FORECASTING
STAT0010

Alexandros Beskos

a.beskos@ucl.ac.uk

Course notes, these slides, exercise sheets all on Moodle.

'Events'

- Practical session: Wilkins Building Gustave Tuck LT, 01/02/19.
- Computer practical session 1:
  Group 1: Foster Court B29 - Public Cluster, 18/02/19.
  Group 2: Christopher Ingold G20 - Public Cluster, 22/02/19.
- Computer practical session 2:
  Group 1: Foster Court B29 - Public Cluster, 11/03/19.
  Group 2: Christopher Ingold G20 - Public Cluster, 17/03/19.
- **ICA:** Wilkins Building Gustave Tuck LT &
  Bedford Way (26) LG04, **25/02/19**.

Assessment

- 20% ICA
- 80% Exam

Office hours: Mondays 11.30-13.00

Some literature...

- **The course notes and course slides (Moodle)**
- Chatfield, C., *The analysis of time series*, Chapman and Hall
- Cryer, J., *Time Series Analysis*, PWS Publishers
- Brockwell, P. & Davies, R.,
  Introduction to Time Series and Forecasting, Springer
- Box, G. & Jenkins, G.,
  Time series analysis; forecasting and control, Holden Day
- Brockwell, P. & Davis, R., *Time Series: Theory and Methods*,
  Springer
- Diggle, P., *Time series; a biostatistical introduction*, Oxford
- Harvey, A., *Time series models*, MIT Press.
- Montgomery, D., Jenning, C. & Kulahci, M.,
  Introduction to Time Series Analysis and Forecasting
- Pena, D., Tiao, G. & Tsay, R.,
  A Course in Time Series Analysis, Wiley

Definition 1 (Time series)

A time series is a sequence of observations measured over consecutive points in time.

Examples

- Economics
  - stock market
  - unemployment figures
  - sales growth
- Medicine
  - flu cases
  - blood pressure
  - ECG
- Social science
  - demographics
  - public opinion
  - judicial decisions
- And many more!

Generally, time series comprise: **trend terms**, **seasonal behaviour**, **short term correlations**, and **purely random 'noise'**.

