Intro and Examples

- (**Time series and finite dimensional distributions**) A discrete time series is a set of random variables, $\{X_t\}$, indexed by $t \in T \subseteq \mathbb{z}$. For any finite subset of $\{t_1, \dots, t_n\} \subset T$, the finite dimensional distributions are the joint distributions

$$F(x_{t_1}, \dots, x_{t_n}) \coloneqq P(X_{t_1} \leq x_{t_1}, \dots, X_{t_n} \leq x_{t_n})$$

- Example (**An i.i.d sequence**) A sequence of independent identically distributed random variables $\{X_t\}$, such that $X_t \sim F_x$ for all t, form a discrete time series, with

$$F(x_{t_1}, \dots, x_{t_n}) = \prod_{i=1}^{n} F_X(x_{t_i})$$

- Example (**White noise**) A sequence of uncorrelated random variables, $\{X_t\}$, each with $E(X_t) = 0, Var(X_t) = \sigma^2$, for all $t \in T$, is called a white noise process and is denoted by WIN(0, σ^2)
  - For this time series, the finite dimensional distributions are not defined explicitly but the second order moment structure (i.e. all means, variances and covariances) is.

- Example (**Gaussian process**) Let $T = \mathbb{z}$, then $\{X_t\}$ is a discrete Gaussian process if any finite subset $\{X_{t_1}, \dots, X_{t_n}\}$ has a n-dimensional multivariate normal distribution
  - This model is completely determined when the mean and variance-covariance structures are known

- Example (**Random walk**) Let $Z_t, t \in \mathbb{z}$ be an i.i.d. sequence of random variables. The series defined by the following is called a random walk where t = 1, 2, …

$$X_t \coloneqq \sum_{i=1}^{t} Z_i$$

- Example (**MA(1) process**) let $Z_t, t \in \mathbb{z}$ be an i.i.d. sequence of $N(0, \sigma^2)$, or more generally $Z_t \sim WN(0, \sigma^2)$, random variables.
- The series defined by the following for all t, is called a first order moving average process, and denoted by MA(1)

$$X_t \coloneqq Z_t + \theta Z_{t-1}$$

- Example (**AR(1) process**) Let $Z_t, t \in \mathbb{z}$ be an i.i.d. sequence of $N(0, \sigma^2)$, or more generally $Z_t \sim WN(0, \sigma^2)$, random variables.
- The series defined by the following for all t is called a first order autoregressive (AR(1)) process

$$X_t = \phi X_{t-1} + Z_t$$

Stationary process

- (**The auto-covariance function**) if $\{X_t\}$ is a time series with $Var(X_t) < \infty$ for all $t \in T$, then the auto-covariance function is defined by

$$\gamma(r,s) = Cov(X_r, X_s), r, s \in T$$

- (**Stationarity**) The time series $\{X_t\}_{t\in T}$ is said to be stationary if
  - (i) $E(|X_t|^2) < \infty$ for all $t \in T$,
  - (ii) $E(X_t) = \mu$ for all $t \in T$
  - (iii) the auto-covariance function satisfies the following for all $r, s, r+t, s+t \in T$

$$\gamma(r,s) = \gamma(r+t, s+t)$$

- (**Strict Stationarity**) The time series $\{X_t\}_{t\in T}$ is said to be strictly stationary if the finite dimensional vectors $(X_{t_1}, \dots, X_{t_n})$ and $(X_{t_1+h}, \dots, X_{t_n+h})$ have the same joint distributions for all finite subsets of T and all h where the translation is defined

- Example (**White noise process**) the white noise process $WN(0, \sigma^2)$ we defined earlier is stationary since
  - (i) $E(|X_t|^2) = \sigma^2 + 0^2 < \infty$ for all t.
  - (ii) $E(X_t) = 0$ for all t and
  - (iii)

$$\gamma(r,s) = \begin{cases} \sigma^2 & if\ r = s \\ 0 & if\ r \neq s \end{cases}$$

$$= \gamma(r+t, s+t)$$

- Since the finite dimensional distributions are not defined it is not strictly stationary

- Example (**Cauchy example**) A Cauchy distributed random variable X has density function

$$f(x) = \frac{1}{\pi(1+x^2)}$$

- Which does not have a mean or variance. So an i.i.d. sequence $\{X_n\}$ is strictly stationary but not stationary

- Example (**Random walk**) The random walk, $X_t := \sum_{i=1}^{t} Z_i$, for T = {1, 2, …}, was defined earlier. It is clear that the auto-covariance function satisfies

$$\gamma(t,t) = t\sigma^2 \neq (t+h)\sigma^2 = \gamma(t+h, t+h)$$

- Which does not satisfy Condition (iii) for stationary property

- Example (**Predictable process**) Let Z_1, Z_2 be two independent $N(0, \sigma^2)$ random variables. We can define the discrete time series

$$X_t = Z_1 \cos(2\pi t/100) + Z_2 \sin(2\pi t/100)$$

- From the definition of the auto-covariance function we have

$$\gamma(r,s) = \sigma^2\{\cos^2(2\pi r/100) + \sin^2(2\pi s/100)\}$$
$$= \sigma^2 \cos(2\pi(r-s)/100)$$
$$= \gamma(r+t, s+t)$$

- Which satisfies (iii) of stationary property. From this we can easily check that $\{X_t\}$ is stationary

- (**Properties of auto-covariance function**) Assume $\gamma(r,s)$ is the auto-covariance function of a stationary process $\{X_t\}$, then the following statements hold
  - (i) The auto-covariance function can be written as the following for all r

$$\gamma(h) \coloneqq \gamma(r+h, r)$$

  - (ii) $\gamma(0) \geq 0$
  - (iii) $|\gamma(h)| \leq \gamma(0)$
  - (iv) The auto-covariance function is an even function, i.e. $\gamma(h) = \gamma(-h)$
- Proof
  - (i) we have $\gamma(r,s) = \gamma(r+t, s+t)$,
    - Let t = -s, and
    - Let h=r-s,
    - $\gamma(r,s) = \gamma(r-s, 0) = \gamma(h, 0) = \gamma(h) = \gamma(h+r, r)$
  - (ii) $\gamma(0) = \gamma(0+t, 0+t) = Var(X_t) \geq 0$
  - (iii)
    - Let hand side, $\gamma(h) = \gamma(h, 0) = \gamma(h+t, t) = \gamma(r,s) = Cov(X_r, X_s)$
    - Right hand side, $\gamma(0) = Var(X_t) = Var(X_s) = Var(X_r)$
    - Now, $|Cov(X_r, X_s)| \leq \sqrt{Var(X_r)Var(X_s)} \rightarrow |\gamma(h)| \leq \gamma(0)$
  - (iv)
    - $\gamma(-h) = \gamma(-h, 0) = \gamma(-h+h, h) = \gamma(0, h) = \gamma(h)$

- (**Auto-correlation function**) For a stationary process $\{X_t\}$, the auto-correlation function is defined by

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

Estimable but flexible models

- (**Multivariate normal**) Let X be a n-dimensional multivariate normal random variable. Then the density of X is given by

$$\frac{1}{(2\pi)^{n/2}}|\det\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

- Where $\mu \in R^n$ is the mean and Σ is the n*n positive-definite variance covariance matrix of X

