ST6006

Time Series Analysis

Lecture notes
Outline

I Introduction to time series analysis

II Stationarity and ARMA modelling

1. Stationarity
   a. Definitions
   b. Strict stationarity
   c. Weak stationarity
2. Autocovariance, autocorrelation and partial autocorrelation
   a. Autocovariance
   b. Autocorrelation
   c. Partial autocorrelation
   d. Estimation of the ACF and PACF
3. ARMA modelling
   a. AR models
   b. MA models
   c. ARMA models
4. Backward Shift Operator and Difference Operator
5. AR(p) models, stationarity and the Yule-Walker equations
   a. The AR(1) model
   b. The AR(p) model and stationarity
   c. Yule-Walker equations
6. MA(q) models and invertibility
   a. The MA(1) model
   b. The MA(q) model and invertibility
7. ARMA(p,q) models
8. ARIMA(p,d,q) models
   a. Non-ARMA processes
   b. The I(d) notation
9. The Markov property

III Non-stationarity: trends and techniques

1. Typical trends
2. Least squares trend removal
3. Differencing
   a. Linear trend removal
   b. Selection of $d$
4. Seasonal differencing
5. Method of moving averages
6. Seasonal means
7. Filtering, smoothing
8. Transformations

IV Box-Jenkins methodology

1. Overview
2. Model selection
   a. Identification of white noise
   b. Identification of MA(q)
   c. Identification of AR(p)
3. Model fitting
   a. Fitting an ARMA(p,q) model
   b. Parameter estimation: LS and ML
   c. Parameter estimation: method of moments
   d. Diagnostic checking
I. Introduction to time series analysis

A time series is a stochastic process in discrete time with a continuous state space.

Notation: \( \{X_1, X_2, ..., X_n\} \) denotes a time series process, whereas \( \{x_1, x_2, ..., x_n\} \) denotes a univariate time series, i.e. a sequence of realisations of the time series process.

![](image1.png)

I.1 Purposes of Time Series Analysis

- **Describe** the observed time series data:
  - mean, variance, correlation structure, ...
  - e.g. correlation coefficient between sales 1 month apart, 2 months apart, etc.
    - Autocorrelation Function (ACF)
    - Partial Autocorrelation Function (PACF)

- **Construct** a model which fits the data
  - From the class of ARMA models, select a model which best fits the data based on ACF and PACF of the observed time series
  - Apply **Box Jenkins Methodology**:
    - Identify tentative model
    - Estimate model parameters
    - Diagnostic checks - does the model fit?

- **Forecast** future values of the time series process
  - easy, once model has been fitted to past data

All ARMA models are **stationary**. If an observed time series is **non-stationary** (e.g. upward trend), it must be converted to stationary time series (e.g. by **differencing**).
1.2 Other forms of analysis

Another important approach to the analysis of time series relies on the Spectral Density Function; the analysis is then based on the autocorrelation function of a time series model. This approach is not covered in this course.
II. Stationarity and ARMA modelling

II.1 Stationarity

a. Definition

A stochastic process is (strictly) stationary if its statistical properties remain unchanged over time.

Joint distribution of $X_{t1}, X_{t2}, ..., X_{tn} = \text{Joint distribution of } X_{k+t1}, X_{k+t2}, ..., X_{k+tn}$, for all $k$ and for all $n$.

*Example:* Joint distribution of $X_5, X_6, ..., X_{10} = \text{Joint distribution of } X_{120}, X_{121}, ..., X_{125}$

- for any ‘chunk’ of variables
- for any ‘shift’ of start

Implications of (strict) stationarity

*Take $n = 1$:*

- Distribution of $X_t = \text{distribution of } X_{t+k}$ for any integers $k$
  
  \begin{align*}
  X_t \text{ discrete: } & \quad P(X_t = i) = P(X_{t+k} = i) \text{ for any } k \\
  X_t \text{ continuous: } & \quad f(X_t) = f(X_{t+k}) \text{ for any } k \\
  \text{In particular, } & \quad E(X_t) = E(X_{t+k}) \text{ for any } k \\
  & \quad \text{Var}(X_t) = \text{Var}(X_{t+k}) \text{ for any } k
  \end{align*}

- A stationary process has constant mean and variance

- The variables $X_t$ in a stationary process must be identically distributed (but not necessarily independent)
Take $n = 2$:

- Joint Distribution of $(X_s, X_t) = \text{Joint Distribution of } (X_{s+k}, X_{t+k})$
  - for all lags $(t-s)$
  - for all integers $k$
  - depends on the lag $(t-s)$

- In particular, $\text{COV}(X_s, X_t) = \text{COV}(X_{s+k}, X_{t+k})$

where $\text{COV}(X_s, X_t) = E[(X_s - E(X_s))(X_t - E(X_t))]$

- Thus $\text{COV}(X_s, X_t)$ depends only on lag $(t-s)$ and not on time $s$

b. **Strict Stationarity**

- Very stringent requirement
- Hard to prove a process is stationary
- To show a process is not stationary show one condition doesn’t hold

*Examples:*

Simple Random Walk: $\{X_t\}$ not identically distributed

⇒ NOT stationary

White Noise Process: $\{Z_t\}$ i.i.d.

⇒ trivially stationary

c. **Weak Stationarity**

- This requires only that $E(X_t)$ is constant AND $\text{COV}(X_s, X_t)$ depends only on $(t-s)$
- Since $\text{Var}(X_t) = \text{COV}(X_t, X_t)$ this implies that $\text{Var}(X_t)$ is constant
- Weak stationarity does not imply strict stationarity
- For weak stationarity, $\text{COV}(X_t, X_{t+k})$ is constant with respect to $t$ for all lags $k$
- Here (and often), stationary is shorthand for weakly stationary
Question: If the joint distribution of the $X_i$’s is multivariate normal, then weak stationarity implies strong stationarity.

Solution: If $X \sim N(\mu, \Sigma)$ then $X$ is completely determined by $\mu$ and $\Sigma$ (property of the multivariate Normal distribution). If these do not depend on $t$, neither does the distribution of $X$.

Example: $X_t = \sin(\omega t + u), U \sim U[0, 2\pi]$ then $E(X_t) = 0$.

Here $\text{COV}(X_t, X_{t+k}) = \cos(\omega k) E(\sin^2(u))$, hence does not depend on $t$.

$\Rightarrow$ $X_t$ is weakly stationary

Question: If we know $X_0$, then we can work out $u$, since $X_0 = \sin(u)$. We then know all the values of $X_t = \sin(\omega t + u)$

$\Rightarrow$ $X_t$ is completely determined by $X_0$

Definition: $X$ is purely indeterministic if values of $X_1, ..., X_n$ are progressively less useful at predicting $X_N$ as $N \to \infty$.

Here stationary time series means weakly stationary, purely indeterministic process.

II.2 Autocovariance, autocorrelation and partial autocorrelation

a. Autocovariance function

For a stationary process, $E(X_t) = \mu_t = \mu$, for any $t$.

We define $\gamma_k = \text{Cov}(X_t, X_{t+k}) = E(X_t X_{t+k}) - E(X_t) E(X_{t+k})$ the “autocovariance at LAG $k$”.

This function does not depend on $t$.

Autocovariance function of $X$: $\{\gamma_0, \gamma_1, \gamma_2, ...\} = \{\gamma_k : k \geq 0\}$

Note: $\gamma_0 = \text{Var}(X_t)$

Question: Properties of covariance – needed when calculating autocovariances for specified models.

b. Autocorrelation function (ACF)

Recall that $\text{corr}(X,Y) = \text{Cov}(X,Y) / (\sigma_X \sigma_Y)$

For a stationary process, we define $\rho_k = \text{corr}(X_t, X_{t+k}) = \gamma_k/\gamma_0$ the “autocorrelation at lag $k$”.

Time Series Analysis

(This is the usual correlation coefficient, since \(\text{Var}(X_t) = \text{Var}(X_{t+k}) = \gamma_0\)

- **Autocorrelation Function (ACF) of \(X\):** \(\{\rho_0, \rho_1, \rho_2, \ldots\} = \{\rho_k : k \geq 0\}\)

- **Note:** \(\rho_0 = 1\)

- For a purely indeterministic process, we expect \(\rho_k \to 0\) as \(k \to \infty\) (i.e. values far apart will not be correlated)

- **Recall** (ST3053): a sequence of i.i.d. random variables \(\{Z_t\}\) is called a white noise process and is trivially stationary.

**Example:** \(\{e_t\}\) is a zero-mean **white noise process** if

- \(\text{E}(e_t) = 0\) for any \(t\) and

- \(\gamma_k = \text{COV}(e_t, e_{t+k}) = \begin{cases} \sigma^2, & \text{if } k = 0 \\ 0, & \text{otherwise} \end{cases}\)

- **Note:** the variables \(e_t\) have **zero mean**, variance \(\sigma^2\) and are **uncorrelated**

- A sequence of i.i.d. variables with zero mean will be a white noise process, according to this definition. In particular, \(Z_t\) independent, \(Z_t \sim N(0, \sigma^2)\) is a white noise process.

- **Result:** \(\gamma_k = \gamma_{-k}\) and \(\rho_k = \rho_{-k}\)

- **Correlogram** = plot of ACF \(\{\rho_k : k \geq 0\}\) as a function of lag \(k\). It is widely used as it tells a lot about the time series.

c. **Partial autocorrelation function (PACF)**

Let \(r(x,y|z) = \text{corr}(x,y|z)\) denote the partial correlation coefficient between \(x\) and \(y\), adjusted for \(z\) (or with \(z\) held constant).

\[
\begin{array}{ccccccc}
X_t & X_{t+1} & X_{t+2} & \ldots & X_{t+k-1} & X_{t+k} \\
\hline
\end{array}
\]

| \(t\) | \(t+1\) | \(t+2\) | \(t+k-1\) | \(t+k\) |

- Denote:
  - \(\varphi_2 = \text{corr}(x_t, x_{t+2}|x_{t+1})\)
  - \(\varphi_3 = \text{corr}(x_t, x_{t+3}|x_{t+1}, x_{t+2})\)
  - \(\varphi_4 = \text{corr}(x_t, x_{t+k}|x_{t+1}, \ldots x_{t+k-1})\)
  - \(\varphi_k = \text{partial autocorrelation coefficient at lag } k\).
• Partial autocorrelation function (PACF):

\[ \{ \varphi_1, \varphi_2, \ldots \} = \{ \varphi_k, k \geq 1 \} \]

• The \( \varphi_k \)'s are related to the \( \rho_k \)'s:

\[
\varphi_1 = \text{corr}(X_t, X_{t+1}) = \rho_1
\]

Recall that

\[
r(x,y|z) = \frac{r(x,y) - r(x,z)r(y,z)}{\sqrt{1-r^2(x,z)}\sqrt{1-r^2(y,z)}}
\]

Applying this here, using \( x = X_t, y = X_{t+2}, z = X_{t+1}, \varphi_2 = \text{corr}(x_t, x_{t+2}|x_{t+1}) = r(x,y|z) \), along with \( \rho_1 = r(x,z) \) and \( \rho_2 = r(x,y) \), yields:

\[
\varphi_2 = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}
\]

d. Estimation of the ACF and PACF

We assume that the sequence of observations \( \{x_1, x_2, \ldots x_n\} \) comes from a stationary time series process.