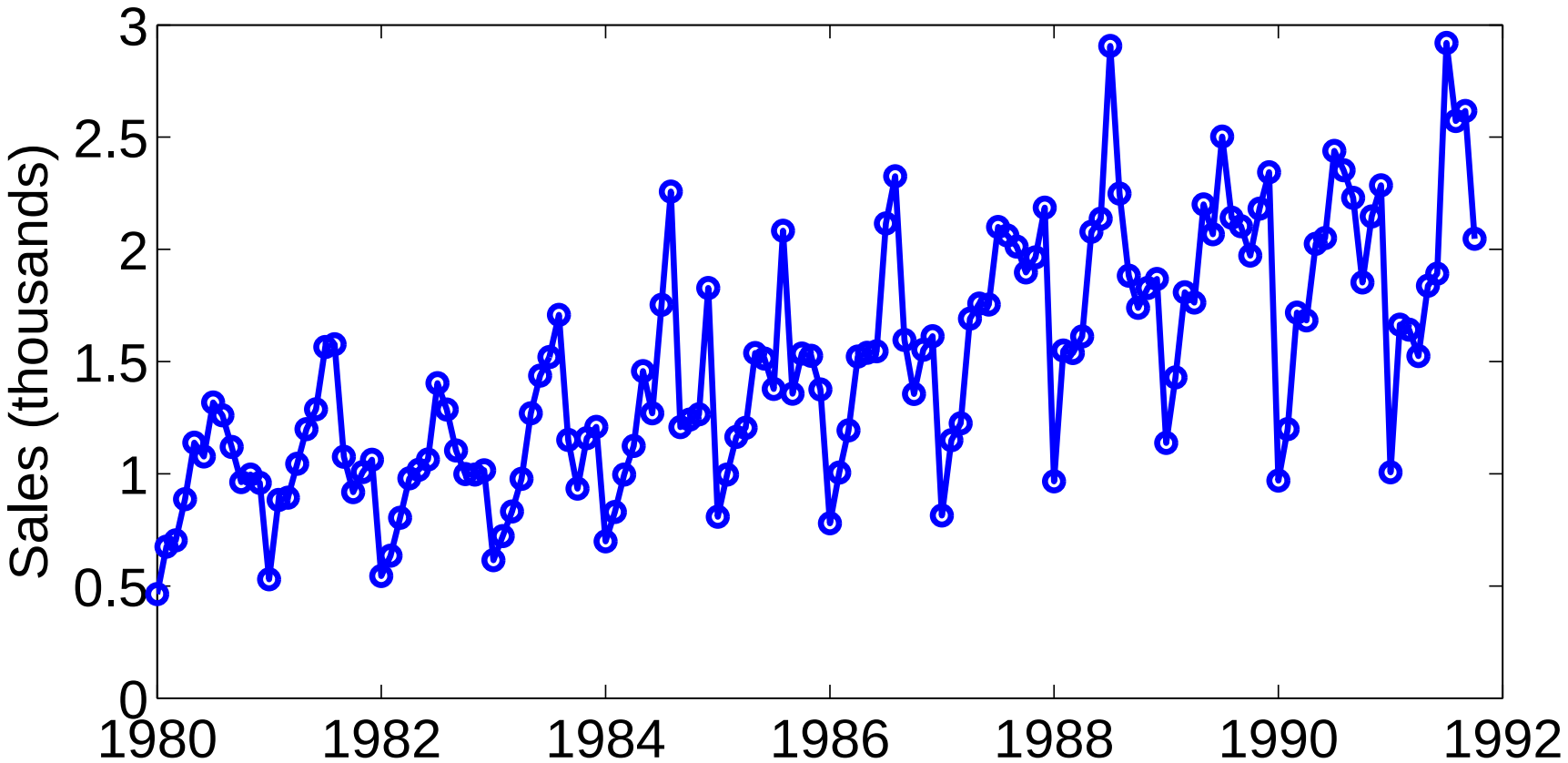

Figure: Australian red wine sales, Jan. '80 — Oct. '91

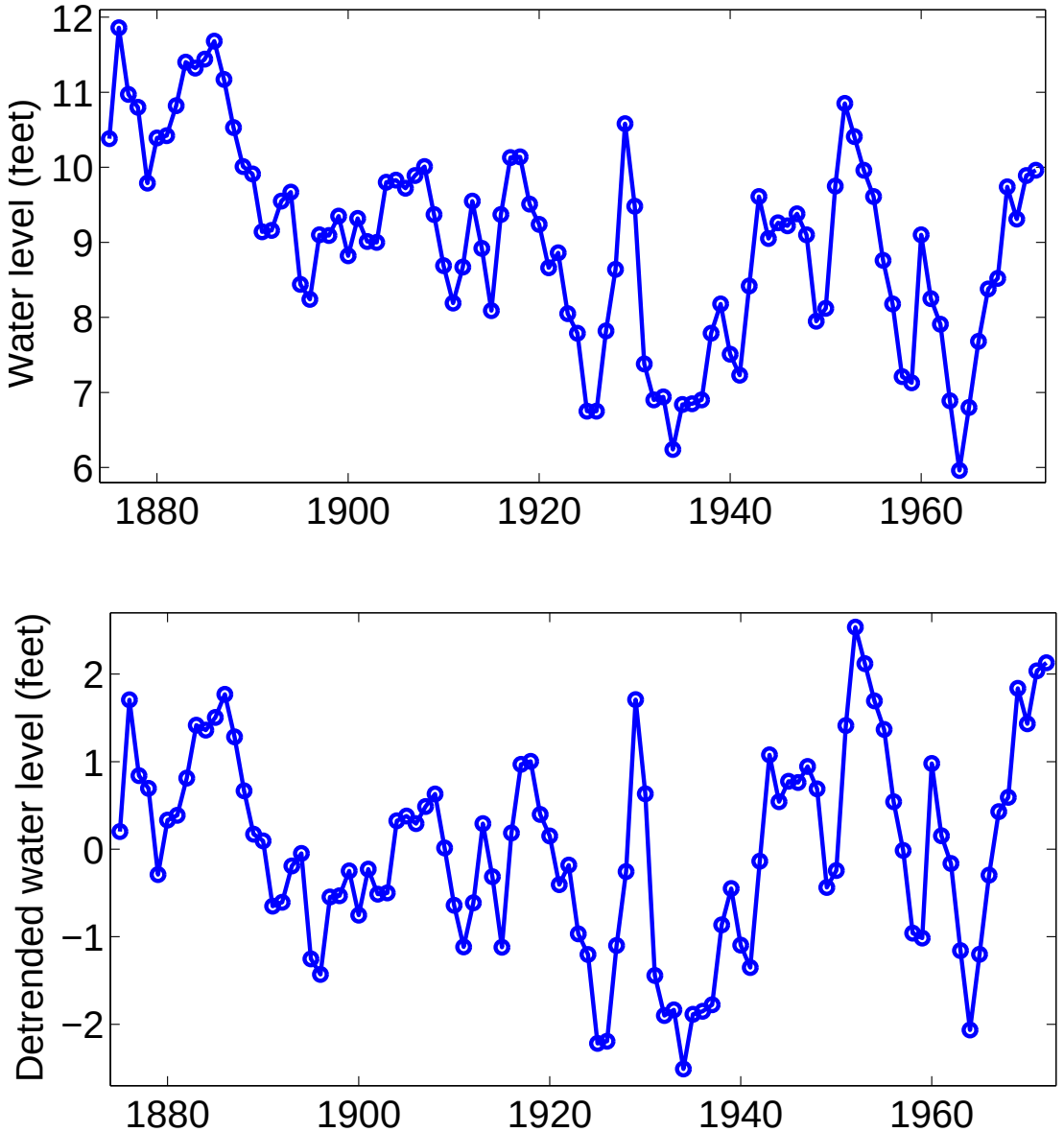

Figure: Water levels in Lake Huron, 1875-1972