- Let's count the number of parameters and the number of observations available in a time series
  - The i.i.d. normal sequence defined previously has 2 parameters μ, σ^2.
    - This in general will not fit commonly observed real data since the independence assumption is too strong
  - The MA(1) process has, in general, three parameters μ, σ^2, θ. Its covariance structure is defined later on
  - The AR(1) process has, in general, three parameters μ, σ^2, ϕ, where here we need to check that the process is not a random walk, since this is not stationary. Its covariance structure is defined later on

Best Linear Predictor

- We already see from Chapter 2 that the best predictor, in terms of mean square error, of a random variable Y given the information in a random variable X is $E(Y|X)$
  - To compute this requires knowledge of the joint distribution and is computationally hard for many examples.
  - An alternative to this is to look at a simpler class of predictors

- (**Best Linear Predictor**) consider two random variables, X and Y, with $E(X) \coloneqq \mu_X, E(Y) \coloneqq \mu_Y$, and all second moments are finite. We want to compute the best linear predictor of Y given X, i.e. find $\hat{Y}(X) \coloneqq a + bX$ which minimizes

$$E\left(\{Y - \hat{Y}(X)\}^2\right)$$

- Minimizing $L(a,b) \coloneqq E(\{Y-(a+bX)\}^2)$ over a, b gives the solution.

$$\frac{dL}{da}(\hat{a},\hat{b}) = 2E(\{Y-(\hat{a}+\hat{b}X)\}) = 2\left(E(Y)-\hat{a}-\hat{b}E(X)\right) = 0$$

$$\hat{a} = E(Y) - \hat{b}E(X)$$

- Substituting into $L(\hat{a},\hat{b})$ gives

$$E\left(\{(Y-\mu_Y)-\hat{b}(X-\mu_X)\}^2\right)$$

$$\frac{dL}{db}(\hat{a},b) = 2E\left(\{(Y-\mu_Y)-\hat{b}(X-\mu_X)\}(X-\mu_X)\right) = 0$$

$$\hat{b} = \frac{Cov(X,Y)}{Var(X)}$$

- Thus the best linear predictor is

$$\hat{Y} = E(Y) + \frac{Cov(X,Y)}{Var(X)}(X - \mu_X)$$

- And the MSE of the predictor is

$$Var(Y)\big(1 - Corr(X,Y)\big)$$

- Note that if X and Y are uncorrelated, the best linear predictor of Y is just $E(Y)$ and its MSE is $Var(Y)$

- Example (**Non-linear prediction**) Suppose $X \sim N(0,1)$ and $Y = X^2 - 1$, then we have $E(X) = E(Y) = 0$ and the best linear predictor is $\hat{Y} = 0$, since $Cov(X,Y) = 0$
  - Of course the best predictor of Y given X is $X^2 - 1$, which has a MSE of 0.
- So the best linear predictor might not be very good

- Example (**Best linear predictor**) Suppose now we wish to predict Y given X_1, X_2. We apply the same method used previously,
  - First write the best linear predictor

$$\hat{Y} = a_0 - a_1 X_2 - a_2 X_1$$

  - Differentiating with respect to a_0 and substituting gives that we want to minimize

$$E\left(\left\{(Y - \mu_Y) - a_1\big(X_2 - \mu_{X_1}\big) - a_2\big(X_1 - \mu_{X_1}\big)\right\}^2\right)$$

- This is minimized by $(\hat{a}_1, \hat{a}_2)$ being the solution to

$$\begin{pmatrix} Var(X_2) & Cov(X_1,X_2) \\ Cov(X_1,X_2) & Var(X_1) \end{pmatrix}\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \begin{pmatrix} Cov(Y,X_2) \\ Cov(Y,X_1) \end{pmatrix}$$

- i.e. when covariance is non-singular the best predictor is

$$\hat{Y} = E(Y) + (X_2 - \mu_{X_2} \quad X_1 - \mu_{X_1})\begin{pmatrix} Var(X_2) & Cov(X_1,X_2) \\ Cov(X_1,X_2) & Var(X_1) \end{pmatrix}^{-1}\begin{pmatrix} Cov(Y,X_2) \\ Cov(Y,X_1) \end{pmatrix}$$

- The methods and results of above examples can be generalized to the following theorem
- Theorem (**Prediction operator**) if $\{X_t\}$ is a stationary time series, with mean μ, and auto-covariance function $\gamma(h)$.
- The best linear predictor of X_{n+h} given the set $X_n, \ldots, X_1$ is

$$Pred(X_{n+h}|X_n, \ldots, X_1) \coloneqq \mu + (a_1, \ldots, a_n)^T\begin{pmatrix} X_n - \mu \\ X_2 - \mu \\ \vdots \\ X_1 - \mu \end{pmatrix}$$

- Where $a \coloneqq (a_1, \ldots, a_n)$ satisfies the equation

$$\Gamma a = \gamma_{(n,h)} \coloneqq \big(\gamma(h), \ldots, \gamma(h+n-1)\big)^T$$

- Where Γ is the n*n matrix with ijth-element, $\Gamma_{ij} = \gamma(|i-j|)$
- The MSE is given by

$$\gamma(0) - a^T\gamma_{(n,h)}$$

- The theorem tells us that as least as far as linear prediction for a stationary process $\{X_t\}$ is concerned, all that is required is knowledge of the mean of the process μ and the auto-correlation function $\gamma(h)$ for a suitable range of h.
  - Often the computation issues are concerned with inverting the n*n matrix Γ

- Example: The following figure shows an example of a optimal linear predictor based on an observed sample with n = 200.

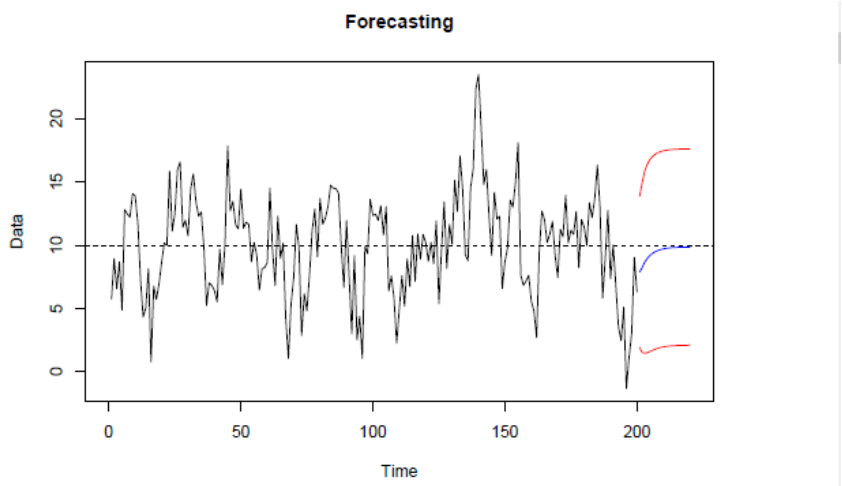


Figure 3.1: Simple forecast example: blue line forecast, red lines ±1.96× standard error. The dashed line is mean of process.