The following functions are central to the analysis of time series:

- \( \{ \gamma_k \} \) - Autocovariance function
- \( \{ \rho_k \} \) - Autocorrelation function (ACF)
- \( \{ \phi_k \} \) - Partial Autocorrelation function (PACF)

To find a model to fit the sequence \( \{x_1, x_2, \ldots x_n\} \), we must be able to estimate the ACF of the process of which the data is a realisation. Since the model underlying the data is assumed to be stationary, its mean can be estimated using the sample mean.

\[
\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} x_t
\]

The autocovariance function, \( \gamma_k \), can be estimated using the sample autocovariance function:
\[ \hat{y}_k = \frac{1}{n} \sum_{t=k+1}^{n} (x_t - \hat{\mu})(x_{t-k} - \hat{\mu}) \]

from which are derived estimates, \( r_k \) for the autocorrelation \( \rho_k \):

\[ r_k = \frac{\hat{y}_k}{\hat{y}_0} \]

The collection \( \{r_k : k \in \mathbb{Z}\} \) is called the **sample autocorrelation function (SACF)**. The plot of \( r_k \) against \( k \) is called a **correlogram**.

**Recall** that the partial autocorrelation coefficients \( \phi_k \) are calculated as follows:

\[
\phi_1 = \rho_1 \\
\phi_2 = \begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}
\]

In general, \( \phi_k \) is given as a **ratio of determinants** involving \( \rho_1, \rho_2, \ldots, \rho_k \). The sample partial autocorrelation coefficients are given by these formulae, but with the \( \rho_k \) replaced by their estimates \( r_k \):

\[
\hat{\phi}_1 = r_1 \\
\hat{\phi}_2 = \frac{r_2 - r_1^2}{1 - r_1^2}
\]

etc.

The collection \( \{\hat{\phi}_k\} \) is called the sample partial autocorrelation function (SPACF). The plot of \( \{\hat{\phi}_k\} \) against \( k \) is called the partial **correlogram**.

These are the main tools in identifying a model for a stationary time series.
II.3 ARMA modelling

Autoregressive moving average (ARMA) models constitute the main class of linear models for time series. More specifically:

- Autoregressive (AR)
- Moving Average (MA)
- Autoregressive Moving Average (ARMA)
- Autoregressive Integrated Moving Average (ARIMA)

\[ \begin{align*}
\text{Last type are non-stationary} \\
\text{Others are stationary}
\end{align*} \]

a. AR models

- **Recall**: Markov Chain = process such that the conditional distribution of \( X_{n+1} \), given \( X_n, X_{n-1}, \ldots, X_0 \) depends only on \( X_n \), i.e. “the future depends on the present, but not on the past”
- The simplest type of autoregressive model (AR(1)) has this property: \( X_t = \alpha X_{t-1} + \varepsilon_t \), where \( \varepsilon_t \) is zero-mean white noise.

\[ \begin{align*}
X_{t-2} & \quad X_{t-1} & \quad X_t \\
t-2 & \quad t-1 & \quad t
\end{align*} \]

- For AR(1), we prove that \( \varphi_2 = \text{corr}(X_t, X_{t-2} | X_{t-1}) = 0 \)
- Similarly, \( \varphi_k = 0 \) for \( k > 2 \).
- A more general form of an AR(1) model is

\[ X_t = \mu + \alpha_1 (X_{t-1} - \mu) + \varepsilon_t \]

where \( \mu = \text{E}(X_t) \) is the process mean

- **Autoregressive process of order p (AR(p))**: 

\[ X_t = \mu + \alpha_1 (X_{t-1} - \mu) + \alpha_2 (X_{t-2} - \mu) + \ldots + \alpha_p (X_{t-p} - \mu) + \varepsilon_t \]

b. MA models

A realisation of a white noise process is very ‘jagged’, since successive observations are realisations of independent variables... Most time series observed in practice have a smoother time series plot than a realisation of a white noise process, since in this process the successive observations are realisations of independent variables. In that respect, taking a “moving average” is a standard way of smoothing an observed time series:

Observed data: \( x_1, x_2, x_3, x_4, \ldots \)
Moving average: \[ \frac{1}{3}(x_1 + x_2 + x_3, x_2 + x_3 + x_4, \ldots) \]

- A moving average process is “smoothed white noise”

- The simplest type of moving average (MA) process is \( X_t = \mu + \epsilon_t + \beta \epsilon_{t-1} \) where \( \epsilon_t \) is zero-mean white noise

- The \( \epsilon_t \)'s are uncorrelated, but the \( X_t \)'s are not:

\[ \begin{array}{c}
X_{t+1} \\
\vdots \\
X_{t+2} \\
X_t \\
\end{array} \]

\[ \begin{array}{c}
\epsilon_{t+1} \\
\epsilon_{t+2} \\
\epsilon_{t+1} \\
\epsilon_t \\
\end{array} \]

- For MA(1) we prove that: \( \rho_2 = \text{corr}(X_t, X_{t+2}) = 0 \)

- Similarly, \( \rho_k = 0 \), for \( k > 2 \)

- Moving average process of order (q) (MA(q)):

\[ X_t = \mu + \epsilon_t + \beta_1 \epsilon_{t-1} + \ldots + \beta_q \epsilon_{t-q} \]

c. **ARMA models**

ARMA processes « combine » AR and MA parts:

\[ X_t = \mu + a_1(X_{t-1} - \mu) + \ldots + a_p(X_{t-p} - \mu) + \epsilon_t + \beta_1 \epsilon_{t-1} + \ldots + \beta_q \epsilon_{t-q} \]

Note: ARMA\((p,0) = AR(p)\)
ARMA\((0,q) = MA(q)\)

### II.4 Backwards Shift Operator and Difference Operator

The following operators will be useful:

- Backwards shift operator: \( B X_t = X_{t-1}, B \mu = \mu \)
• Difference operator: $\nabla = 1 - B$, hence

\[
\begin{align*}
\nabla X_t &= X_t - X_{t-1} \\
B^2 X_t &= BBX_t = BX_{t-1} = X_{t-2} \\
\nabla^2 X_t &= \nabla X_t - \nabla X_{t-1} \\
&= X_t - X_{t-1} - (X_{t-1} - X_{t-2}) \\
&= (1 - B)^2 X_t \\
&= (1 - 2B + B^2) X_t \\
&= X_t - 2X_{t-1} + X_{t-2}
\end{align*}
\]

II.5 AR(p) models, stationarity and the Yule-Walker equations

a. The AR(1) Model

• Recall $X_t = \mu + \alpha (X_{t-1} - \mu) + \varepsilon_t$
• Substituting in for $X_{t-1}$, then for $X_{t-2}$,

\[
\begin{align*}
X_t &= \mu + \alpha [\alpha (X_{t-2} - \mu) + \varepsilon_{t-1}] + \varepsilon_t \\
&= \mu + \alpha^2 (X_{t-2} - \mu) + \varepsilon_t + \alpha \varepsilon_{t-1} \\
X_t &= \mu + \alpha^i (X_0 - \mu) + \varepsilon_t + \alpha \varepsilon_{t-1} + \ldots + \alpha^{i-1} \varepsilon_1 = \mu + \alpha^i (X_0 - \mu) + \sum_{j=0}^{i-1} \alpha^j \varepsilon_{t-j}
\end{align*}
\]

• Note: $X_0$ is a Random Variable
• Since $E(\varepsilon_t) = 0$ for any $t$, $\mu_t = E(X_t) = \mu + \alpha^i (\mu_0 - \mu)$
• Since the $\varepsilon_i$’s are uncorrelated with each other and with $X_0$,

\[
\begin{align*}
\text{Var}(X_t) &= \text{Var} \left( \mu + \alpha^i (X_0 - \mu) + \sum_{j=0}^{i-1} \alpha^j \varepsilon_{t-j} \right) \\
&= \alpha^{2i} \text{Var}(X_0) + \sum_{j=0}^{i-1} \alpha^{2j} \sigma^2 \\
&= \alpha^{2i} \text{Var}(X_0) + \sigma^2 \frac{1 - \alpha^{2i}}{1 - \alpha^2}
\end{align*}
\]

Question: When will AR(1) process be stationary?

Answer: This will require constant mean and variance.

If $\mu_0 = \mu$ then $\mu_t = \mu + \alpha \ (\mu_0 - \mu) = \mu$.

If $\text{Var}(X_0) = \frac{\sigma^2}{1 - \alpha^2}$ then $\text{Var}(X_t) = \sigma^2 \frac{1 - \alpha^{2i}}{1 - \alpha^2} + \alpha^2 \frac{\sigma^2}{1 - \alpha^2} = \frac{\sigma^2}{1 - \alpha^2}$
Neither \( \mu_t \) nor \( \text{Var}(X_t) \) depend on \( t \). We also require that \( |\alpha| < 1 \) so that the AR(1) process be stationary, in which case

\[
\mu_t = \mu + \alpha(\mu_0 - \mu) \quad \text{AND} \quad \text{Var}(X_t) = \sigma^2 \left( \frac{1}{1 - \alpha^2} \right) = \alpha^2 \left( \text{Var}(X_0) - \sigma^2 \frac{1}{1 - \alpha^2} \right)
\]

- If \( |\alpha| < 1 \), both terms will decay away to zero for large \( t \)
  \[ \Rightarrow \text{X is almost stationary for large } t \]

- Equivalently, if we assume that the process has already been running for a very long time, it will be stationary

- Any AR(1) process with infinite history and \( |\alpha| < 1 \) will be stationary:

\[ \epsilon_{-2}, \epsilon_{-1}, \epsilon_0, \epsilon_1, \ldots, \epsilon_t \]
\[ \ldots X_2, X_1, X_0, X_1, \ldots, X_t \]

Steady State reached  Observed time series

- An AR(1) process can be represented as:

\[ X_t = \mu + \sum_{j=0}^{\infty} \alpha^j \epsilon_{t-j} \]

and this converges only if \( |\alpha| < 1 \).

