Chapter 4

Parameter Estimation for ARMA models

4.1 ACVF of an ARMA processes

Recall that a zero-mean time series $\{X_t,\, t \in \mathbb{Z}\}$ is an ARMA(p,q) process if it is stationary and

$$\phi(B)X_t = \theta(B)\epsilon_t, \quad t \in \mathbb{Z} \text{ and } \{\epsilon_t\} \sim \text{WN}(0, \sigma^2),$$

where

$$\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p \qquad \text{and} \qquad \theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q.$$

If the process is causal, then X_t has representation $X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$. The ACVF is:

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

4.2 System of Equations for the ACVF of an ARMA Process

Firstly, suppose that the ϕ's and θ's are given; we compute the ACVF γ in terms of ϕ's and θ's.

The ARMA equation is:

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

Multiply both sides by X_{t-k}, for $k \geq p$, and take expectations. This gives (for mean zero ARMA process):

$$\begin{aligned}
&\gamma(k) - \sum_{j=1}^{p} \phi_j \gamma(k-j) \\
&\quad = \sum_{j=0}^{q} \theta_j \mathbb{E}[X_{t-k}\epsilon_{t-j}] = \sum_{j=0}^{q}\sum_{i=0}^{\infty} \psi_i\theta_j \mathbb{E}[\epsilon_{t-k-i}\epsilon_{t-j}] \\
&\quad = \sigma^2 \sum_{j=0}^{q}\sum_{i=0}^{\infty} \psi_i\theta_j \mathbf{1}_{k+i=j} = \begin{cases} \sigma^2 \sum_{j=k}^{q} \psi_{j-k}\theta_j & k \le q \\ 0 & k \ge q+1. \end{cases}
\end{aligned} \tag{4.1}$$

where $\theta_0 = 1$ and $\theta_j = 0$ for $j > q$. Consider the set of equations

$$\gamma(k) - \sum_{j=1}^{p} \phi_j \gamma(k-j) = 0 \qquad k \ge \max(p, q+1) \tag{4.2}$$

and look for solutions $\gamma(k) = az^k$. An expression of this form gives a solution provided:

$$z^k - \sum_{j=1}^{p} \phi_j z^{k-j} = 0 \qquad k \ge \max(p, q+1)$$

and hence

$$1 - \sum_{j=1}^{p} \phi_j z^{-j} = 0.$$

Let $y = z^{-1}$, then $1 - \sum_{j=1}^{p} \phi_j y^j = 0$. This is a polynomial of degree p. Let $\xi_1, \ldots, \xi_p$ denote its roots then, by linearity,

$$\gamma(k) = a_1\xi_1^{-k} + \ldots + a_p\xi_p^{-k} \qquad k \ge m - p$$

provides a solution to Equation (4.2), where $a_1, \ldots, a_p$ are arbitrary constants. If the roots are distinct, then this provides all the solutions, otherwise there are others. For this discussion, we only consider the case of distinct roots. The constants $a_1, \ldots, a_p$ and the remaining covariances $\gamma(h)$ for $h = 0, \ldots, m-p$ are determined by Equation (4.1).

This can be used as the basis of a method of moments estimation procedure. Assume we have estimated $\widehat{\gamma}$ for as many lags as necessary.

Example 4.1 (MA(1) process)**.**

Let $\{X_t\}$ be a MA(1) process:

$$X_t = \epsilon_t + \theta\epsilon_{t-1}, \quad \{\epsilon_t\} \sim \mathrm{WN}(0, \sigma^2),$$

where $|\theta| < 1$. In this case the equations are:

$$\begin{cases} \gamma(0) = \sigma^2(1+\theta^2) & \gamma(1) = \sigma^2\theta \\ \gamma(k) = 0 & |k| \ge 2. \end{cases}$$

This is the autocovariance function. Using

$$\rho(1) = \frac{\gamma(1)}{\gamma(0)} = \frac{\theta}{1+\theta^2},$$

it is natural to estimate θ by the method of moments:

$$\widehat{\rho}(1) = \frac{\widehat{\gamma}(1)}{\widehat{\gamma}(0)} = \frac{\widehat{\theta}_n^{(1)}}{1+\left(\widehat{\theta}_n^{(1)}\right)^2}.$$