  - In this plot the auto-covariance function was estimated and the best linear predictor for h = 1, …, 20 was computed and plotted in blue.
  - The error associated with the forecast is shown by the 95% prediction interval, shown in red
- For values of h where the auto-correlation is near zero, the best linear predictor is simply the (sample) mean which is shown with the horizontal dashed line.
  - We see that, in this example, the blue line converges to this for roughly h = 20
  - The error in the forecast also converges to a limit for large h. This is just determined by the variance of the sample
- The time between observation and prediction before the best estimate is just the $sample\ mean \pm 1.96 * sample\ standard\ error$ will depend on the shape of the auto-covariance function.
  - If it rapidly decreases to zero then the convergence is fast.
  - The following figure shows an example where there is a slower decay to zero.
    - In this case we see that the convergence to the mean has not happened by h = 20

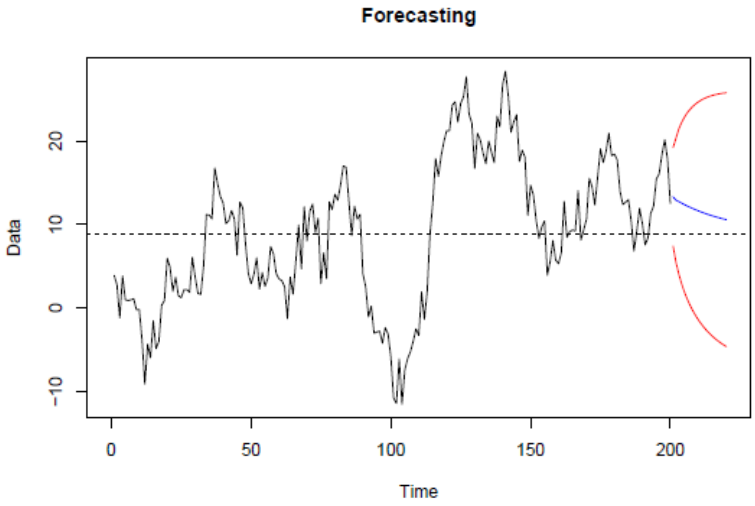


Figure 3.2: Simple forecast example: blue line forecast, red lines ±1.96× standard error. The dashed line is mean of process.

Estimating the mean and auto-covariance functions

- Since the **Prediction operator** theorem shows that for optimal linear h-step forecasting we only need the mean and auto-covariance functions, can these functions be directly estimated from observed data?

- Definition (**Sample moments**) Let $x_1, \ldots, x_n$ be observed values of a stationary time series. The sample mean is defined by: $\bar{x} := \frac{1}{n}\sum_{t=1}^{n} x_t$, the sample auto-covariance function is defined by

$$\hat{\gamma}(h) := \frac{1}{n}\sum_{t=1}^{n-|h|}\left(x_{t+|h|} - \bar{x}\right)(x_t - \bar{x})$$

  - And the sample auto-correlation function is $\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$

- Since the data is not i.i.d. in genera, the usual properties of the sample means do not apply, but they are well behaved statistically

- Theorem (**Properties of sample mean**) For a stationary process $\{X_t\}$ the estimator $\bar{X}_n \coloneqq \frac{1}{n}\sum_{t=1}^{n} X_t$ is an unbiased estimate of $E(X_t) \coloneqq \mu$. Further, its mean squared error is

$$E(\{\bar{X}_n - \mu\}^2) = \frac{1}{n}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma(h)$$

  - If $\gamma(h) \to 0$ as $h \to \infty$, then $\bar{X}_n$ convergences, in mean square, to μ and

$$nE(\{\bar{X}_n - \mu\}^2) \to \sum_{|h|<\infty}\gamma(h)$$

  - If $\sum_{h=-\infty}^{\infty}|\gamma(h)| < \infty$

- Theorem (**Properties of sample auto-correlation function**) For each $k \geq 1$, the k-dimensional sample covariance matrix

$$\hat{\Gamma}_k \coloneqq \begin{bmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \dots & \hat{\gamma}(k-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \dots & \hat{\gamma}(k-2) \\ \vdots & \vdots & \dots & \vdots \\ \hat{\gamma}(k-1) & \hat{\gamma}(k-2) & \dots & \hat{\gamma}(0) \end{bmatrix}$$

  - Is non-negative definite

Computing the auto-covariance function

- Rather than estimating the mean and variance structure purely from the data we can assume a model and compute the auto-covariance function exactly.
  - This gains efficiency when the model is correct but, of course, could be a problem when the model is incorrectly specified.

- Theorem **(MA(1))** Consider a MA(1) – process of the form $X_t \coloneqq Z_t + \theta Z_{t-1}, Z_t \sim (i.i.d)\ N(0, \sigma^2)$, random variables, then we have the following results
  - (i) The auto-covariance function is defined by

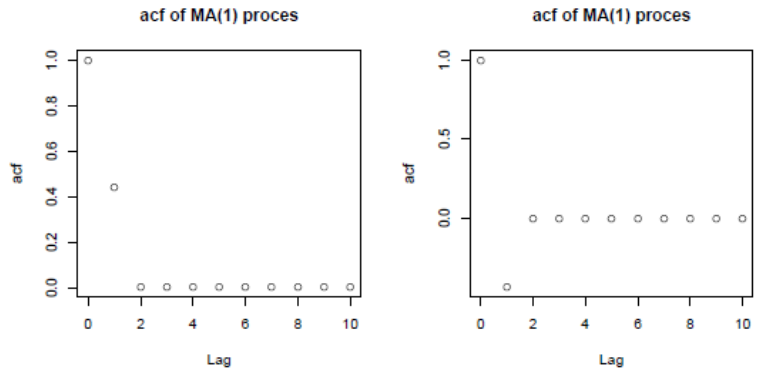
$$\gamma(r,s) = \begin{cases} \sigma^2(1+\theta^2), & for\ r-s = 0 \\ \sigma^2\theta, & for\ |r-s| = 1 \\ 0, & for\ |r-s| > 1 \end{cases}$$

  - (ii) X_t is a stationary process for all values of θ and $\sigma > 0$
- Proof: for r – s = 0

$$\gamma(r,s) = Cov(X_r, X_r)$$
$$= Cov(Z_r + \theta Z_{r-1}, Z_r + \theta Z_{r-1})$$
$$= Cov(Z_r, Z_r) + 2\theta Cov(Z_r, Z_{r-1}) + \theta^2 Cov(Z_{r-1}, Z_{r-1})$$
$$= \sigma^2 + 0 + \sigma^2\theta^2$$

- Example (**MA(1)**) The following figures show an example of the autocorrelation function for

two MA(1) process.