- The AR(1) model \( X_t = \mu + \alpha (X_{t-1} - \mu) + \epsilon_t \) can be written as

\[ (1 - \alpha B)(X_t - \mu) = \epsilon_t \]

If \( |\alpha| < 1 \), then \( 1 - \alpha B \) is invertible and

\[ X_t - \mu = (1 - \alpha B)^{-1} \epsilon_t = (1 + \alpha B + \alpha^2 B^2 + \ldots) \epsilon_t \]
\[ = \epsilon_t + a \epsilon_{t-1} + a^2 \epsilon_{t-2} + \ldots, \]

- So

\[ X_t = \mu + \sum_{j=0}^{\infty} \alpha^j \epsilon_{t-j} \]

From this representation,

\[ \mu_t = E(X_t) = \mu \] and \( \text{Var}(X_t) = \sum_{j=0}^{\infty} \alpha^j \sigma^2 = \frac{\sigma^2}{1 - \alpha^2} \] if \( |\alpha| < 1 \).
• So, if $|\alpha| < 1$, the mean and variance are **constant**, as required for stationarity

• We must calculate the **autocovariance** $\gamma_k = \text{Cov}(X_t, X_{t+k})$ and show that this depends **only** on the lag $k$. We need properties of covariance:

$$\text{Cov}(X+Y, W) = \text{Cov}(X, W) + \text{Cov}(Y, W)$$

$$\text{Cov}(X, e) = 0$$

From the following diagram

```
... ε_{t-2}, ε_{t-1}, ε_t (uncorrelated) and ... ε_{t-2}, ε_{t-1}, ε_t
X_t
X_{t-1}
```

we can tell that $\epsilon_t$ and $X_{t-1}$ are uncorrelated, hence

$$\text{Cov}(\epsilon_t, X_{t-1}) = 0$$

$$\text{Cov}(\epsilon_t, X_{t-k}) = 0, \quad k \geq 1$$

$$\text{Cov}(\epsilon_t, X_t) = \sigma^2$$

$$\gamma_1 = \text{Cov}(X_t, X_{t-1}) = \text{Cov}(\mu + \alpha(X_{t-1} - \mu) + \epsilon_t, X_{t-1})$$

$$= \alpha \text{Cov}(X_{t-1}, X_{t-1}) + \text{Cov}(\epsilon_t, X_{t-1})$$

$$= \alpha \gamma_0 + 0$$

$$\gamma_2 = \text{Cov}(X_t, X_{t-2}) = \text{Cov}(\mu + \alpha(X_{t-1} - \mu) + \epsilon_t, X_{t-2})$$

$$= \alpha \text{Cov}(X_{t-1}, X_{t-2}) + \text{Cov}(\epsilon_t, X_{t-2})$$

$$= \alpha \gamma_1 + 0$$

$$= \alpha^2 \gamma_0$$

Similarly,

$$\gamma_k = \alpha^k \gamma_0, \quad k \geq 0$$

In general,

$$\gamma_k = \text{Cov}(X_t, X_{t+k}) = \text{Cov}(\mu + \alpha(X_{t-1} - \mu) + \epsilon_t, X_{t+k})$$

$$= \alpha \text{Cov}(X_{t-1}, X_{t+k}) + \text{Cov}(\epsilon_t, X_{t+k})$$

$$= \alpha \gamma_{k-1} + 0$$

Hence,

$$\gamma_k = \alpha^k \gamma_0 = \alpha^k \frac{\sigma^2}{1-\alpha^2}, \quad \text{for } k \geq 0$$

and

$$\rho_k = \gamma_k / \gamma_0 = \alpha^k \quad \text{for } k \geq 0$$

$\Rightarrow$ **ACF decreases geometrically with $k$**
Recall the partial autocorrelations $\phi_1$ and $\phi_2$ satisfy

$$\phi_1 = \rho_1 \quad \text{and} \quad \phi_2 = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$$

Here $\phi_1 = \rho_1 = \alpha$ and

$$\phi_2 = \frac{\alpha^2 - \alpha^2}{1 - \alpha^2} = 0$$

In fact, $\phi_k = 0$ for $k > 1$

In summary, for the AR(1) model,

- ACF “tails off” to zero
- PACF “cuts off” after lag 1

Example: Consumer price index $Q_t$

$$r_t = \ln(Q_t/Q_{t-1})$$ models the force of inflation

Assume $r_t$ is an AR(1) process:

$$r_t = \mu + \alpha(r_{t-1} - \mu) + e_t$$

Note: Here $\mu$ is the long-run mean

$$r_t - \mu = \alpha(r_{t-1} - \mu), \text{ ignoring } e_t$$

If $|\alpha| < 1$, then $r_t - \mu \to 0$ and so $r_t \to \mu$ as $t \to \infty$. In this case $r_t$ is said to be mean-reverting.

b. The AR(P) model and stationarity

Recall that the AR(p) model can be written either in its generic form

$$X_t = \mu + \alpha_1(X_{t-1} - \mu) + \alpha_2(X_{t-2} - \mu) + \ldots + \alpha_p(X_{t-p} - \mu) + e_t$$

or using the $B$ operator as

$$(1 - \alpha_1 B - \alpha_2 B^2 - \alpha_3 \ldots - \alpha_p B^p) (X_t - \mu) = e_t$$

Result: AR(p) is stationary IFF the roots of the characteristic equation

$$1 - \alpha_1 z - \alpha_2 z^2 - \ldots - \alpha_p z^p = 0$$

are all greater than 1 in absolute value.
1 - \alpha_1z - \alpha_2z^2 - ... - \alpha_pz^p \iff \text{Characteristic Polynomial}

Explanation for this result: write the AR(p) process in the form

\[ \left( \frac{1 - B}{z_1} \right) \left( \frac{1 - B}{z_2} \right) ... \left( \frac{1 - B}{z_p} \right) (X_t - \mu) = e_i \]

where \( z_1 ... z_p \) are roots of the characteristic polynomial:

\[ 1 - \alpha_1z ... - \alpha_pz^p = \left( 1 - \frac{z}{z_1} \right) \left( 1 - \frac{z}{z_2} \right) ... \left( 1 - \frac{z}{z_p} \right) \]

In the AR(1) case,

\[ 1 - \alpha z = 1 - z/z_1, \text{ where } z_1 = 1/\alpha \]

In AR(1) case, we can invert the term

\[ \left( \frac{1 - B}{z_1} \right) \]

in

\[ \left( \frac{1 - B}{z_1} \right) (X_t - \mu) = e_i \]

IFF \(|z_1| > 1\). In the AR(p) case, we need to be able to invert all of the factors

\[ \left( \frac{1 - B}{z_i} \right) \]

This will be the case IFF \(|z_i| > 1\) for \( i = 1, 2, ..., p \).

Example: AR(2)

\[ X_t = 5 - 2(X_{t-1} - 5) + 3(X_{t-2} - 5) + e_t \quad \text{or} \quad (1 + 2B - 3B^2)(X_t - 5) = e_t \]

\[ 1 + 2z - 3z^2 = 0 \] is the characteristic equation here

Question: when is an AR(1) process stationary ?

Answer: we have

\[ X_t = \mu + \alpha (X_{t-1} - \mu) + e_t. \]

i.e. \((1 - \alpha B)(X_t - \mu) = e_t\), so \(1 - \alpha z = 0\) is the characteristic equation with solution \( z = 1/\alpha \). So \(|\alpha| < 1\) is equivalent to \(|z| > 1\), as required.

Question: Consider the AR(2) process \( X_n = X_{n-1} - \frac{1}{2} X_{n-2} + e_n \). Is it stationary ?
Answer: Use B-operator: 

\[(1 - B + \frac{1}{2} B^2)X_n = e_n\]  

So characteristic equation is 

\[1 - z + \frac{1}{2} z^2 = 0, \text{ with roots } 1 \pm i \text{ and } |1 \pm i| = \sqrt{2} > 1\]  

Since both roots satisfy \(|z| > 1\), the process is stationary.

In the AR(1) model, we had \(\gamma_1 = \alpha \gamma_0\) and \(\gamma_0 = \sigma^2\). These are a particular case of the Yule-Walker Equations for AR(p):

\[
\text{Cov}(X_t, X_{t-k}) = \text{Cov}(\mu + \alpha_1 (X_{t-1} - \mu) + ... + \alpha_p (X_{t-p} - \mu) + e_t, X_{t-k})
\]

\[= \alpha_1 \text{Cov}(X_{t-1}, X_{t-k}) + ... + \alpha_p \text{Cov}(X_{t-p}, X_{t-k}) + \begin{cases} \sigma^2, & \text{if } k=0 \\ 0, & \text{otherwise} \end{cases}\]

c. Yule-Walker equations

The Yule-Walker equations are defined by the following relationship:

\[
\gamma_k = \alpha_1 \gamma_{k-1} + \alpha_2 \gamma_{k-2} + ... + \alpha_p \gamma_{k-p} + \begin{cases} \sigma^2, & \text{if } k=0 \\ 0, & \text{otherwise} \end{cases}, \text{ for } 0 \leq k \leq p
\]

Considering the AR(1) (i.e. \(p = 1\)), for \(k = 1\), we get \(\gamma_1 = \alpha \gamma_0\), and for \(k = 0\), we get \(\gamma_0 = \sigma^2\).

Example (\(p=3\)):

\[
\begin{align*}
\gamma_3 & = \alpha_1 \gamma_2 + \alpha_2 \gamma_1 + \alpha_3 \gamma_0 \\
\gamma_2 & = \alpha_1 \gamma_1 + \alpha_2 \gamma_0 + \alpha_3 \gamma_1 \\
\gamma_1 & = \alpha_1 \gamma_0 + \alpha_2 \gamma_1 + \alpha_3 \gamma_2 \\
\gamma_0 & = \alpha_1 \gamma_1 + \alpha_2 \gamma_2 + \alpha_3 \gamma_3 + \sigma^2
\end{align*}
\]

Example: consider the AR(3) model \(X_t = 0.6X_{t-1} + 0.4X_{t-2} - 0.1X_{t-3} + e_t\)

Yule-Walker Equations:

\[
\begin{align*}
\gamma_0 &= 0.6 \gamma_1 + 0.4 \gamma_2 - 0.1 \gamma_3 + \sigma^2 \quad (0) \\
\gamma_1 &= 0.6 \gamma_0 + 0.4 \gamma_1 - 0.1 \gamma_2 \quad (1) \\
\gamma_2 &= 0.6 \gamma_1 + 0.4 \gamma_0 - 0.1 \gamma_1 \quad (2) \\
\gamma_3 &= 0.6 \gamma_2 + 0.4 \gamma_1 - 0.1 \gamma_0 \quad (3)
\end{align*}
\]

From (1), \(\gamma_2 = 6\gamma_0 - 6\gamma_1\)

From (2), \(\gamma_2 = 0.4 \gamma_0 + 0.56 \gamma_1\), hence \(\gamma_1 = \frac{56}{65} \gamma_0\), and hence \(\gamma_2 = \frac{54}{65} \gamma_0\).

