\ \Rightarrow S_{X_jX_k}(f)\, df &= \mathrm{E}\{dZ^*_{X_j}(f) dZ_{X_k}(f)\} \\ \Rightarrow S^*_{X_kX_j}(f) &= S_{X_jX_k}(f). \end{aligned}$$

The complete spectral properties are given by the spectral matrix

$$S(f) = \begin{pmatrix} S_{X_1}(f) & S_{X_1X_2}(f) \\ S_{X_2X_1}(f) & S_{X_2}(f) \end{pmatrix}.$$

Since $S_{X_jX_k}(f)$ is a complex quantity we can write it as

$$S_{X_jX_k}(f) = |S_{X_jX_k}(f)|e^{i\theta_{X_jX_k}(f)},$$

where $|S_{X_jX_k}(f)|$ is the cross-amplitude spectrum

$\theta_{X_jX_k}(f)$ is the phase spectrum.

$\theta_{X_jX_k}(f)$ is defined only up to an integer multiple of 2π (since $e^{i2\pi} = e^{i4\pi} = \ldots = 1$).

Worked example: group delay

Coherence

The quantity

$$\gamma^2_{X_jX_k}(f) = \frac{|S_{X_jX_k}(f)|^2}{S_{X_j}(f)S_{X_k}(f)},$$

is called the magnitude squared coherence at f.

It is a real valued coefficient such that

$$0 \leq \gamma^2_{X_jX_k}(f) \leq 1.$$

It measures the linear correlation between the components of $\{X_{j,t}\}$ and $\{X_{k,t}\}$ at frequency f in the same sense as the coefficient of determination R^2 does in ordinary regression.

7.2.1 Example

Figure 37 shows measurements of ocean waves versus time recorded by two different instruments. Figure 38 shows the elements of the estimated spectral matrix for all frequencies. The Nyquist frequency is $1/(2 \times (4/30)) = 3.75$Hz but the frequency axis has been truncated at 0.5Hz. Figure 39 shows the estimated cross-amplitude and phase spectra. Figure 40 shows the estimated coherence. The coherence between the datasets is highest around 0.1-0.2Hz, so this is the frequency range where the instruments behave most similarly (since they are measuring the same waves).

7.3 Linear filtering with noise

Video 43

The model is

$$X_{2,t} = \sum_{u=-k}^{k} g_u X_{1,t-u} + \eta_t.$$

Then

$$\begin{aligned} S_{X_1X_2}(f) &= \sum_{\tau=-\infty}^{\infty} s_{X_1X_2,\tau} e^{-i2\pi f\tau} \\ &= \sum_{u=-k}^{k} g_u \sum_{\tau=-\infty}^{\infty} s_{X_1,\tau-u} e^{-i2\pi f\tau} \\ &= \sum_{u=-k}^{k} g_u e^{-i2\pi fu} \sum_{\tau=-\infty}^{\infty} s_{X_1,\tau-u} e^{-i2\pi f(\tau-u)} \\ &= G(f) S_{X_1}(f). \end{aligned}$$

We can write the model as:

$$\int_{-1/2}^{1/2} e^{i2\pi ft}\, dZ_{X_2}(f) = \sum_{u=-k}^{k} g_u \int_{-1/2}^{1/2} e^{i2\pi f(t-u)}\, dZ_{X_1}(f) + \int_{-1/2}^{1/2} e^{i2\pi ft}\, dZ_\eta(f).$$

Hence,

$$dZ_{X_2}(f) = \sum_{u=-k}^{k} g_u e^{-i2\pi fu} dZ_{X_1}(f) + dZ_\eta(f).$$

Thus,

$$\mathrm{E}\{|dZ_{X_2}(f)|^2\} = \sum_{u=-k}^{k} g_u e^{-i2\pi fu} \sum_{v=-k}^{k} g_v e^{i2\pi fv} \mathrm{E}\{|dZ_{X_1}(f)|^2\} + \mathrm{E}\{|dZ_\eta(f)|^2\}$$

since cross-products have expectation zero.

Hence,

$$S_{X_2}(f) = |G(f)|^2 S_{X_1}(f) + S_\eta(f).$$

Then,

$$\begin{aligned} \gamma^2_{X_1X_2}(f) &= \frac{|G(f)|^2 S^2_{X_1}(f)}{S_{X_1}(f)[|G(f)|^2 S_{X_1}(f) + S_\eta(f)]} \\ &= \left[1 + \frac{S_\eta(f)}{|G(f)|^2 S_{X_1}(f)}\right]^{-1}. \end{aligned}$$

Now,

$$\begin{aligned} S_\eta(f) &= S_{X_2}(f) - |G(f)|^2 S_{X_1}(f) \\ &= S_{X_2}(f)\left[1 - \frac{|G(f)|^2}{S_{X_2}(f)} S_{X_1}(f)\right]. \end{aligned}$$

But,

$$\gamma^2_{X_1X_2}(f) = \frac{|G(f)|^2 S^2_{X_1}(f)}{S_{X_1}(f)S_{X_2}(f)} = \frac{|G(f)|^2 S_{X_1}(f)}{S_{X_2}(f)},$$

so,

$$\begin{aligned} S_\eta(f) &= S_{X_2}(f)[1 - \gamma^2_{X_1X_2}(f)] \\ \text{“noise”} \qquad & \text{“total times unexplained proportion”} \end{aligned}$$

7.4 Bivariate autoregressive processes

Video 44

A bivariate model arises as an extension to the univariate AR(p) process. Let

$$\boldsymbol{X}_t = \begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon}_t = \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix}.$$

The VAR(p) model can be expressed as

$$\begin{aligned} \boldsymbol{X}_t &= \phi_{1,p}\boldsymbol{X}_{t-1} + \ldots + \phi_{p,p}\boldsymbol{X}_{t-p} + \boldsymbol{\epsilon}_t, \\ \Phi(B)\boldsymbol{X}_t &= \boldsymbol{\epsilon}_t \end{aligned}$$

where,

$$\Phi(B) = \mathrm{I} - \phi_{1,p}B - \phi_{2,p}B^2 - \ldots - \phi_{p,p}B^p,$$

where I is the (2×2) identity matrix, and now $\{\phi_{i,p}\}$ are (2×2) matrices of parameters.

$\boldsymbol{\epsilon}_t$ is a bivariate white noise process, such that

$$\mathrm{E}\{\boldsymbol{\epsilon}_t\} = 0$$

and

$$\mathrm{E}\{\boldsymbol{\epsilon}_s \boldsymbol{\epsilon}_t^T\} = \begin{cases} \Sigma, & t = s \\ 0 & \text{otherwise} \end{cases}$$

and Σ is a (2×2) covariance matrix. Thus the elements of $\boldsymbol{\epsilon}_t$ may be correlated.

The condition for stationarity is that the roots of the *determinantal polynomial*, $|\Phi(z)|$, lie outside the unit circle.

Worked example: bivariate VAR(1)

Determine whether the bivariate system

$$\begin{aligned} X_{1,t} &= \frac{1}{2} X_{1,t-1} + \frac{1}{10} X_{2,t-1} + \epsilon_{1,t} \\ X_{2,t} &= \frac{1}{2} X_{1,t-1} + \frac{1}{2} X_{2,t-1} + \epsilon_{2,t} \end{aligned}$$

is stationary.