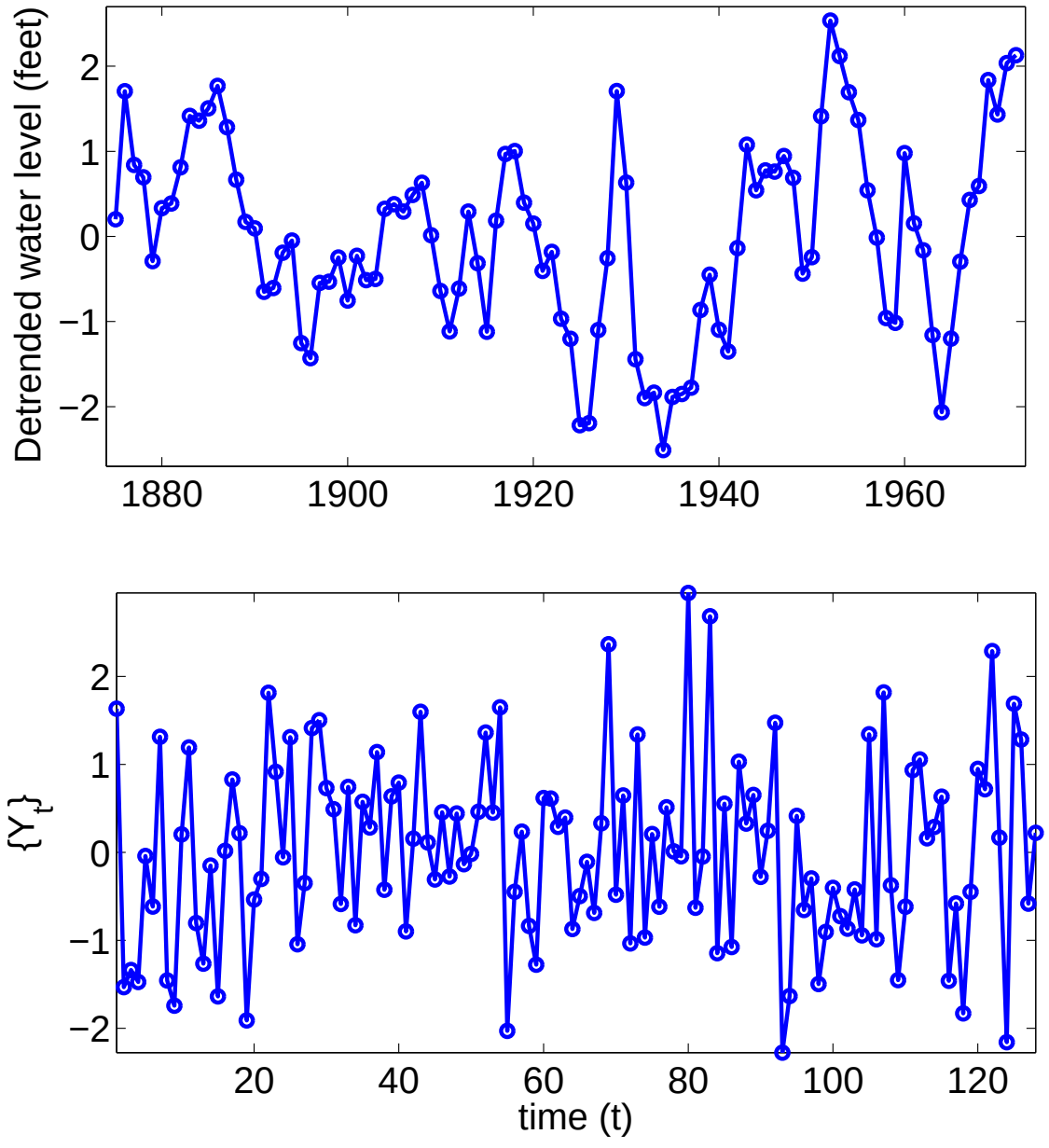

Figure: Comparison: (top) detrended Lake Huron; (bottom) white noise

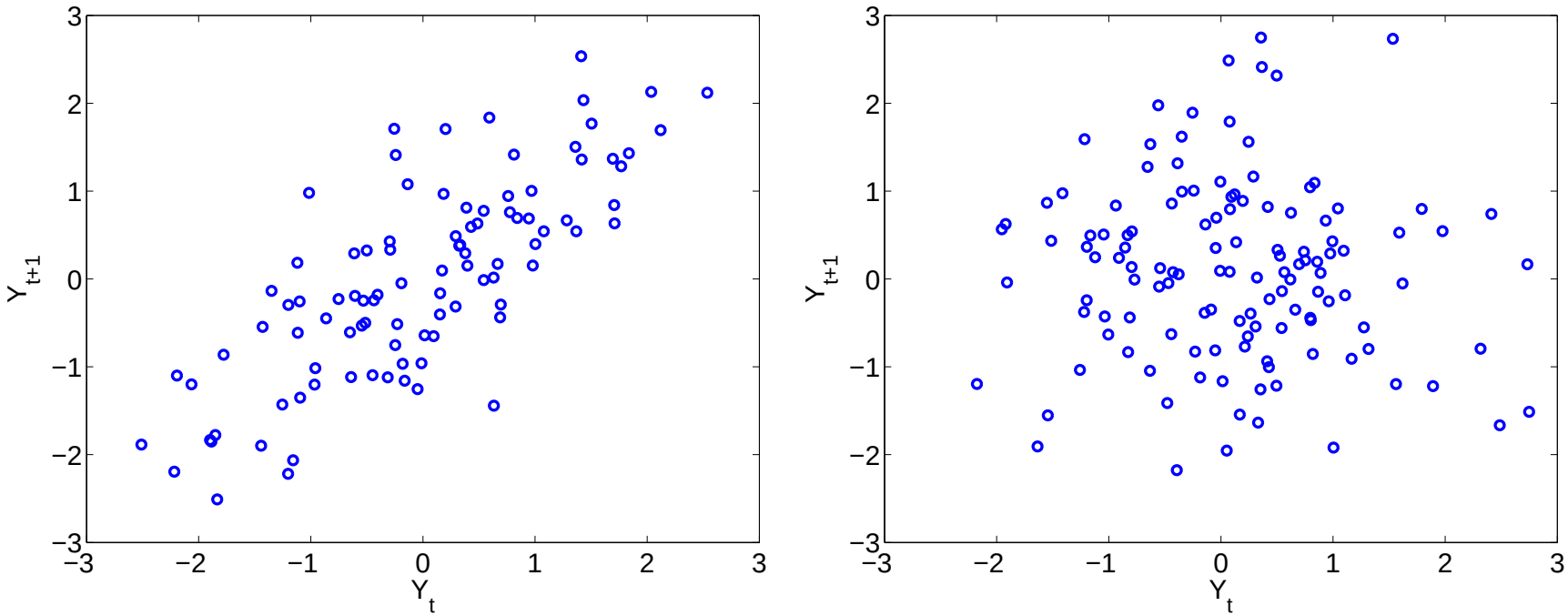

Figure: Scatter plots: (left) detrended Huron; (right) white noise

It is always very difficult to predict the future. Indeed it has been likened to driving a car blindfolded while following instructions given by a person looking out of the back window.