From (3), \(\gamma_3 = \frac{483}{650} \gamma_0\)

From (0), \(\sigma^2 = 0.22508 \gamma_0\)

Hence, \(\gamma_0 = 4.4429 \sigma^2\), \(\gamma_1 = 3.8278 \sigma^2\), \(\gamma_2 = 3.6910 \sigma^2\), \(\gamma_3 = 3.3014 \sigma^2\)
and so, since $\rho_k = \gamma_k / \gamma_0$, $\rho_0 = 1$, $\rho_1 = 0.862$, $\rho_2 = 0.831$, $\rho_3 = 0.743$.

It may be shown that for AR(p) models,

- ACF “tails off” to zero,
- PACF “cuts off” after lag p, i.e. $\varphi_k = 0$ for $k > p$

II.6 MA(q) models and invertibility

a. The MA(1) model

The model is given by $X_t = \mu + e_t + \beta e_{t-1}$, where $\mu_t = E(X_t) = \mu$, and

$\gamma_0 = \text{Var}(e_t + \beta e_{t-1}) = (1 + \beta^2)\sigma^2$

$\gamma_1 = \text{Cov}(e_t + \beta e_{t-1}, e_{t-1} + \beta e_{t-2}) = \beta \sigma^2$

$\gamma_k = 0$ for $k > 1$

Hence, the ACF for MA(1) is:

$\rho_0 = 1$

$\rho_1 = \frac{\beta}{1 + \beta^2}$

$\rho_k = 0$ for $k > 1$

Since the mean $E(X_t)$ and covariance $\gamma_k = E(X_t, X_{t-k})$ do not depend on $t$, the MA(1) process is (weakly) stationary - for all values of the parameter $\beta$.

However, we require MA models to be invertible and this imposes conditions on the parameters.

Recall: If $|\alpha| < 1$ then in the AR(1) model

$$(1 - \alpha B)(X_t - \mu) = e_t,$$

$$(1 - \alpha B)$$ is invertible and

$$X_t = \mu - \sum_{j=0}^{\infty} \alpha^j e_{t-j} = \mu + e_t + \alpha e_{t-1} + \alpha^2 e_{t-2} + \ldots$$

i.e. an AR(1) process is MA($\infty$). An MA(1) process can be written as

$$X_t - \mu = (1 + \beta B)e_t$$

or

$$(1 + \beta B)^{-1}(X_t - \mu) = e_t$$

i.e.
$X_t - \mu - \beta(X_{t-1} - \mu) + \beta^2(X_{t-2} - \mu) + ... = e_t$

So an MA(1) process is represented as an AR(\infty) one – but only if $|\beta| < 1$, in which case the MA(1) process is invertible.

**Example:** MA(1) with $\beta = 0.5$ or $\beta = 2$

For both values of $\beta$ we have:

$$\rho_1 = \frac{\beta}{1 + \beta^2} = \frac{0.5}{1 + (0.5)^2} = \frac{2}{1 + 2^2} = 0.4,$$

So both models have the same ACF. However, only the model with $\beta=0.5$ is invertible.

**Question:** Interpretation of invertibility

Consider the MA(1) model $X_n - \mu - \beta e_{n-1}$. We have

$$e_n = X_n - \mu - \beta e_{n-1} = X_n - \mu - \beta(X_{n-1} - \mu - \beta e_{n-2})$$

$$= ...$$

$$= X_n - \mu - \beta(X_{n-1} - \mu) + \beta^2(X_{n-2} - \mu) ... + (-\beta)^{n-1}(X_1 - \mu) + (-\beta)^ne_0$$

As $n$ gets large, the dependence of $e_n$ on $e_0$ will be small if $|\beta| < 1$.

**Note:**

AR(1) is stationary IFF $|\alpha| < 1$.

MA(1) is invertible IFF $|\beta| < 1$.

For an MA(1) process, we have $\rho_k = 0$ for $k > 1$, so for an MA(1) process, the ACF “cuts off” after lag1. It may be shown that PACF “tails off” to zero.

<table>
<thead>
<tr>
<th></th>
<th>AR(1)</th>
<th>MA(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF</td>
<td>Tails off to zero</td>
<td>Cuts off after lag 1</td>
</tr>
<tr>
<td>PACF</td>
<td>Cuts off after lag 1</td>
<td>Tails off to zero</td>
</tr>
</tbody>
</table>

**b. The MA(q) model and invertibility**

An MA(q) process is modeled by $X_t = \mu + e_t + \beta_1e_{t-1} + ... + \beta_qe_{t-q}$, where $\{e_t\}$ is a sequence of uncorrelated realisations. For this model we have $\gamma_k = \text{Cov}(X_t, X_{t+k}) = 0$ for $k > q$.

$$\gamma_k = \text{Cov}(X_t, X_{t+k})$$

$$= E[(e_t + \beta_1e_{t-1} + ... + \beta_qe_{t-q}) (e_{t+k} + \beta_1e_{t+k-1} + ... + \beta_qe_{t+k-q})]$$

$$= \sum_{i=0}^q \sum_{j=0}^q \beta_i \beta_j E(e_i e_{i-j-k})$$

[where $\beta_0 = 1$]
\[ \sigma^2 \sum_{j=k}^{q-k} \beta_{j+k} \beta_j, \quad [\text{since } j = i-k \leq q-k] \]

since the only non-zero terms occur when the subscripts of \(e_{t-i}\) and \(e_{t-j-k}\) match, i.e. when \(i = j+k\), for \(k \leq q\).

In summary, for \(k > q\), \(\gamma_k = 0\):
- For MA(q), ACF cuts off after lag q
- For AR(p), PACF cuts off after lag p

**Question:** ACF of the MA(2) process \(X_t = 1 + e_t - 5e_{t-1} + 6e_{t-2}\).

\[ \gamma_0 = \text{Cov}(1 + e_n - 5e_{n-1} + 6e_{n-2}, 1 + e_n - 5e_{n-1} + 6e_{n-2}) = (1 + 25 + 36) \cdot 1 = 62 \]

If \(E(e_n) = 0\) and \(\text{Var}(e_n) = 1\).

\[ \gamma_1 = \text{Cov} (1 + e_n - 5e_{n-1} + 6e_{n-2}, 1 + e_n - 5e_{n-1} + 6e_{n-2}) = (-5)(1) + (6)(-5) = -35 \]
\[ \gamma_2 = \text{Cov} (1 + e_n - 5e_{n-1} + 6e_{n-2}, 1 + e_n - 5e_{n-1} + 6e_{n-2}) = (6)(1) = 6 \]
\[ \gamma_k = 0, \quad k > 2 \]

**Recall** that an AR(p) process is **stationary** IFF roots \(z\) of the characteristic eq satisfy \(|z| > 1\). For an MA(q) process, we have

\[ X_t - \mu = (1 + \beta_1B + \beta_2B^2 + \ldots + \beta_pB^p) e_t \]

Consider the equation \(1 + \beta_1z + \beta_2z^2 + \ldots + \beta_pz^p = 0\). The MA(q) process is **invertible** IFF all roots \(z\) of this equation satisfy \(|z| > 1\).

In summary:
- If AR(p) **stationary**, then AR(p) = MA(\(\infty\))
- If MA(q) is **invertible**, then MA(q) = AR(\(\infty\))

**Question:** Assess invertibility of the MA(2) process \(X_t = 2 + e_t - 5e_{t-1} + 6e_{t-2}\).

We have \(X_t = 2 + (1-5B + 6B^2)e_t\).

The characteristic equation is \(1 - 5z + 6z^2 = 0\) with roots \((1-2z)(1-3z) = 0\), i.e. roots \(z = \frac{1}{2}\) and \(z = \frac{1}{3}\)

\(\Rightarrow\) **Not** invertible

### II.7 ARMA(p,q) models

Recall that the ARMA(p,q) model can be written either in its generic form
\[ X_t = \mu + \alpha_1 (X_{t-1} - \mu) + \ldots + \alpha_p (X_{t-p} - \mu) + e_t + \beta_1 e_{t-1} + \ldots + \beta_q e_{t-q} \]

or using the B operator:
\[(1 - \alpha_1 B \ldots - \alpha_p B^p) (X_t - \mu) = (1 + \beta_1 B \ldots + \beta_q B^q) e_t \]
i.e. \[\Phi(B)(X_t - \mu) = \Theta(B)e_t\]

where
\[\Phi(\lambda) = 1 - \alpha_1 \lambda - \ldots - \alpha_p \lambda^p\]
\[\Theta(\lambda) = 1 + \beta_1 \lambda + \ldots + \beta_q \lambda^q\]

If \(\Phi(\lambda)\) and \(\Theta(\lambda)\) have factors in common, we simplify the defining relation.

Consider the simple ARMA(1,1) process with \(\beta = -\alpha\), written either
\[X_t = \alpha X_{t-1} + e_t - \alpha e_{t-1}\]

or
\[(1 - \alpha B)X_t = (1 - \alpha B)e_t, \text{ with } |\alpha| < 1\]

Dividing through by \((1 - \alpha B)\), we obtain \(X_t = e_t\). Therefore the process is actually an ARMA(0,0), also called white noise.

We assume that \(\Phi(\lambda)\) and \(\Theta(\lambda)\) have no common factors. Properties of ARMA(p,q) are a mixture of those of AR(p) and those of MA(q).

- Characteristic polynomial of ARMA(p,q) = \(1 - \alpha_1 z \ldots - \alpha_p z^p\) (as for AR(p))
- ARMA(p,q) is stationary IFF all the roots \(z\) of \(1 - \alpha_1 z \ldots - \alpha_p z^p = 0\) satisfy \(|z| > 1\)
- ARMA(p,q) is invertible IFF all the roots \(z\) of \(1 - \beta_1 z \ldots - \beta_q z^q = 0\) satisfy \(|z| > 1\)

**Example:** the ARMA(1,1) process \(X_t = \alpha X_{t-1} + e_t + \beta e_{t-1}\) is stationary if \(|\alpha| < 1\) and invertible if \(|\beta| < 1\).