- The left hand panel, $\theta = 0.6$, while in the right panel $\theta = -0.6$
- We see that after h=1, the values of the auto-correlation function are exactly 0

- Theorem (**AR(1)**) Let $\{X_t\}$ be an AR(1) process defined by the following for $Z_t \sim WN(0, \sigma^2)$

$$X_t = \phi X_{t-1} + Z_t$$

- Assuming that the process is stationary, $|\phi| < 1$ and X_t is uncorrelated with Z_{t+h} for h > 0, then the auto-covariance function is given by

$$\gamma(h) = \frac{\sigma^2 \phi^{|h|}}{1 - \phi^2}$$

- Proof: by definition, and for h > 0

$$\gamma(h) = Cov(X_t, X_{t-h})$$
$$= Cov(\phi X_{t-1} + Z_t, X_{t-h})$$
$$= \phi Cov(X_{t-1}, X_{t-h}) + Cov(Z_t, X_{t-h})$$
$$= \phi\gamma(h-1) + 0$$
$$\ldots$$
$$= \phi^h \gamma(0)$$

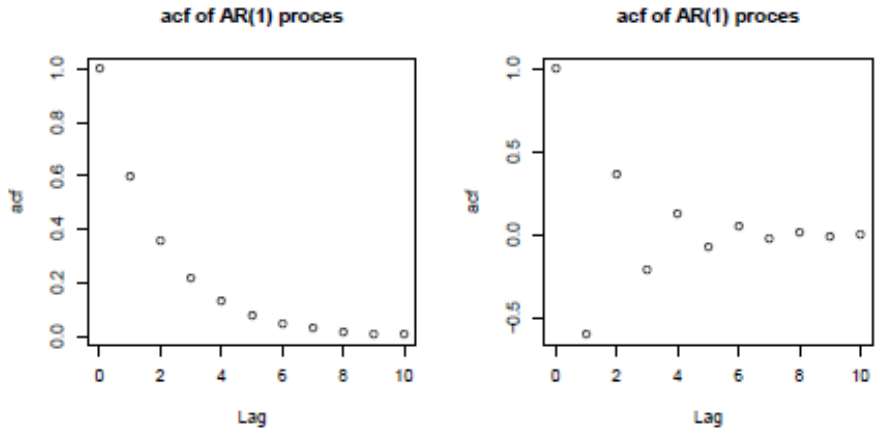
- Note that

$$\gamma(0) = Cov(X_t, X_t) = Cov(\phi X_{t-1} + Z_t, \phi X_{t-1} + Z_t) = \phi^2\gamma(0) + \sigma^2$$
$$\gamma(0) = \frac{\sigma^2}{1 - \phi^2}$$

- Combining the results and using the fact that $\gamma()$ is an even function

$$\gamma(h) = \frac{\sigma^2 \phi^{|h|}}{1 - \phi^2}$$

- Example (**AR(1)**) The following figure shows plots of the auto-correlation function for two AR(1) models.



- The left hand panel has $\phi = 0.6$ and the right hand panel has $\phi = -0.6$
- In both plots we see exponential decay, which is characteristic of the auto-correlation of AR-process. In the right hand plot the values of the correlation alternate between being positive and negative due to the sign of $\phi^{|h|}$
- Note, this never touches 0

MA(q) Process

- Definition (**Backward shift operator**) The *backward shift* operator B acts on a time series $\{X_t\}$ and is defined as

$$BX_t = X_{t-1}$$

- The difference operator ∇ is also define on $\{X_t\}$ via

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

  - Where 1 here represents the identity operator

- We can combine operators as "polynomials", thus

$$B^0X_t = X_t, B^1X_t = X_{t-1}, B^2X_t = X_{t-2}, \dots$$

  - Then define

$$\theta(B) \coloneqq 1 + \theta_1 B + \cdots + \theta_q B^q$$

  - Thus

$$\theta(B)X_t = X_t + \theta_1 X_{t-1} + \cdots + qX_{t-q}$$

- Definition (**MA(q) process**) Let $\{Z_t\}, t \in Z$ be an i.i.d. sequence of $N(0, \sigma^2), or\ Z_t \sim WN(0, \sigma^2)$ random variables, the series defined by

$$X_t \coloneqq \theta(B)Z_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

- Is a q^{th} – order moving average, and denoted by the notation MA(q)

- Theorem (**moments** of MA(q) process)
  - (i) for all t, the mean of an MA(q) process is

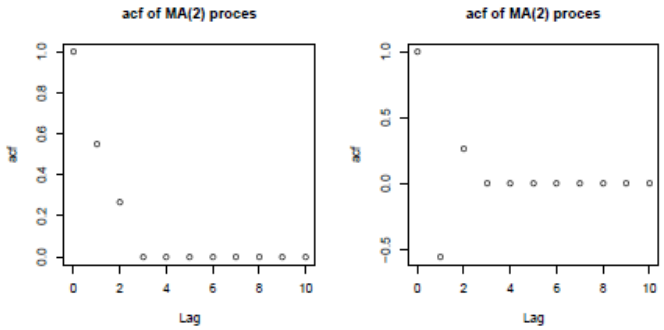
$$E(X_t) = E(\theta(B)Z_t) = 0$$

  - (ii) the auto-covariance of an MA(q) process is given by

$$\gamma(r,s) = \begin{cases} \sigma^2 \sum_{j=0}^{q-|r-s|} \theta_j \theta_{j+|r-s|}, & if\ |r-s| \le q \\ 0, & Otherwise \end{cases}$$

  - (iii) All MA(q) processes are stationary

- Example (**Auto-correlation of MA(2) process**) The plots below show the auto-correlation function $\rho(h)$ for two MA(2) process.



  - The left hand panel shows the case $\theta_1 = 0.6, \theta_2 = 0.4$ and the right hand panel shows the case $\theta_1 = -0.6, \theta_2 - 0.4$
  - Both have the property that all correlations, and hence covariances, are zero after lag 2

AR(p) Process

- Just as polynomials can be extended to infinite series, as long as we are careful about convergence, moving average process can have infinite order.
  - As an example of such an infinite moving average process consider the following argument about an AR(1)

- Example (**Solving the AR(1) equation**) The AR(1) process defined previously was assumed to be stationary when $|\phi| < 1$. The process is defined implicitly as the solution of

$$X_t = \phi X_{t-1} + Z_t$$

  - For $\{Z_t\} \sim WN(0, \sigma^2)$
- We can write it as

$$\begin{aligned} X_t = \phi X_{t-1} + Z_t &= \phi(\phi X_{t-2} + Z_{t-1}) + Z_t \\ &= \phi^2 X_{t-2} + (Z_t + \phi Z_{t-1}) \\ &= \phi^3 X_{t-3} + (Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2}) \end{aligned}$$

...

$$Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \phi^3 Z_{t-3} + \cdots$$

- For this to be MA(∞) process, we need the coefficients to form an absolutely continuous sum. i.e. $\sum_{j=0}^{\infty} |\phi^j| < \infty$, but we have that, when $|\phi| < 1$ the standard result that

$$\sum_{j=0}^{\infty} |\phi|^j = (1 - |\phi|)^{-1} < \infty$$

- Also notice that X_t is only a function of Z_s random variables where $s \leq t$
- Hence, from the properties of $WN(0, \sigma^2)$, we have that X_t is uncorrelated with Z_{t+h} for h > 0, the last regularity condition of our previous theory

- Definition (**The AR(p) process**) Define the polynomial operator

$$\phi(B) \coloneqq 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^P$$

  - Where B is the backward shift operator.
- The AR(p) process is the process which is the stationary solution to the difference equations

$$\phi(B) X_t = Z_t$$

  - For $\{Z_t\} \sim WN(0, \sigma^2)$, when such a solution exists

- Definition (**Causal Process**) A causal process $\{X_t\}$ generated by $\{Z_t\} \sim WN(0, \sigma^2)$ is one where each X_t is only a function of those Z_s where $s \leq t$