Professor A.C. Harvey (Dept of Statistical and Mathematical Sciences, LSE)

. . . the record of forecasters in the past thirty or forty years, whatever their professional qualification as prophets, has been so spectacularly bad that only governments and economic research institutes still have, or pretend to have, much confidence in it.

Eric Hobsbawm (historian), in Age of Extremes: the Short Twentieth Century 1914–1991

The Box-Jenkins methodology for forecasting

1. **Model identification**
2. **Parameter estimation**
3. **Verification**

   Check model obtained from 1 & 2

   - Good? Goto 4
   - Bad? Goto 1 & decide on new model

4. **Forecasting**

Course outline

1. Revision
2. Stationarity
3. Stationary models (ARMA)
4. Non-stationary models (ARIMA)
5. Seasonal models (SARIMA)
6. Box-Jenkins
7. **Forecasting**
8. Structural models

‘Lecture 1’ Outline

1. Preliminaries
2. Stationarity
3. Autocorrelation function

Let X and Y be RVs then the covariance of X and Y is

$$\operatorname{cov}(X,Y) := \mathbb{E}\big((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\,.$$

Covariance has the following properties:

1. $\operatorname{cov}(X,Y) \in \mathbb{R}$
2. $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$
3. If X and Y are independent (i.e. $X \perp Y$), then $\operatorname{cov}(X,Y) = 0$
4. If $\operatorname{cov}(X,Y) = 0$ this does not necessarily mean that X and Y are independent
5. $\operatorname{cov}(a+bX, c+dY) = bd\operatorname{cov}(X,Y)$
6. If Z is RV $\Rightarrow \operatorname{cov}(X+Z,Y) = \operatorname{cov}(X,Y) + \operatorname{cov}(Z,Y)$
7. $X_1,\ldots,X_n$, and $Y_1,\ldots,Y_m$ RVs, then:

$$\operatorname{cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \operatorname{cov}(X_i, Y_j)$$

8. $\text{cov}(X, X) = \text{var}(X)$
9. $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\ \text{cov}(X, Y)$
10. $X \perp Y \Rightarrow \text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$
11. The correlation coefficient between X and Y is defined as

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

12. $|\text{corr}(X, Y)| \leq 1$

A matrix $\mathbf{A}$ is a rectangular array of numbers with m rows and n columns. Sometimes written as $\mathbf{A} \in \mathbb{R}^{m \times n}$.
Denote the element (number) in row i and column j by A_{ij}. Then

$$\mathbf{A} = \begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,n} \\ A_{2,1} & A_{2,2} & \dots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \dots & A_{m,n} \end{pmatrix}$$

For $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{m \times n}$ matrix addition $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is performed element-wise. I.e. $C_{i,j} = A_{i,j} + B_{i,j}$.
For $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times p}$

$$\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times p}$$

given by

$$C_{i,j} = \sum_{k=1}^{n} A_{i,k} B_{k,j}$$

In particular, for $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{c} = \mathbf{Ab} \in \mathbb{R}^m$ we have $c_i = \sum_{k=1}^{n} A_{i,k} b_k$

Matrix multiplication does not commute. I.e.

$$\mathbf{AB} \neq \mathbf{BA}$$

in general. However, it is associative:

$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$$

$$(\mathbf{B}+\mathbf{C})\mathbf{A} = \mathbf{BA}+\mathbf{CB}$$

$$\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$$

The (i,j)th element of the transpose $\mathbf{A}^T$ of a matrix $\mathbf{A}$ is defined by

$$(\mathbf{A}^T)_{i,j} = \mathbf{A}_{j,i}$$

(I.e. rows of $\mathbf{A}$ become columns of $\mathbf{A}^T$ (and vice versa)

$$\mathbf{A} = \begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,n} \\ A_{2,1} & A_{2,2} & \dots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \dots & A_{m,n} \end{pmatrix}, \quad \mathbf{A}^T = \begin{pmatrix} A_{1,1} & A_{2,1} & \dots & A_{m,1} \\ A_{1,2} & A_{2,2} & \dots & A_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,n} & A_{2,n} & \dots & A_{m,n} \end{pmatrix}$$

Some useful properties include:

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T, \qquad (\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$$

Unless used as an index (!) the symbol i will mean the square root of -1. By definition, $i^2 = -1$.