**Example:** ACF of ARMA(1,1). For the model given by \(X_t = \alpha X_{t-1} + e_t + \beta e_{t-1}\) we have
\[
\text{Cov}(e_t, X_{t-1}) = 0
\]
\[
\text{Cov}(e_t, e_{t-1}) = 0
\]
\[
\text{Cov}(e_t, X_t) = \alpha \text{Cov}(e_t, X_{t-1}) + \text{Cov}(e_t, e_t) + \beta \text{Cov}(e_t, e_{t-1}) = \sigma^2
\]
\[
\text{Cov}(e_{t-1}, X_t) = \alpha \text{Cov}(e_{t-1}, X_{t-1}) + \text{Cov}(e_{t-1}, e_t) + \beta \text{Cov}(e_{t-1}, e_{t-1})
\]
\[
= \alpha \sigma^2 + 0 + \beta \sigma^2 = (\alpha + \beta) \sigma^2
\]
\[ \gamma_0 = \text{Cov}(X_t, X_t) = \alpha \text{Cov}(X_t, X_{t-1}) + \beta \text{Cov}(X_t, e_{t-1}) \]
\[ = \alpha \gamma_1 + \sigma^2 + \beta (\alpha + \beta) \sigma^2 \]
\[ = \alpha \gamma_1 + (1 + \alpha \beta + \beta^2) \sigma^2 \]

\[ \gamma_1 = \text{Cov}(X_{t-1}, X_t) \]
\[ = \alpha \text{Cov}(X_{t-1}, X_{t-1}) + \text{Cov}(X_{t-1}, e_t) + \beta \text{Cov}(X_{t-1}, e_{t-1}) \]
\[ = \alpha \gamma_0 + \beta \sigma^2 \]

For \( k > 1 \),

\[ \gamma_k = \text{Cov}(X_{t-1}, X_t) \]
\[ = \alpha \text{Cov}(X_{t-k}, X_{t-1}) + \text{Cov}(X_{t-k}, e_t) + \beta \text{Cov}(X_{t-k}, e_{t-1}) \]
\[ = \alpha \gamma_{k-1} \]

(Appalogues of Yule-Walker Equations)

\[ \Rightarrow \] Solve for \( \gamma_0 \) and \( \gamma_1 \):

\[ \gamma_0 = \frac{1 + 2 \alpha \beta + \beta^2}{1 - \alpha^2} \sigma^2 \]

\[ \gamma_1 = \frac{(\alpha + \beta)(1 + \alpha \beta)}{1 - \alpha^2} \sigma^2 \]

\[ \gamma_k = \alpha^{k-1} \gamma_1, \text{ for } k > 1 \]

Hence \[ \frac{\gamma_1}{\gamma_0} = \frac{(1 + \alpha \beta)(\alpha + \beta)}{1 + 2 \alpha \beta + \beta^2}, \rho_k = \alpha^{k-1} \rho_1, \text{ for } k > 1 \] (compare \( \rho_k = \alpha^k \), for \( k \geq 0 \) for AR(1)).

For (stationary) ARMA(p,q),

- ACF tails off to zero
- PACF tails off to zero

**Question:** ARMA(2,2) process

\[ 12 X_t = 10X_{t-1} - 2X_{t-2} + 12e_t - 11e_{t-1} + 2e_{t-2} \]
(12 − 10B + 2B²)X_i = (12 − 11B + 2B²)e_i

The roots of
12 − 10z + 2z² = 2(z − 2)(z − 3) = 0

Are z = 2 and z = 3, |z| > 1 for both roots; process stationary.

II.8 ARIMA(p,d,q) models

a. Non-ARMA processes

• Given time series data X_1 ... X_n, find a model for this data.
• Calculate sample statistics: sample mean, sample ACF, sample PACF.
• Compare with known ACF/PACF of class of ARMA models to select suitable model.
• All ARMA models considered are stationary – so can only be used for stationary time series data.
• If time-series data is non-stationary, transform it to a stationary time series (e.g. by differencing)
• Model this transformed series using an ARMA model
• Take the “inverse transform” of this model as model for the original non-stationary time series.

Example: Random Walk X_0 = 0, X_n = X_{n-1} + Z_n, where Z_n is a white noise process.
X_n is non-stationary, but X_n = X_n − X_{n-1} = Z_n is stationary.

Question: Given X_0, X_1 ... X_n the first order differences are w_i = x_i − x_{i-1}, i = 1, ..., N

From the differences w_1, w_2, ..., w_N and x_0 we can calculate the original time series:

w_1 = x_1 − x_0 , so x_1 = x_0 + w_1
w_2 = x_2 − x_1 , so x_2 = x_1 + w_2

= x_0 + w_1 + w_2, etc.

The inverse process of differencing is integration, since we must sum the differences to obtain the original time series.

b. The I(d) notation ("integrated of order d")

• X is said to be I(0) if X is stationary
• X is said to be I(1) if X is not stationary but \( Y_t = X_t - X_{t-1} \) is stationary

• X is said to be I(2) if X is not stationary, but Y is I(1).

Thus X is I(\(d\)) if X must be “differenced” \(d\) times to make it stationary.

\textit{Example:} If the first differences \( x_n = x_n - x_{n-1} \) of \( x_1, x_2 \ldots x_n \) are modelled by an AR(1) model (stationary)

\[
\nabla X_n = 0.5 \nabla X_{n-1} + e_n,
\]

Then, \( X_n - X_{n-1} = 0.5(X_{n-1} - X_{n-2}) + e_n \), so \( X_n = 1.5X_{n-1} - 0.5X_{n-2} + e_n \) is the model for the original time series.

This AR(2) model is non-stationary since written as \((1 - 1.5B + 0.5B^2)X_n = e_n\), for which the characteristic equation is:

\[
1 - 1.5z + 0.5z^2 = 0
\]

with roots \( z = 1 \) and \( z = 2 \). The model is non-stationary since \(|z| > 1\) does not hold for BOTH roots.

X is ARIMA(p,1,q) if X is non-stationary, but \( \nabla X \) (the first difference of X) is a stationary ARMA(p,q) process

• \textbf{Recall} that a process X is I(1) if X is non-stationary, but \( \nabla X = X_t - X_{t-1} \) is stationary

\textbf{Note:} If \( X_t \) is ARIMA(p,1,q) then \( X_t \) is I(1).

\textit{Example: Random Walk}. \( X_t - X_{t-1} = e_t \), where \( e_t \) is a white noise process.

We have

\[
X_t = X_0 + \sum_{j=1}^{t} e_j
\]

So \( E(X_t) = E(X_0) \), if \( E(e_t) = 0 \), but \( \text{Var}(X_t) = \text{Var}(X_0) + t\sigma^2 \). Hence \( X_t \) is non-stationary, but \( \nabla X_t = e_t \), where \( e_t \) is a stationary white noise process.

\textit{Example:} \( Z_t = \text{closing share price on day } t \). Here the model is given by

\[
Z_t = Z_{t-1} \exp(\mu + e_t)
\]

Let \( Y_t = \ln Z_t \), then \( Y_t = \mu + Y_{t-1} + e_t \). This is a \textit{random walk with drift}.

Now consider the daily returns \( Y_t - Y_{t-1} = \ln(Z_t/Z_{t-1}) \). Since \( Y_t - Y_{t-1} = \mu + e_t \) and the \( e_t \)'s are independent, then \( Y_t - Y_{t-1} \) is independent of \( Y_1 \ldots Y_{t-1} \) or \( \ln(Z_t/Z_{t-1}) \) is independent of past prices \( Z_0, Z_1, \ldots Z_{t-1} \).
Example: Recall the example of $Q_t =$ consumer price index at time $t$. We have

$$ r_t = \ln(Q_t/Q_{t-1}) \text{ follows AR(1) model} $$

$$ r_t = \mu + \alpha (r_{t-1} - \mu) + \epsilon_t $$

$$ \ln(Q_t/Q_{t-1}) = \mu + \alpha (\ln(Q_t/Q_{t-1}) - \mu) + \epsilon_t $$

$$ \nabla \ln(Q_t) = \mu + \alpha (\nabla \ln(Q_{t-1}) - \mu) + \epsilon_t $$

thus $\nabla \ln(Q_t)$ is AR(1) and so $\ln(Q_t)$ is ARIMA(1,1,0)

If

• $X$ needs to be differenced at least $d$ times to reduce it to stationarity,
• and $Y = \nabla^d X$ is stationary ARMA(p,q),

then

$X$ is an ARIMA(p,d,q) process.

An ARIMA(p,d,q) process is I(d)

Example: Identify as ARIMA(p,d,q) the following model

$$ X_t = 0.6X_{t-1} + 0.3X_{t-2} + 0.1X_{t-3} + \epsilon_t - 0.25\epsilon_t $$

$$(1 - 0.6B - 0.3B^2 - 0.1B^3) X_t = (1 - 0.25B) \epsilon_t $$

Check for factor $(1 - B)$ on LHS: $$(1 - B)(1 - 0.4B + 0.1B^2)X_t = (1 - 0.25B) \epsilon_t $$

$\Rightarrow$ Model is ARIMA(2,1,1)

Characteristic equation: $1 + 0.4z + 0.1z^2 = 0$ with roots $-2 \pm i \sqrt{6}$

Since $|z| = \sqrt{10} > 1$ for both roots $\nabla X_t$ is stationary, as required.

Alternative method: Write model in terms of $\nabla X_t = X_t - X_{t-1}$, $\nabla X_{t-1}$, etc

$$ X_t - X_{t-1} = -0.4X_{t-1} + 0.4X_{t-2} $$

$$ = -0.1X_{t-2} + 0.1X_{t-3} + \epsilon_t - 0.25\epsilon_t $$

$$ \nabla X_t = -0.4 \nabla X_{t-1} - 0.1 \nabla X_{t-2} + \epsilon_t - 0.25\epsilon_{t-1} $$

Hence, $\nabla X_t$ is ARMA(2,1) (check for stationarity as above), and so $X_t$ is ARIMA(2,1,1)

Note: if $\nabla^d X_t$ is ARMA(1,q), to check for stationarity, we only need to see that $|\alpha_1| < 1$. 
II.9 The Markov Property

AR(1) Model:

\[ X_t = \mu + \alpha (X_{t-1} - \mu) + e_t \]

Conditional distribution of \( X_{n+1} \), given \( X_n, X_{n-1}, \ldots, X_0 \) depends only on \( X_n \)

\[ \Rightarrow \text{ AR(1) has markov property} \]

AR(2) Model:

\[ X_t = \mu + \alpha_1 (X_{t-1} - \mu) + \alpha_2 (X_{t-2} - \mu) + e_t \]

Conditional distribution of \( X_{n+1} \), given \( X_n, X_{n-1}, \ldots, X_0 \) depends on \( X_{n-1} \) as well as \( X_n \).

\[ \Rightarrow \text{ AR(2) does not have the Markov Property} \]

Consider now

\[ X_{n+1} = \mu + \alpha_1 X_n + \alpha_2 X_{n-1} + e_{n+1} \]

or

\[
\begin{pmatrix}
X_{n+1} \\
X_n \\
\end{pmatrix} =
\begin{pmatrix}
\mu & \alpha_1 & \alpha_2 \\
0 & 1 & 0 \\
\end{pmatrix}
\begin{pmatrix}
X_n \\
X_{n-1} \\
e_{n+1}
\end{pmatrix}
\]

Define \( Y_n = \begin{pmatrix} X_n \\ X_{n-1} \end{pmatrix} = (X_n, X_{n-1})^T \)

then \( Y_{n+1} = \begin{pmatrix}
\mu \\
0 \\
\end{pmatrix} + \begin{pmatrix}
\alpha_1 & \alpha_2 \\
1 & 0 \\
\end{pmatrix} Y_n + \begin{pmatrix}
e_{n+1} \\
0 \\
\end{pmatrix} \)

- \( Y \) is said to be a vector autoregressive process of order 1.
- **Notation:** \( \text{Var}(1) \)

- \( Y \) has the Markov property
- In general, \( \text{AR}(P) \) does not have the Markov property for \( p > 1 \), but \( Y = (X_t, X_{t-1}, \ldots, X_{t-p+1})^T \) does
- **Recall:** Random walk – \( \text{ARIMA}(0,1,0) \) defined by \( X_t - X_{t-1} = e_t \) has independent increments and hence does have the Markov property

It may be shown that for \( p+d > 1 \), \( \text{ARIMA}(p,d,0) \) does not have the Markov property, but \( Y_t = (X_t, X_{t-1}, \ldots, X_{t-p-d+1})^T \) does.
Consider the **MA(1)** process \( X_t = \mu + e_t + \beta e_{t-1} \). It is clear that “knowing \( X_n \) will never be enough to deduce the value of \( e_n \), on which the distribution of \( X_{n+1} \) depends”. Hence an MA(1) process **does not** have the Markov property.