- Theorem (**Existence of stationary solution**) there exists a stationary solution to $\phi(B) X_t = Z_t$ when, for $z \in C$, the following polynomial has no roots which lie on the unit circle $\{z \in C | \, |Z| = 1\}$

$$\phi(z) \coloneqq 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^P$$

  - If all roots lie strictly outside the unit circle, we say there is a causal solution which can be written as

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

  - i.e. X_t is a function of "previous" Z_t values

- Example (**Example of AR(p) processes**) Assume that $\{Z_t\} \sim WN(0, \sigma^2)$
  - (i) Let p = 2, and consider

$$X_t - \frac{1}{2} X_{t-1} + \frac{1}{4} X_{t-2} = Z_t$$

    - Look at the roots of the equation $1 - \frac{1}{2} z + \frac{1}{4} z^2 = 0$, since the roots, $1 \pm \sqrt{3} i$ lie outside the units circle there is a stationary solution and it has the causal form
  - (ii) let p = 2, and consider

$$X_t - X_{t-1} + \frac{1}{4}X_{t-2} = Z_t$$

- ✧ The corresponding polynomial, $1 - z + \frac{1}{4}z^2$, has roots, 2, 2 which lie outside the unit circle, so there is a stationary causal solution

➢ (iii) let p = 3, and consider

$$X_t - 5X_{t-1} + 7X_{t-2} - 3X_{t-3} = Z_t$$

- ✧ The corresponding polynomial has roots, 1, 1, 1/3, so there is not a stationary solution to these equations

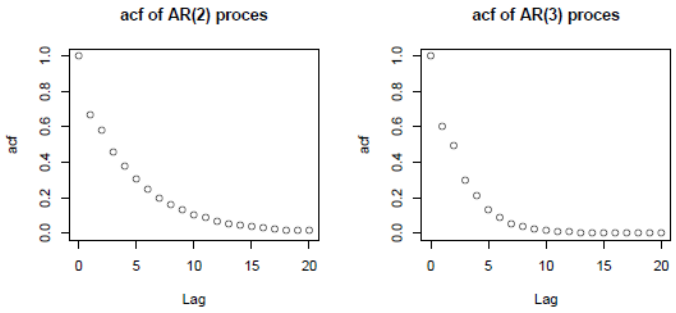
- Theorem (**Auto-covariance function for AR(p) process**) If X_t is a causal, stationary AR(p) process it can be written as

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

- And hence its auto-covariance function can be written as

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}$$

- Example (**AR(p) process**) The following figure shows the auto-correlation function for the two stationary processes.



➢ The left hand panel is an AR(2) process defined by

$$X_t = 0.5X_{t-1} + 0.25X_{t-2} + Z_t$$

➢ And

$$X_t = 0.5X_{t-1} + 0.25X_{t-2} - 0.1X_{t-3} + Z_t$$

- Numerical checks of the polynomials $\phi(z)$ shows all roots lie outside the unit circle, so there exists stationary solutions.

➢ We see both panels show the characteristic exponential decay

The ARMA(p, q) process

- Definition (**The ARMA process**) Let $Z_t \sim WN(0, \sigma^2)$, or i.i.d. $N(0, \sigma^2)$ random variables, and define the polynomial operators

$$\phi(B) \coloneqq 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

- And

$$\theta(B) \coloneqq 1 + \theta_1 B + \cdots + \theta_q B^q$$

  - Where B is the backward shift operator
- The ARMA (p, q) process is the process which is the stationary solution to the difference equations

$$\phi(B) X_t = \theta(B) Z_t$$

  - When such a solution exists, and we assume that the polynomials $\phi(z)$ and $\theta(z)$ share no common factors

- Theorem (**Existence of ARMA process**) The ARMA(p, q) process has a stationary solution if none of the roots of the polynomial $\phi(z)$ lie on the unit circle $|z| = 1$. Further, if the roots lie outside the unit circle (i.e. $|z| > 1$) then there is a causal stationary solution

Likelihood estimation

- Note: we will assume that we have a zero mean stationary process.
  - This is really without loss since if $E(X_t) = \mu$, we estimate μ with the sample mean and work with the process $X_t - \hat{\mu}$

- In order to use likelihood methods for inference we need to make assumptions about the distributions involved in the ARMA(p, q) models.
  - For simplicity, we assume that the innovations $Z_t \sim N(0, \sigma^2)$ i.i.d.
  - We also assume that we have observed $X_1 = x_1, \ldots, X_n = x_n$.
    - For any ARMA(p, q) process the observed (X_1, …, X_n) are a linear function of the unobserved
  - $(Z_1, \ldots, Z_n) \sim N(0_n, \sigma^2 I_{n*n})$
  - Thus the distribution of (X_1, …, X_n) is also normal, and the computational issue is to compute the mean and the variance-covariance function

- Definition (**The likelihood function**) The likelihood function for a mean zero, stationary ARMA (p, q) process is

$$Lik(\phi, \theta, \sigma^2) = \frac{1}{(2\pi)^{n/2} \det(\Gamma_n)^{1/2}} \exp\left(-\frac{1}{2} X^T \Gamma_n^{-1} X\right)$$

  - Where Γ_n is variance covariance matrix for X of the form

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n-2) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{pmatrix}.$$

- The optimization of the log-likelihood can not be done exactly, rather numerical methods such as the Newton-Raphson algorithm have to be used.
- In particular there are two non-linear functions which are involved.
  - The first is the function which takes the parameters to the auto-covariance function
    $$\phi, \theta, \sigma^2 \to \gamma(h) \to \Gamma_n$$
  - The second is the function which inverts the matrix
    $$\Gamma \to \Gamma_n^{-1}$$

- Finally we estimate σ^2 with the residual sum of squares, as in regression

- Theorem (**Properties of MLE**) Let the parameter of an ARMA (p, q) model be given by
  $$\beta = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)$$
- And let $\hat{\beta}$ be the maximum likelihood estimates, then for large n, we have that the sampling distribution of $\hat{\beta}$ is
  $$N\left(\beta, n^{-1}V(\beta)\right)$$
  - Where $V(\beta)$ is a (p + q) * (p + q) positive definite matrix
- Rather than give the general form of $V(\beta)$, let us look at some special cases

Theorem 3.10.3. (MLE for AR(p)) For an AR(p) model we have $V(\phi) = \sigma^2\Gamma_p^{-1}$, so that

1. When $p = 1$
   $$V(\phi) = \left(1 - \phi_1^2\right).$$
2. When $p = 2$
   $$V(\phi) = \begin{pmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{pmatrix}.$$

Theorem 3.10.4. (MLE for MA(q)) For an MA(q) model two important special cases are

1. When $q = 1$
   $$V(\theta) = \left(1 - \theta_1^2\right).$$
2. When $q = 2$
   $$V(\theta) = \begin{pmatrix} 1-\theta_2^2 & -\theta_1(1+\theta_2) \\ -\theta_1(1+\theta_2) & 1-\theta_2^2 \end{pmatrix}.$$

Theorem 3.10.5. (MLE for ARMA(p, q)) For a causal and invertible ARMA(p, q) model we have

$$V(\phi, \theta) = \frac{1+\phi\theta}{(\phi+\theta)^2} \begin{pmatrix} (1-\phi^2)(1+\phi\theta) & -1(1-\theta^2)(1-\phi^2) \\ -1(1-\theta^2)(1-\phi^2) & (1-\phi^2)(1+\phi\theta) \end{pmatrix}.$$

- Of course to use the likelihood approach we have to assume Normally distributed errors, if weaker assumptions are made inference can still be done: moments method called the Yile-

Walker algorithm, least squares methods also exists

- In general all algorithms require numerically inverting large matrices
  - When computing power was less than it is today algorithms such the innovations and the Durbin-Levinson algorithm have being developed.