- Any $z \in \mathbb{C}$ can be written as $x + iy$, for $x, y \in \mathbb{R}$, where x is the real part and y is the imaginary part.
- Let $z_1 = x_1 + iy_1, z_2 = x_2 + iy_2$. Then

$$z_1 + z_2 = x_1 + x_2 + i(y_1 + y_2)$$

- Complex multiplication

$$z_1 z_2 = (x_1 + iy_1)(x_2 + iy_2) = x_1 x_2 + ix_1 y_2 + ix_2 y_1 + i^2 y_1 y_2$$
$$= x_1 x_2 - y_1 y_2 + i(x_1 y_2 + x_2 y_1)$$

- Complex division

$$\frac{z_1}{z_2} = \frac{x_1 + iy_1}{x_2 + iy_2} = \frac{(x_1 + iy_1)(x_2 - iy_2)}{(x_2 + iy_2)(x_2 - iy_2)}$$
$$= \frac{x_1 x_2 + y_1 y_2}{x_2^2 + y_2^2} + i\frac{(y_1 x_2 - x_1 y_2)}{x_2^2 + y_2^2}$$

There is a nice geometric interpretation of complex numbers (see **argand.pdf**):

$$x = R\cos\theta$$
$$y = R\sin\theta$$

i.e. $z = R(\cos\theta + i\sin\theta) = Re^{i\theta} = |z|\, e^{i\theta}$. Here, R is the modulus

$$R = |z| = \sqrt{(x^2 + y^2)}$$

and θ is the argument. Now, let $z_1 = R_1 e^{i\theta_1}$, $z_2 = R_2 e^{i\theta_2}$. Then

$$z_1 z_2 = R_1 R_2 e^{i(\theta_1+\theta_2)}$$

and

$$\frac{z_1}{z_2} = \frac{R_1}{R_2} e^{i(\theta_1-\theta_2)}$$

The set $S = \{z \in \mathbb{C} : |z| = 1\}$ is called the unit circle. Example $\pm 1, \pm i, \sqrt{2}(\pm 1 \pm i)/2, \pm\sqrt{3}/2 \pm i/2 \in S$. If $|z| < 1$ then z is said to be inside the unit circle; if $|z| > 1$ it is outside the unit circle; if $|z| = 1$ it is on the unit circle.

A time series can be thought of as a stochastic process:

$$\{Y_t\} = \{Y_1, \ldots, Y_T\},$$

where Y_t are RVs and $t = 1, \ldots, T$ denotes (discrete) time. In practice, the observed data $\{y_1, \ldots, y_T\}$ is a realisation of $\{Y_t\}$.

Important points to note:

- the order of observations is determined by time
- it is (usually) impossible to make multiple observations at one instant of time
- it is 'difficult' and 'often' not necessary to find joint distrib. of $\{Y_t\}$

i.e., for some fixed t_0 we cannot take lots of samples of Y_{t_0}.

Initially restrict discussion to a specific type of stochastic process.

Definition 2

Let $\{Y_t\}$ be a time series. The mean function of Y_t is defined by:

$$\mu_Y(t) := \mathbb{E}(Y_t)$$

The variance function of Y_t is defined by:

$$\sigma_Y^2(t) := \text{var}(Y_t)$$

The autocovariance function (acf) of Y_t is defined by:

$$\gamma_Y(t,k) := \text{cov}(Y_t, Y_k)$$

Remark 3

$$\text{var}(Y_t) = \gamma_Y(t,t)$$

Now ready to introduce stationarity (restriction).

Definition 4

$\{Y_t\}$ *is* strictly stationary *if* $\forall k, m, t_1, \ldots t_m$*:*

$$(Y_{t_1}, Y_{t_2}, \ldots, Y_{t_m}) \stackrel{d}{=} (Y_{t_1+k}, Y_{t_2+k}, \ldots, Y_{t_m+k})$$

Remark 5

Strict stationarity $\Rightarrow$ *time shift has no effect on joint distribution.*

Remark 6

Put $m = 1$*. Then strict stationarity* $\Rightarrow Y_s \stackrel{d}{=} Y_t, \forall s, t$ *and*

$$\mathbb{E}(Y_s) = \mathbb{E}(Y_t), \quad \mu(t) = \mu, \quad \text{var}(Y_s) = \text{var}(Y_t), \quad \sigma^2(t) = \sigma^2$$