Now consider an **MA(q) = AR(\infty) process**. It is known that AR(p) processes \( Y = (X_t, X_{t-1}, ..., X_{t-p+1})^T \) have the Markov property if considered as a p-dimensional vector process (p finite). It follows that an MA(q) process has no **finite** dimensional Markov representation.

**Question:** Associate a vector-valued Markov process with \( 2X_t = 5X_{t-1} - 4X_{t-2} + X_{t-3} + e_t \)

We have

\[
2 (X_t - X_{t-1}) = 3 (X_{t-1} - X_{t-2}) - (X_{t-2} - X_{t-3}) + e_t
\]

\[
2\nabla X_t = 3 \nabla X_{t-1} - \nabla X_{t-2} + e_t
\]

\[
\nabla^2 X_t = \nabla^2 X_{t-1} + e_t
\]

\( \Rightarrow \) **ARIMA(1,2,0) or ARIMA(p,d,q)** with \( p = 1 \) and \( d = 2 \).

Since \( p+d = 3 > 1 \), \( Y_t = (X_t, X_{t-1}, ..., X_{t-p+d+1})^T = (X_t, X_{t-1}, X_{t-2})^T \) is Markov

**Question:** Let the MA(1) process \( X_n = e_n + e_{n-1} \), where

\[
e_n = \begin{cases} 
1 & \text{with probability } \frac{1}{2} \\
-1 & \text{with probability } \frac{1}{2}
\end{cases}
\]

\[
P(X_n = 2 \mid X_{n-1} = 0)
\]

\[
= P(e_n = 1, e_{n-1} = 1 \mid e_{n-1} + e_{n-2} = 0)
\]

\[
= P(e_n = 1) P(e_{n-1} = 1 \mid e_{n-1} + e_{n-2} = 0)
\]

\[
= \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}
\]

\[
P(X_n = 2 \mid X_{n-1} = 0, X_{n-2} = 2)
\]

\[
= P(e_n = 1, e_{n-1} = 1 \mid e_{n-1} + e_{n-2} = 0, e_{n-2} + e_{n-3} = 2)
\]

\[
= 0
\]

\( \Rightarrow \) **Not Markov:** since the two probabilities differ, value of \( X_n \) does not depend on the immediate past \( n-1 \) only.
III. Non-stationarity: trends and techniques

III.1 Typical trends

Possible causes of non-stationarity in a time series are:

- Deterministic trend (e.g. linear or exponential growth)
- Deterministic cycle (e.g. seasonal effects)
- Time series is integrated (as opposed to differenced)

Example:

\[ X_n = X_{n-1} + Z_n \]

where \( Z_n = \begin{cases} +1, & \text{probability 0.6} \\ -1, & \text{probability 0.4} \end{cases} \)

Here \( X_n \) is I(1), since \( Z_n = X_n - X_{n-1} \) is stationary. Also, \( E(X_n) = E(X_{n-1}) + 0.2 \), so the process has a deterministic trend.

Many techniques allow to detect non-stationary series; among the simplest methods:

- Plot of time series against \( t \)
- Sample ACF

The sample ACF is an estimate of the theoretical ACF, based on the sample data and is defined later. A plot of the time series will highlight a trend in the data and will show up any cyclic variation.

Recall: For a stationary time series, \( \rho_k \to 0 \) as \( k \to \infty \), i.e. (theoretical) ACF converges toward zero.

Hence, the sample ACF should also converge toward zero. If the sample ACF decreases slowly, the time series is non-stationary, and needs to be differenced before fitting a model.
If sample ACF exhibits **periodic oscillation**, there is probably a **seasonal pattern** in the data. This should be removed **before** fitting a model (see Figures 7.3a and 7.3b). The following graph (Fig 7.3(a)) shows the number of hotel rooms occupied over several years. Inspection shows the clear seasonal dependence, manifested as a cyclic effect.

The next graph (Fig 7.3(b)) shows the sample autocorrelation function for this data. It is clear that the seasonal effect shows up as a cycle in this function. In particular, the period of this cycle looks to be 12 months, reinforcing the idea that it is a seasonal effect.

**Methods for removing a linear trend:**
• Least squares
• Differencing

Methods for removing a seasonal effect

• Seasonal differencing
• Method of Moving Averages
• Method of seasonal means

III.2 Least squares trend removal

Fit a model,

\[ \hat{X}_t = a + bt + \hat{Y}_t \]

where \( \hat{Y}_t \) is a zero-mean, stationary process.

Recall: \( e_t \) = error variables (“true residuals”) in a regression model. Assume \( e_t \sim \text{IN}(0,\sigma^2) \)

• Estimate parameters \( a \) and \( b \) using linear regression
• Fit a stationary model to the residuals:

\[ \hat{y}_t = x_t - (\hat{a} - \hat{b}t) \]

Note: least squares may also be used to remove nonlinear trends from a time series. It is naturally possible to model any observed nonlinear trend by some term \( \tau(t) \) within

\[ X_t = \tau(t) + Y_t \]

which can be estimated using least squares. For example, a plot of hourly data of daily energy loads against temperature, over a one-daytime frame, may indicate quadratic variations over the day; in this case one could use \( \tau(t) = a + bt^2 \).

III.3 Differencing

a. Differencing and linear trend removal

Use differencing if the sample ACF decreases slowly. If there is a linear trend, e.g. \( x_t = a + bt + y_t \), then

\[ \nabla x_t = x_t - x_{t-1} = b + \nabla y_t, \]

so differencing has removed the linear trend. If \( x_t \) is I(\( d \)), then differencing \( x_t \) \( d \) times will make it stationary.

Differencing \( x_t \) once will remove any linear trend, as above.
Suppose \( x_t \) is I(1) with a linear trend. If we difference \( x_t \) once, then \( \nabla x_t \) is stationary and we have removed the trend.

However, if we remove the trend using linear regression we will still be left with an I(1) process that is non-stationary.

Example:

\[
X_n = X_{n-1} + Z_n, \quad \text{where } Z_n = \begin{cases} 
+1, & \text{prob. 0.6} \\
-1, & \text{prob. 0.4}
\end{cases}
\]

Let \( X_0 = 0 \). Then \( E(X_1) = 0.2 \), since \( E(Z_1) = 0.2 \), and

\[
\begin{align*}
E(X_2) &= 0.2(2) \\
E(X_n) &= 0.2(n).
\end{align*}
\]

Then \( X_n \) is I(1) AND \( X_n \) has a linear trend.

Let \( Y_n = X_n - 0.2(n) \). Then \( E(Y_n) = 0 \), so we have removed the linear trend but

\[
Y_n - Y_{n-1} = X_n - X_{n-1} - 0.2 \\
= Z_n - 0.2
\]

Hence \( Y_n \) is a random walk (which is non-stationary) and \( \nabla Y_n \) is stationary, so \( Y_n \) is an I(1) process.

**b. Selection of \( d \)**

How many times \((d)\) do we have to difference the time series \( X_t \) to convert it to stationarity? This will determine the parameter \( d \) in the fitted ARIMA\((p,d,q)\) model.

Recall the three causes of non-stationarity:

- Trend
- Cycle
- Time series is an integrated series

We are assuming that linear trends and cycles have been removed, so if the plot of the time series and its SACF indicate non-stationarity, it could be that the time series is a realisation of an integrated process and so must be differenced a number of times to achieve stationarity.

**Choosing an appropriate value of \( d \):**

- Look at the SACF. If the SACF decays slowly to zero, this indicates a need for differencing (for a stationary ARMA model, the SACF decays rapidly to zero).

- Look at the sample variance of the original time series \( X \) and its difference.

Let \( \hat{\sigma}^2 \) be the sample variance of \( z^{(d)} = \nabla^d X \). It is normally the case that \( \hat{\sigma}^2 \) first decreases with \( d \) until stationarity is reached, and then starts to increase, since differencing too much introduces correlation.
Take $d$ equal to the value that minimises $\hat{\sigma}^2$.

In the above example, take $d=2$, which is the value for which the estimated variance is minimised.

### III.4 Seasonal differencing

**Example:** Let $X$ be the monthly average temperature in London. Suppose that the model

$$x_t = \mu + \theta_t + y_t$$

applies, where $\theta_t$ is a periodic function with period 12 and $y_t$ is stationary. The **seasonal difference** of $X$ is defined as:

$$\left( \nabla_{12} x \right)_t = x_t - x_{t-12}$$

But: $x_t - x_{t-12} = (\mu + \theta_t + y_t) - (\mu + \theta_{t-12} + y_{t-12}) = y_t - y_{t-12}$ since $\theta_t = \theta_{t-12}$.

Hence $x_t - x_{t-12}$ is a **stationary process**. We can model $x_t - x_{t-12}$ as a stationary process and thus get a model for $x_t$.

**Example:** In the UK, monthly inflation figures are obtained by seasonal differencing of the retail prices index (RPI). If $x_t$ is the value of RPI for month $t$, then annual inflation figure for month $t$ is

$$\frac{x_t - x_{t+12}}{x_{t+12}} \times 100\%$$

**Remark 1:** the number of seasonal differences taken is denoted by $D$. For example, for the seasonal differencing $X_t - X_{t-12} = \nabla_{12} X_t$ we have $D=1$.