Identification of ARMA models

- The previous section worked with the assumption that p and q in the ARMA model were known.
- In practice, we need to go through a model selection and checking procedure just as in regression modelling.
- As we have seen in Chapter 2 one major problem which can seriously affect the ability of a model to forecast correctly is over-fitting.
- Just as in Chapter 2, penalized likelihood methods such as AIC, AICC can be very helpful in finding the model structure.
  - However, in general, all such "black-box" model section procedures must be used with care and the analyst should always check that the fitted model makes sense in the context of the given problem.

- Definition (**AIC**) The Akaike information criterion (AIC) is defined as

$$AIC := -2l(\hat{\beta}) + 2k$$

  - Where $l(\hat{\beta})$ is the maximum value of the log-likelihood for a given model, and k is the number of parameters in the model
- The AICc corrects for small samples and is given by

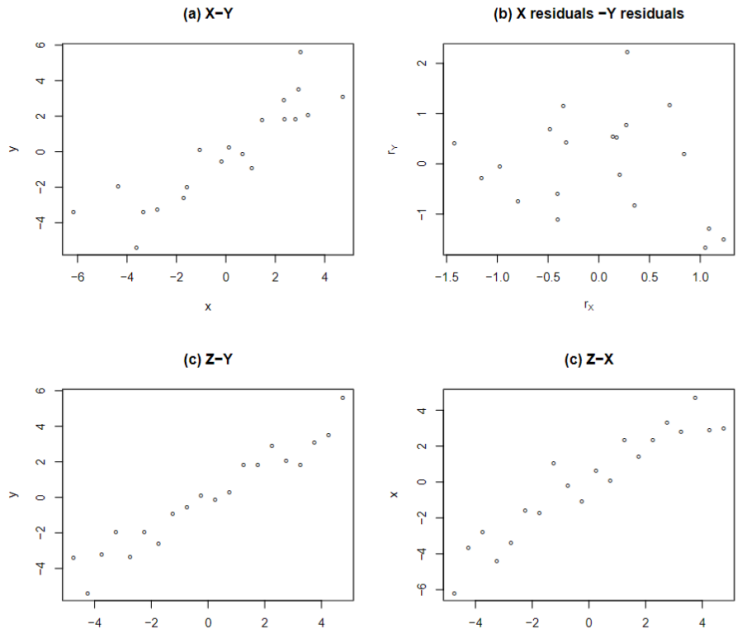
$$AICc := AIC + \frac{2k(k+1)}{n-k-1}$$

- For both criteria in a given set of models the preferred model is the one with the minimum criterion value.
  - The idea is to penalize more complicated model and reward simple models, as long as they fit of course

- We can also use graphical methods to get some ideal of the structure of certain types of ARMA (p, q) models
- Example (**Using the ACF**) notice that if the process is MA(q) then its auto-covariance function will be zero after lag q. Hence inspection of the sample auto-covariance (or auto-correlation) plot gives information about q.
  - This method will not work for an AR(p) model since its auto-covariance function never equal to zero. There is a plot that will help to find p and this is called the partial auto-correlation plot.

- **(PACF)**

- What is partial covariance?
- **Example.** (Adapted from Mardia et al. (1979) [page 170]) Suppose we had 20 observations on verbal skills **(x)**, weight **(y)** and age **(z)** for a group of children.
- If we plot the verbal skills score against weight (see Fig. 1 (a)) we see a high positive correlation (0.89).
- Do we believe there is a 'real' effect between weight and verbal skills?



- We think the verbal skills – weight correlation is **spurious** and just due to the effect of age.
- We can control for the effect of age; regress *x* and *y* on age *z*.
- If we look at the plot of the corresponding residuals (see Fig. 1 (b)) we see almost no correlation left (-0.21).
- Panels (c) and (d) show the very strong common relationship between the two variables and age, which is giving the confounding.
- The covariance between the residuals of X given Z and Y given Z is call the **partial covariance** between X and Y given Z.

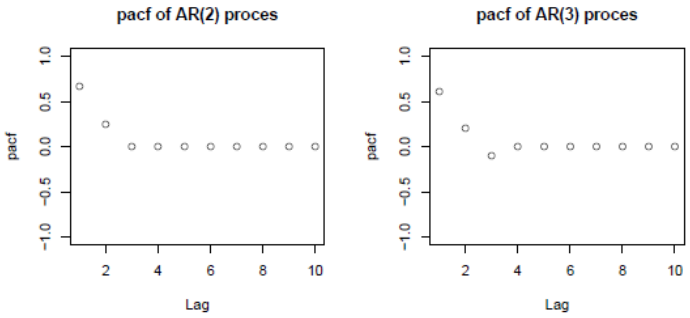
Definition. The partial auto-correlation function (pacf) for a stationary process, $\{X_t\}$, is defined by

$$\alpha(h) = \begin{cases} Cor(X_1, X_0) & \text{for } |h| = 1 \\ \rho_{X_h X_0 \cdot \{X_{h-1} \ldots X_1\}} & \text{for } |h| > 1 \end{cases}$$

where $\rho_{X_h X_0 \cdot \{X_{h-1} \ldots X_1\}}$ is the partial correlation of X_h and X_0 given the set $\{X_{h-1} \ldots X_1\}$.

- Theorem (**Partial auto-correlation function**) if $\alpha(h)$ is the partial auto-correlation function for a stationary AR(p) process, then $\alpha(k) = 0$ for $|k| > p$

- Example (**Using the PACF**) the following figures shows the pacf function for the same models as previous AR() model.



- The plots of the acf above both had the characteristic exponential decay of AR(p) process.
- In contrast the pacf loos different
  - The one for the AR(2) – process has non-zero only values for h = 1, 2
  - The one for the AR(3) process has non-zero only values for h = 1, 2, 3.
  - This shows how the pacf can be used to give information about p in a AR(p) process

Using R for ARMA modelling

- In this section we look at some basic R commands for working with ARMA – process. We start by simulation and plotting of realization of ARMA models

Example 3.11.1. (Using R) The R command:
`sim.ar2 <- arima.sim(n = 100, model=list(ar = c(0.5, 0.25)), sd = sqrt(2))`
simulates a realisation of an AR(2)-process of the form

$$\phi(B)X_t = Z_t$$

where $Z_t \sim N(0, 2)$ and

$$\phi(B) := 1 - 0.5B - 0.25B^2,$$

that is

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t = 0.5X_{t-1} + 0.25X_{t-2} + Z_t$$

The object `sim.ar2` is a `ts` object in R and has some built in attributes, for example

```
> print.ts(sim.ar2)
Time Series:
Start = 1
End = 100
Frequency = 1
1.48885878  4.21586053  4.06171606  3.63421248
  [ ... ]
[97]  1.05016201 -1.54411515 -0.21416876 -0.79918265
```