Remark 7

Put $m = 2$*. Then strict stationarity* $\Rightarrow (Y_t, Y_{t+k}) \stackrel{d}{=} (Y_s, Y_{s+k}) \forall k, s, t$ *and*

$$\text{cov}(Y_t, Y_{t+k}) = \text{cov}(Y_s, Y_{s+k})$$

Definition 8

$\{Y_t\}$ *is weakly stationary if* $\forall k, s, t$*:*

1. $\mathbb{E}(Y_s) = \mathbb{E}(Y_t)$, *i.e.* $\mu(t) = \mu$ *(a constant)*
2. $\text{cov}(Y_t, Y_{t+k}) = \text{cov}(Y_s, Y_{s+k})$

Corollary 9

strict stationarity $\Rightarrow$ *weak stationarity*

Remark 10

2 $\Rightarrow \gamma(t, t+k) = \gamma(s, s+k)$ *is a function of lag* k *and we write (define)*

$$\gamma(k) := \text{cov}(Y_t, Y_{t+k})$$

In particular, note that 2 $\Rightarrow \text{var}(Y_t) = \gamma(0)$, *i.e.* $\sigma_Y^2(t) = \sigma_Y^2$ *(const.)*

Definition 11

The autocorrelation function $\rho(k)$ *is defined by:*

$$\rho(k) := \frac{\gamma(k)}{\gamma(0)}$$

Example 12

A series $\{\epsilon_t\}$ *is called white noise if it has* $\mathbb{E}(\epsilon_t) = 0$*, constant variance & is uncorrelated:* $\gamma(k) = \delta_{0,k}\sigma^2$*. We write* $\{\epsilon_t\} \sim \mathcal{WN}(0, \sigma^2)$*.*

Proposition 13 (autocorrelation properties)

1. $\rho(0) = 1$
2. $\rho(-k) = \rho(k)$
3. $|\rho(k)| \leq 1$
4. *Non-uniqueness:* $\{Y_t\} \neq \{X_t\} \nRightarrow \rho_Y \neq \rho_X$

Proof ❶ (by definition) and ❷ (by symmetry of the covariance operator) are trivial, we'll see an example of ❹ later on in the course. For property ❸: for $a, b \in \mathbb{R}$: On Board

In practice, we do not know μ, γ, ρ, but rather we have data and must use estimates.

Definition 14

Let $\{y_t\} = \{y_1, \ldots, y_T\}$ be the observed values of a time series $\{Y_t\}$. The sample mean of $\{y_t\}$ is defined as:

$$\hat{\mu} = \overline{Y} := \frac{1}{T} \sum_{t=1}^{T} y_t$$

The sample autocovariance function is:

$$c(k) = \hat{\gamma}(k) := \frac{1}{T} \sum_{t=k+1}^{T} (y_{t-k} - \hat{\mu})(y_t - \hat{\mu})$$

The sample autocorrelation function is:

$$r(k) = \hat{\rho}(k) := \frac{c(k)}{c(0)} = \frac{\sum_{t=k+1}^{T} (y_{t-k} - \hat{\mu})(y_t - \hat{\mu})}{\sum_{t=k+1}^{T} (y_t - \hat{\mu})^2}$$

A plot of $r(k)$ against lag k is called the correlogram.

Example 15 (White noise)

Consider $\{\epsilon_t\} \sim \mathcal{WN}(0, \sigma^2)$. *Then*

$$\rho(k) = \begin{cases} 1\,, & \text{for } k = 0 \\ 0\,, & \text{otherwise} \end{cases}$$

Remark 16 (See **acf white noise eg**)

It can be shown that, for $k \neq 0$ *and a large sample size* T*:*

$$r_\epsilon(k) \sim \mathcal{N}(0, 1/T)$$

$\Rightarrow$ 95% *confidence interval is* $(-1.96/\sqrt{T}, 1.96/\sqrt{T})$. *I.e.,*

- *for each* k*, we expect* 95% *of realisations of this time series to have* $r_\epsilon(k)$ *inside interval*
- *observed values of* $r_\epsilon(k)$ *that fall outside these limits are considered 'significantly' different from zero at the 5% level*
- *expect to get* 5% *of* $r_\epsilon(k)$ *coeffs outside the* 95% *confidence limits*
- *if you plot* 20 *values of* $r_\epsilon(k)$ *you would expect to see one value outside the limits by chance*