**Remark 2:** in practice, for most time series we would need at most $d=1$ and $D=1$.

### III.5 Method of moving averages

This method makes use of a simple **linear filter** to eliminate the effects of periodic variation. If $X$ is a time series with seasonal effects with **even period** $d = 2h$, we define a **smoothed** process $Y$ by

$$y_t = \frac{1}{2h} \left( \frac{1}{2} x_{t-h} + x_{t-h+1} + \ldots + x_{t-1} + x_t + \ldots + x_{t+h-1} + \frac{1}{2} x_{t+h} \right)$$

This ensures that each period makes equal contribution to $y_t$.

**Example with quarterly data:** A yearly period will have $d = 4 = 2h$, so $h = 2$, and
\[ y_t = \frac{1}{4} \left( \frac{1}{2} x_{t-2} + x_{t-1} + x_t + x_{t+1} + \frac{1}{2} x_{t+2} \right) \]

This is a **centred** moving average, since the average is taken symmetrically around the time \( t \). Such an average can only be calculated retrospectively.

For **odd periods** \( d = 2h + 1 \), the end terms \( x_{t-h} \) and \( x_{t+h} \) need not be halved:

\[ y_t = \frac{1}{2h+1} \left( x_{t-h} + x_{t-h+1} + \ldots + x_{t-1} + x_t + \ldots + x_{t+h-1} + x_{t+h} \right) \]

**Example:** with data every 4 months, a **yearly** period will have \( d = 3 = 2h+1 \), so \( h = 1 \) and

\[ y_t = \frac{1}{3} \left( x_{t-1} + x_t + x_{t+1} \right) \]

### III.6 Seasonal means

In fitting the seasonal model

\[ x_t = \mu + \theta_t + y_t \quad \text{with } E(Y_t)=0 \quad \text{(additive model)} \]

to a **monthly** time series, \( x \) extending over 10 years from January 1990, the estimate of \( \mu \) is \( \bar{X} \) (the average over all 120 observations) and the estimate of \( \theta_{\text{January}} \) is

\[ \hat{\theta}_{\text{January}} = \frac{1}{10} (x_1 + x_{13} + \ldots + x_{109}) - \bar{X} \]

the difference between the **average value for January**, and the overall average over all the months.

Recall that \( \theta_t \) is a **periodic** function with period 12 and \( y_t \) is stationary. Thus, \( \theta_t \) contains the **deviation** of the model (from the overall mean \( \mu \)) at time \( t \) due to the **seasonal** effect.

<table>
<thead>
<tr>
<th>Month/Year</th>
<th>1</th>
<th>2</th>
<th>....</th>
<th>10</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>( x_1 )</td>
<td>( x_{13} )</td>
<td>( \ldots )</td>
<td>( x_{109} )</td>
<td>( \hat{\theta}_1 )</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>\ldots</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>\ldots</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>December</td>
<td>( x_{12} )</td>
<td>( x_{24} )</td>
<td>( \ldots )</td>
<td>( x_{120} )</td>
<td>( \hat{\theta}_{12} )</td>
</tr>
</tbody>
</table>

#### overall mean \( \bar{X} \)

### III.7 Filtering, smoothing

Filtering and exponential smoothing techniques are commonly applied to time series in order to “clean” the original series from undesired artifacts. The moving average is an example of a filtering technique. Other filters may be applied depending on the nature of the input series.
Exponential smoothing is another common set of techniques. It is used typically to “simplify” the input time series by dampening its variations so as to retain in priority the underlying dynamics.

III.8 Transformations

**Recall:** In the simple linear model

\[ y_i = \beta_0 + \beta_1 x_i + e_i \]

where \( e_i \sim \text{IN} \left(0, \sigma^2\right) \), we use **regression diagnostic plots** of the residuals, \( \hat{e}_i \), to test the assumptions about the model (e.g. the normality of the error variables \( e_i \) or the constant variance of the error variables \( e_i \)). To test the later assumption we plot the residuals against the fitted values.

If the plot does not appear as above, the data is transformed, and the most common transformation is the **logarithmic transformation**.

Similarly, if after fitting an ARMA model to a time series \( x_t \), a plot of the “residuals” versus the “fitted values” indicates a **dependence**, then we should consider modelling a **transformation** of the time series \( x_t \) and the most common transformation is the **logarithmic Transformation**

\[ Y_t = \ln(X_t) \]
IV. Box-Jenkins methodology

IV.1 Overview

We consider how to fit an ARIMA(p,d,q) model to historical data \( \{x_1, x_2, \ldots x_n\} \). We assume that trends and seasonal effects have been removed from the data.

The methodology developed by Box and Jenkins consists in 3 distinct steps:

- Tentative identification of an ARIMA model
- Estimation of the parameters of the identified model
- Diagnostic checks

If the tentatively identified model passes the diagnostic tests, it can be used for forecasting.

If it does not, the diagnostic tests should indicate how the model should be modified, and a new cycle of

- Identification
- Estimation
- Diagnostic checks

is performed.

IV.2 Model selection

a. Identification of white noise

Recall: in a simple linear regression model, \( y_i = \beta_0 + \beta_1 X_i + e_i, e_i \sim \text{IN}(0,\sigma^2) \), we use regression diagnostic plots of the residuals \( \hat{e} \) to test the goodness of fit of the model, i.e. if the assumptions \( e_i \sim \text{IN}(0,\sigma^2) \) are justified.

The error variables \( e_i \) form a zero-mean white noise process: they are uncorrelated, with common variance \( \sigma^2 \).

Recall: \( \{e_t : t \in \mathbb{E} \} \) is a zero-mean white noise process if

\[
E(e_t) = 0 \quad \forall t \\
\gamma_k = \text{Cov}(e_t, e_{t-k}) = \begin{cases} \sigma^2, & k = 0 \\ 0, & \text{otherwise} \end{cases}
\]

Thus the ACF and PACF of a white noise process (when plotted against \( k \)) look like this:

![ACF and PACF plots](https://example.com/acf_pacf.png)
i.e. apart from $\rho_0 = 1$, we have $\rho_k = 0$ for $k = 1, 2, \ldots$ and $\phi_k = 0$ for $k = 1, 2, \ldots$.

**Question:** how do we test if the residuals from a time series model look like a realisation of a white noise process?

**Answer:** we look at the SACF and SPACF of the residuals. In studying the SACF and SPACF, we realise that even if the original process was white noise, we would not expect $r_k = 0$ for $k = 1, 2, \ldots$ and $\hat{\phi}_k = 0$ for $k = 1, 2, \ldots$ as $r_k$ is only an estimate of $\rho_k$ and $\hat{\phi}_k$ is only an estimate of $\phi_k$.

**Question:** how close to 0 should $r_k$ and $\hat{\phi}_k$ be, if $r_k = 0$ for $k = 1, 2, \ldots$ and $\hat{\phi}_k = 0$ for $k = 1, 2, \ldots$?

**Answer:** If the original model is white noise, $X_t = \mu + e_t$, then for each $k$, the SACF and SPACF satisfy

$$r_k \sim N\left(0, \frac{1}{n}\right) \quad \text{and} \quad \hat{\phi}_k \sim N\left(0, \frac{1}{n}\right)$$

This is true for large samples, i.e. for large values of $n$.

Values of $r_k$ or $\hat{\phi}_k$ outside the range $\left[-\frac{2}{\sqrt{n}}, \frac{2}{\sqrt{n}}\right]$ can be taken as suggesting that a white noise model is inappropriate.

However, these are only approximate 95% confidence intervals. If $\rho_k = 0$, we can be 95% certain that $r_k$ lies between these limits. This means that 1 value in 20 will lie outside these limits even if the white noise model is correct.

Hence a single value of $r_k$ or $\hat{\phi}_k$ outside these limits would not be regarded as significant on its own, but three such values might well be significant.

There is an overall Goodness of Fit test, based on all the $r_k$’s in the SACF, rather than on individual $r_k$’s, called the Portmanautteau test by Ljung and Box. It consists in checking whether the $m$ sample autocorrelation coefficients of the residuals are too large to resemble those of a white noise process (which should all be negligible).

Given residuals from an estimated ARMA($p,q$) model, under the null hypothesis that all values of $r_k = 0$, and the Q-statistic is asymptotically $\chi^2$-distributed with $s = m - p - q$ degrees of freedom, or, if a constant (say $\mu$) is included, $s = m - p - q - 1$ degrees of freedom.

If the white noise model is correct then

$$Q = n(n + 2) \sum_{k=1}^{m} \frac{r_k^2}{n-k} : \chi^2_s \quad \text{for each } s = m - p - q.$$

That is, under the null hypothesis that all values of $r_k = 0$, the Q-statistic given above is asymptotically $\chi^2$-distributed with $m$ degrees of freedom. If the Q-statistic is found to be greater than the 95th percentile of that $\chi^2$ distribution, the null hypothesis is rejected, which means that the alternative hypothesis that “at least one autocorrelation is non-zero” is accepted. Statistical packages print these statistics. For large $n$, the Ljung-Box Q-statistic tends to closely approximate the Box-Pierce statistic:
The overall diagnostic test is therefore performed as follows (for centred realisations):

- Fit ARMA(p,q) model
- Estimate (p+q) parameters
- Test if

\[ Q = n(n + 2) \sum_{k=1}^{m} \frac{r_k^2}{n-k} \sim \chi^2_{m-p-q} \]

**Remark:** the above Ljung-Box Q-statistic was first suggested to improve upon the simpler Box-Pierce test statistic

\[ Q = n \sum_{k=1}^{m} r_k^2 \]

which was found to perform poorly even for moderately large sample sizes.

**b. Identification of MA(q)**

**Recall:** for an MA(q) process, \( \rho_k = 0 \) for all \( k > q \), i.e. the “ACF cuts off after lag \( q \)”.

To test if an MA(q) model is appropriate, we see if \( r_k \) is close to 0 for all \( k > q \). If the data do come from an MA(q) model, then for \( k > q \) (since the first \( q+1 \) coefficients are significant),

\[ r_k \sim N \left( 0, \frac{1}{n} \left( 1 + \sum_{i=1}^{q} \rho_i^2 \right) \right) \]

and 95% of the \( r_k \)’s should lie in the interval

\[ \left[ -1.96 \sqrt{\frac{1}{n} \left( 1 + 2 \sum_{i=1}^{q} \rho_i^2 \right)}, +1.96 \sqrt{\frac{1}{n} \left( 1 + 2 \sum_{i=1}^{q} \rho_i^2 \right)} \right] \]

(note that it is common to use 2 instead of 1.96 in the above formula). We would expect 1 in 20 values to lie outside the interval. In practise, the \( \rho_i \)’s are replaced by \( r_i \)’s. The “confidence limits” on SACF plots are based on this. If \( r_k \) lies outside these limits it is “significantly different from zero” and we conclude that \( \rho_k \neq 0 \). Otherwise, \( r_k \) is not significantly different to zero and we conclude that \( \rho_k = 0 \).