- We can plot the time series, its auto-correlation function and its partial auto-correlation function using

```
plot.ts(sim.ar2)
acf(sim.ar2)
pacf(sim.ar2)
```

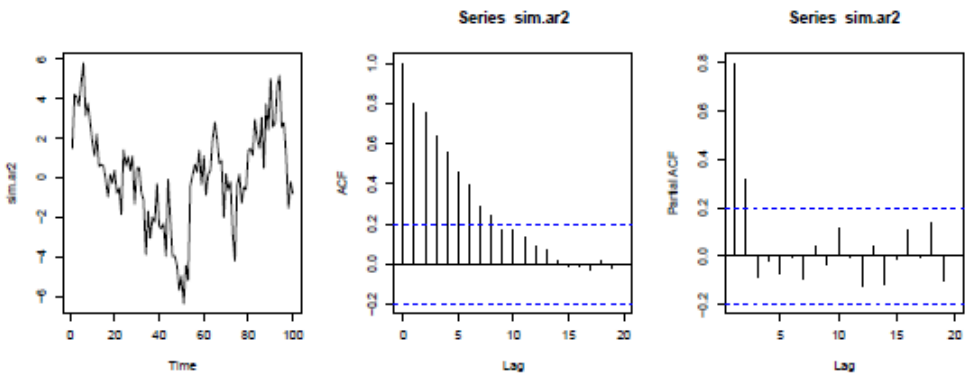


Figure 3.9: Simulated AR(2) model: the time series, its acf and pacf plots

- If you try and simulate from parameter values which do not have a stationary solution you get an error message and no output

```
> arima.sim(n = 100, model=list(ar = c(0.5, 0.5)), sd = sqrt(2))
Error in arima.sim(n = 100, model = list(ar = c(0.5, 0.5)), sd = sqrt(2)) :
  'ar' part of model is not stationary
```

Example 3.11.2. Generating a MA(q) time series is similar, for example we can generate a sample of size 500 from an MA(2) model such that

$$X_t = Z_t + 2Z_{t-1} + 5Z_{t-2}$$

where $Z_t \sim N(0, 10)$ via:

```
sim.ma2 <- arima.sim(n = 500, model=list(ma = c(2, 5)), sd = sqrt(10))
```

Example 3.11.3. For a ARMA(1, 2) example we can generate a sample of size 500 from a model such that

$$X_t + 0.6X_{t-1} = Z_t + 0.6Z_{t-1} - 0.3Z_{t-2}$$

we would use

```
sim.arma12 <- arima.sim(n = 500, list(ar=c(-0.6), ma = c(0.6, -0.3)),
                sd = sqrt(1))
```

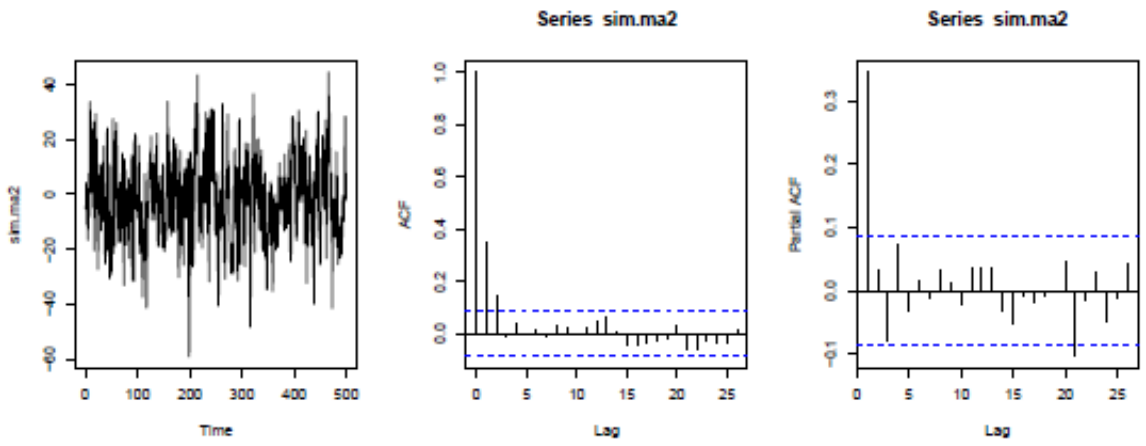


Figure 3.10: Simulated MA(2) model: the time series, its acf and pacf plots

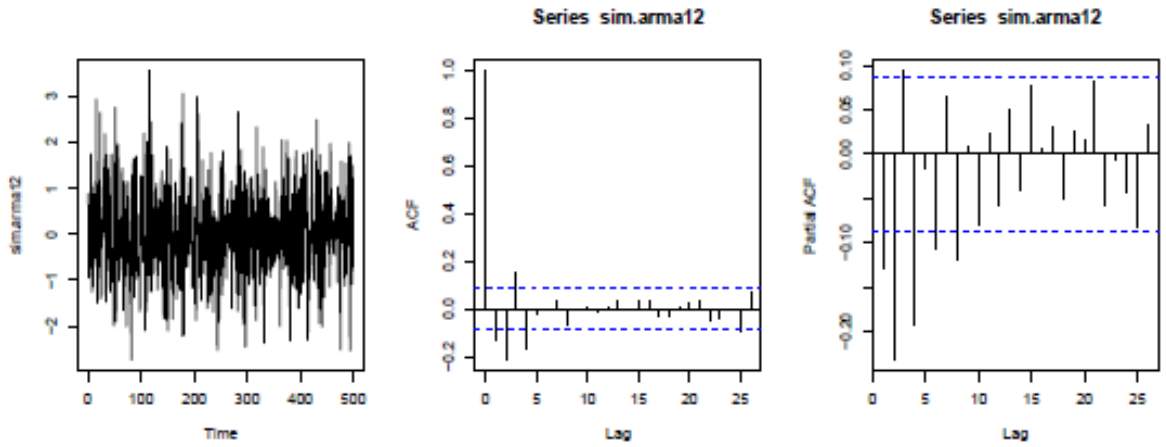


Figure 3.11: Simulated ARMA(1, 2) model: the time series, its acf and pacf plots

Fitting Models

- If we assume, for the moment, that we know the values of p and q in the $ARMA(p,q)$ structure then we can estimate the values of the parameters, using the *arima()* function

```
> arima(sim.arma12, order=c(1, 0,2), include.mean=F)

Coefficients:
          ar1     ma1      ma2
      -0.5891  0.5152  -0.3640
s.e.   0.0653  0.0701   0.0512

sigma^2 estimated as 1.015:  log likelihood = -713.57,  aic = 1435.14
>
```

- From this output we see the following.