**SACF**

\[
\begin{array}{cccc}
   & \cdots & \cdots & \cdots \\
1 & 2 & k & \\
\cdots & \cdots & \cdots \\
\end{array}
\]
For $q=0$, the limits for $k=1$ are

$$
\left[ \frac{-1.96}{\sqrt{n}}, \frac{1.96}{\sqrt{n}} \right]
$$

as for testing for white noise model. Coefficient $r_1$ is compared with these limits. For $q = 1$, the limits for $k = 2$ are

$$
\left[ -1.96 \left( \frac{1}{n} \right) \left( 1 + 2r_1^2 \right), 1.96 \left( \frac{1}{n} \right) \left( 1 + 2r_1^2 \right) \right]
$$

and $r_2$ is compared with these limits. Again, 2 is often used in place of 1.96.

c. Identification of AR(p)

Recall: for an AR(p) process, we have $\phi_k = 0$ for all $k > p$, i.e. the “PACF cuts off after lag p”.

To test if an AR(p) model is appropriate, we see if the sample estimate of $\phi_k$ is close to 0 for all $k > p$. If the data do come from an AR(p) model, then for $k > p$,

$$
\hat{\phi}_k \sim N \left( 0, \frac{1}{n} \right)
$$

and 95% of the sample estimates should lie in the interval

$$
\left[ -\frac{2}{\sqrt{n}}, \frac{2}{\sqrt{n}} \right]
$$

The “confidence limits” on SPACF plots are based on this: if the sample estimate of $\phi_k$ lies outside these limits, it is “significant”.

Sample PACF of AR(1)
IV.3 Model fitting

a. Fitting an ARMA(p,q) model

We make the following assumptions:

- An appropriate value of $d$ has been found and $\{z_{d+1}, z_{d+2}, ... z_n\}$ is stationary.
- Sample mean $\bar{z} = 0$; if not, subtract $\hat{\mu} = \bar{z}$ from each $z_i$.
- For simplicity, we assume that $d = 0$ (to simplify upper and lower limits of sums).

We look for an ARMA(p,q) model for the data $z$:

- If the SACF appears to cut off after lag $q$, an MA(q) model is indicated (we use the tests of significance described previously).
- If the SPACF appears to cut off after lag $p$, and AR(p) model is indicated.

If neither the SACF nor the SPACF cut off, mixed models must be considered, starting with ARMA(1,1).

b. Parameter estimation: LS and ML

Having identified the values for the parameters $p$ and $q$, we must now estimate the values of the parameters $\alpha_1, \alpha_2, ... \alpha_p$ and $\beta_1, \beta_2, ..., \beta_q$ in the model

$$Z_t = \alpha_1 Z_{t-1} + ... + \alpha_p Z_{t-p} + e_t + \beta_1 e_{t-1} + \beta_q e_{t-q}$$

Least squares (LS) estimation is equivalent to maximum likelihood (ML) estimation if $e_t$ is assumed normally distributed.

Example: in the AR(p) model, $e_t = Z_t - \alpha_1 Z_{t-1} - ... - \alpha_p Z_{t-p}$. The estimators $\hat{\alpha}_1, ..., \hat{\alpha}_p$ are chosen to minimise

$$\sum_{t=p+1}^{n} (z_t - \hat{\alpha}_1 z_{t-1} - ... - \hat{\alpha}_p z_{t-p})^2$$

Once these estimates obtained, the residual at time $t$ is given by

$$\hat{e}_t = z_t - \hat{\alpha}_1 z_{t-1} - ... - \hat{\alpha}_p z_{t-p}$$

For general ARMA models, $\hat{e}_t$ cannot be deduced from the $z_t$. In the MA(1) model for instance,

$$\hat{e}_t = z_t - \hat{\beta}_1 \hat{e}_{t-1}$$

We can solve this iteratively for $\hat{e}_t$ as long as some starting value $\hat{e}_0$ is assumed. For an ARMA($p,q$) model, the list of starting values is $(\hat{e}_0, \hat{e}_1, ..., \hat{e}_{q-1})$. The starting values are estimated recursively by backforecasting:
0. Assume \((\hat{e}_0, \hat{e}_1, ..., \hat{e}_{q-1})\) are all zero
1. Estimate the \(a_i\) and \(\beta_j\)
2. Use forecasting on the time-reversed process \(\{z_n, ..., z_1\}\) to predict values for \((\hat{e}_0, \hat{e}_1, ..., \hat{e}_{q-1})\)
3. Repeat cycle (1)-(2) until the estimates converge.

### c. Parameter estimation: method of moments

- Calculate theoretical ACF or ARMA(p,q): \(\rho_k\)'s will be a function of the \(a'\)s and \(\beta'\)s.
- Set \(\rho_k = r_k\) and solve for the \(a'\)s and \(\beta'\)s. These are the method of moments estimators.

**Example:** you have decided to fit the following MA(1) model
\[
x_n = e_n + \beta e_{n-1}, \quad e_n \sim N(0,1)
\]
You have calculated \(\hat{\gamma}_0 = 1, \hat{\gamma}_1 = -0.25\). Estimate \(\beta\).

We have \(r_1 = \frac{\hat{r}_1}{\hat{\gamma}_0} = -0.25\).

**Recall:** \(\gamma_0 = (1 + \beta^2) \sigma^2 = 1 + \beta^2\) and \(\gamma_1 = \beta \sigma^2 = \beta\) here, from which \(\rho_1 = \frac{\beta}{1 + \beta^2}\).

Setting \(\rho_1 = r_1 = \frac{\beta}{1 + \beta^2} = -0.25\) and solving for \(\beta\) gives \(\beta = -0.268\) or \(\beta = -3.732\).

**Recall:** the MA(1) process is invertible IFF \(|\beta| < 1\). So for \(\beta = -0.268\), the model is invertible. But for \(\beta = -3.732\) the model is **not** invertible.

**Note:** If \(\hat{\gamma}_1 = -0.5\) here, then \(\rho_1 = r_1 = \frac{\beta}{1 + \beta^2} = -0.5\), which gives \((\beta + 1)^2 = 0\), so \(\beta = -1\), and neither estimate gives an invertible model.

Now, let us estimate \(\sigma^2 = \text{Var}(e_i)\).

**Recall** that in the simple linear model \(Y_i = \beta_0 + \beta_1 X_i + e_i, e_i \sim IN(0, \sigma^2)\), \(\sigma^2\) is estimated by
\[
\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2
\]
where \(e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\) is the \(i^{th}\) residual. Here we use
\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=p+1}^n e_i^2
= \frac{1}{n} \sum_{i=p+1}^n (z_i - \hat{\alpha}_0 z_{i-1} - \ldots - \hat{\alpha}_p z_{i-p} - \hat{\beta}_1 \hat{e}_{i-1} - \ldots - \hat{\beta}_q \hat{e}_{i-q})
\]
No matter which estimation method is used this parameter is estimated last, as estimates of the $\alpha$’s and $\beta$’s are required first.

**Note:** In using either Least Squares or Maximum Likelihood Estimation we also find the residuals, $\hat{e}_t$, whereas using the Method of Moments to estimate the $\alpha$’s and $\beta$’s these residuals have to be calculated afterwards.

**Note:** for large $n$, there will be little difference between LS, ML and Method of Moments estimators.

d. Diagnostic checking

Assume we have identified a tentative ARIMA(p,d,q) model and calculated the estimates

$\hat{\mu}, \hat{\sigma}, \hat{\alpha}_1, ..., \hat{\alpha}_p, \hat{\beta}_1, ..., \hat{\beta}_q$.

We must perform **diagnostic checks** based on the residuals. If the ARMA(p,q) model is a good approximation to the underlying time series process, then the residuals $\hat{e}_t$ will form a good approximation to a **white noise process**.

**(I) Tests to see if the residuals are white noise:**

- Study SACF and SPACF of residuals. Do $r_k$ and $\hat{\phi}_k$ lie outside $\left[-\frac{1.96}{\sqrt{n}}, \frac{1.96}{\sqrt{n}}\right]$?
- Portmanteau test of residuals (carried out on the residual SACF):

$$n(n + 2) \sum_{k=1}^{n} \frac{r_k^2}{n-k} \sim \chi^2_{m-s}, \text{ for } s = \text{number of parameters of the model}$$

If the SACF or SPACF of the residuals has too many values outside the interval $\left[-\frac{1.96}{\sqrt{n}}, \frac{1.96}{\sqrt{n}}\right]$ we conclude that the fitted model does not have enough parameters and a **new model with additional parameters** should be fitted.

The Portmanteau test may also be used for this purpose. Other tests are:

- Inspection of the graph of $\{\hat{e}_t\}$
- Counting turning points
- Study the **sample** spectral density function of the residuals

**(II) Inspection of the graph of $\{\hat{e}_t\}$:**

- plot $\hat{e}_t$ against $t$
- plot $\hat{e}_t$ against $z_t$

Any patterns evident in these plots may indicate that the residuals are not a realisation of a set of **independent** (uncorrelated) variables and so the model is inadequate.
(III) Counting Turning Points:

This is a test of independence. Are the residuals a realisation of a set of independent variables? Possible configurations for a turning point are:

- In the diagram above, there exists a turning point for all configurations except (a) and (b). Since four out of the six possible configurations exhibit a turning point, the probability to observe one is \( \frac{4}{6} = \frac{2}{3} \).

If \( y_1, y_2, ..., y_n \) is a sequence of numbers, the sequence has a turning point at time \( k \) if

either
- \( y_{k-1} < y_k \) AND \( y_k > y_{k+1} \)

or
- \( y_{k-1} > y_k \) AND \( y_k < y_{k+1} \)

Result: if \( Y_1, Y_2, ..., Y_N \) is a sequence of independent random variables, then

- the probability of a turning point at time \( k \) is \( \frac{2}{3} \)
- The expected number of turning points is \( \frac{2}{3} (N - 2) \)
- The variance is \( \frac{(16N - 29)}{90} \)


therefore, the number of turning points in a realisation of \( Y_1, Y_2, ..., Y_N \) should lie within the 95% confidence interval:

\[
\left[ \frac{2}{3} (N - 2) - 1.96 \sqrt{\frac{16N - 29}{90}}, \frac{2}{3} (N - 2) + 1.96 \sqrt{\frac{16N - 29}{90}} \right]
\]

Study the sample spectral density function of the residuals:

Recall: the spectral density function on white noise process is \( f(\omega) = \frac{\sigma^2}{2\pi}, -\pi < \omega < \pi \). So the sample spectral density function of the residuals should be roughly constant for a white noise process.