True	Estimated	S.E.	Confidence Interval
$\phi_1 = -0.6$	-0.5891	0.0653	(-0.72, -0.46)
$\theta_1 = 0.6$	0.5152	0.0701	(0.38, 0.65)
$\theta_2 = -0.3$	-0.3640	0.0512	(-0.46, -0.26)

- First, we can compare the true parameters values, which we know since this is simulated data to the estimated values and standard errors – and so 95% - confidence intervals.
  - ➢ We see that the true values lie inside the confidence intervals and the width of the confidence intervals are quite small since this example uses a lot of data

- Note that in this fit we used the option *include.mean = F* since I was assuming I knew that it was a mean zero times series. If mean also had to be estimated then we can drop this option

from the function call

- We can now look at the quality of the fit. Just as in regression analysis we use residuals, here estimates of the innovation process Z_t which should be white noise, or if we assume a Gaussian process, i.i.d. Normal data. We do this by the command

```
arima(sim.arma12, order=c(1, 0,2))$residual
```

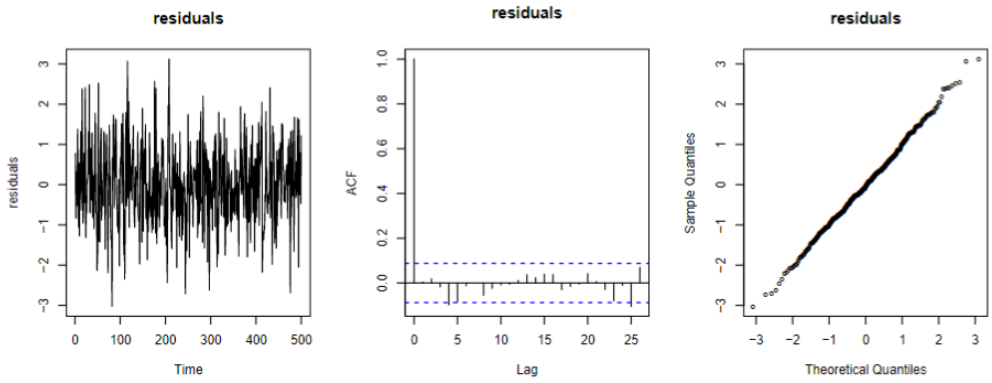


Figure: 3.12 Residual plots: the time series, its acf and QQ norm

- Which output the 500 residuals values. Again, as in regression, we can use plots to evaluate the model fit.
  - For example, we plot the estimated residuals, the acf plot of the residuals – which looks like white noise in this case, and a QQ-plot which shows the residuals do indeed look like they come from a normal distribution

Estimating the Structure

- The previous section assumed that p and q are known. Of course in practice these are not known and a model selection procedure has to be undertaken
  - This is typically done using a mixture of graphical and numerical tools.
  - The blue dashed lines in previous examples can be thought of as pointwise 95%-confidence intervals for the hypothesis that the correlation at h is zero.

Example 3.11.5. For an MA(q) we would expect that the auto-correlation function is zero for all lags greater than q, and it can be shown that the partial auto-correlation function has exponential decay.

For example consider Fig. 3.10. The model which generated the data was MA(2) and we see that in the sample acf plots we have estimated values outside the dashed lines for $h = 0, 1, 2$. The partial auto-correlation function is trivial for all $h > 1$ apart from $h = 22$ and we can plausibly argue that is is just because we would expect 5% of the estimated values to lie outside the lies even if the true values were actually zero.

Example 3.11.6. For an AR(p) model we expect that the auto-correlation function shows exponential decay, while the partial auto-correlation function is zero for all $h > p$. We can see this in Fig. 3.8, where we have in the pacf non-zero values for $h = 0, 1$ and exponential decay in the acf plot.

For general ARMA(p, q) models unfortunately there are no simple patterns that always appear, and even the patterns described in Examples 3.11.5 and 3.11.6 tend to require quite large sample sizes to be used reliably.

The second way of doing model selection is to use the AIC or AICc values from the fit of the models. Since these use the number of the parameters the will give the same penalty to, for example an AR(2), ARMA(1, 1) and MA(2) model. In general there will rarely be the case where the model can be uniquely identified and it is advisable to compare forecasts from different plausible models to see if there are any major differences.

Forecasting

Once a model has been fitted using the `ARIMA( )` function then the `predict( , n.ahead= )` can be used to make forecasts for h steps a head and compute the standard error of the forecast. This has already been shown in Fig. 3.1 and 3.2. These were generated with the following code: The data was a `ts`-object called `sim.forecast`, and I assumed this this came from an $AR(2)$ model, this was then fitted and its (estimated) mean and auto-covariance function was used, as in §3.4, to make a point forecast and estimate the variance around this. Forecasts for $h = 1, 2, 3$ are given by the code

```
> predict(arima(sim.ma2.forecast, order=c(2, 0,0)), n.ahead=3)
$pred
Time Series:
Start = 201
End = 203
Frequency = 1
[1] 13.30448 12.90422 12.75071

$se
Time Series:
Start = 201
End = 203
Frequency = 1
[1] 3.017765 3.780846 4.418923
```

MA(∞) Process

- Just as polynomials can be extended to infinite series, as long as we are careful about convergence, moving average processes can have infinite order.
- This might sound rather abstract but it is the way we can link MA and AR models
- It will give us a way of characterising when AR(p) equations have solutions which are both *stationary* and *causal*

Definition. Let Z_t, $t \in \mathbb{Z}$ be an i.i.d. sequence of $N(0, \sigma^2)$, (or $Z_t \sim \mathrm{WN}(0, \sigma^2)$), random variables. Let $\{\psi_j\}$, $j = 0, 1, \cdots$ be a sequence which is absolutely convergent, i.e.

$$\sum_{j=0}^{\infty} |\psi_j| < \infty,$$

then the process defined by

$$X_t := \sum_{j=0}^{\infty} \psi_j Z_{t-j},$$

for all t, is called a infinite-order moving average, and denoted by $(\mathrm{MA}(\infty))$ process.

Example.3.8.1 (Solving the AR(1) equation)
The process is defined implicitly as the solution of

$$X_t = \phi X_{t-1} + Z_t,$$

for $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$. We can, at least informally, write it as

$$\begin{aligned} X_t &= \phi X_{t-1} + Z_t = \phi(\phi X_{t-2} + Z_{t-1}) + Z_t \\ &= \phi^2 X_{t-2} + (Z_t + \phi Z_{t-1}) = \phi^2(\phi X_{t-3} + Z_{t-2}) + (Z_t + \phi Z_{t-1}) \\ &= \phi^3 X_{t-3} + \left(Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2}\right) \\ &\vdots \\ &= Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \phi^3 Z_{t-3} + \cdots \end{aligned} \tag{1}$$

For this to be an MA(∞) process (Definition 1) we need the coefficients to form an absolutely continuous sum. i.e. $\sum_{j=0}^{\infty} |\phi^j| < \infty$, but we have that, when $|\phi| < 1$ the standard result that

$$\sum_{j=0}^{\infty} |\phi|^j = (1 - |\phi|)^{-1} < \infty.$$

We also see that in Equation (1) that X_t is only a function of Z_s random variables where $s \leq t$. Hence, from the properties of $\mathrm{WN}(0, \sigma^2)$ we have that X_t is uncorrelated with Z_{t+h} for $h > 0$, the last regularity condition of Theorem 3.6.3.

Theorem. The MA(∞) process of Definition 1 is stationary with zero mean and auto-covariance function

$$\gamma(h) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|h|}.$$

Theorem. *Wold decomposition theorem* Any stationary process can be written as the sum of an MA(∞) process and an independent deterministic process, where a deterministic process is any process whose complete realisation is a deterministic function of a finite number of its values.