This equation has a solution for $|\widehat{\rho}(1)| < \frac{1}{2}$ and it is natural to set:

$$\widehat{\theta}_n^{(1)} = \begin{cases} -1 & \text{if } \widehat{\rho}(1) < -\frac{1}{2}\,, \\ \dfrac{1-\sqrt{1-4\widehat{\rho}(1)^2}}{2\widehat{\rho}(1)} & \text{if } |\widehat{\rho}(1)| < \frac{1}{2}\,, \\ 1 & \text{if } \widehat{\rho}(1) > \frac{1}{2}\,. \end{cases}$$

The estimate $\widehat{\theta}_n^{(1)}$ is consistent. Furthermore, it can be shown that:

$$\widehat{\theta}_n^{(1)} \sim \mathrm{AN}(\theta, n^{-1}\sigma_1^2(\theta)), \quad \text{for large values of } n,$$

where

$$\sigma_1^2(\theta) = \frac{1+\theta^2+4\theta^4+\theta^6+\theta^8}{(1-\theta^2)^2}.$$

□

Example 4.2.

Let $\{X_t\}$ be a causal AR(p) process. Them $m = p$ and Equation (4.1) reduces to

$$\gamma(k) - \phi_1\gamma(k-1) - \ldots - \phi_p\gamma(k-p) = \begin{cases} 0, & k = 1, \ldots, p, \\ \sigma^2, & k = 0, \end{cases}$$

which are the *Yule-Walker equations*. When $p = 2$, this is:

$$\begin{cases} \gamma(k) - \phi_1\gamma(k-1) - \phi_2\gamma(k-2) = 0 & k \geq 2 \\ \gamma(1) - \phi_1\gamma(0) - \phi_2\gamma(1) = 0 \\ \gamma(0) - \phi_1\gamma(1) - \phi_2\gamma(2) = \sigma^2 \end{cases}$$

Here

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2$$

and the two solutions to $\phi(z) = 0$ are:

$$z = -\frac{\phi_1}{2\phi_2} \pm \sqrt{\frac{\phi_1^2}{4\phi_2^2} + \frac{1}{\phi_2}} = (\xi_1, \xi_2).$$

The general solution is therefore:

$$\gamma(h) = a_1 \frac{1}{\xi_1^h} + a_2 \frac{1}{\xi_2^h}.$$

The constants a_1 and a_2 are then computed from the second and third listed equations by:

$$\begin{cases} (1-\phi_2)(a_1\frac{1}{\xi_1} + a_2\frac{1}{\xi_2}) - \phi_1(a_1+a_2) = 0 \\ (a_1+a_2) - \phi_1(\frac{a_1}{\xi_1} + \frac{a_2}{\xi_2}) - \phi_2(\frac{a_1}{\xi_1^2} + \frac{a_2}{\xi_2^2}) = \sigma^2. \end{cases}$$

This gives two linear equations with two unknowns and therefore has a unique solution. □

The roots $\xi_1, \ldots, \xi_p$ may be obtained numerically and then the remaining equations form a linear system from which $a_1, \ldots, a_p$ may be computed. □

4.3 Yule-Walker estimation

Consider a causal zero-mean AR(p) process $\{X_t\}$:

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = \epsilon_t, \quad \{\epsilon_t\} \sim \text{IIDN}(0, \sigma^2).$$

The *Yule-Walker equations* are defined as:

$$\gamma(j) - \phi_1\gamma(j-1) - \ldots - \phi_p\gamma(j-p) = \begin{cases} 0 & j = 1, \ldots, p, \\ \sigma^2 & j = 0, \end{cases}.$$

and these may be obtained quite simply, by considering

$$\mathbb{E}[X_{t-h}(X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p})] = \mathbb{E}[X_{t-h}\epsilon_t].$$

These equations may be rewritten as:

$$\phi_1\gamma(j-1) + \ldots + \phi_p\gamma(j-p) = \begin{cases} \gamma(j), & j = 1, \ldots, p, \\ \gamma(0) - \sigma^2, & j = 0, \end{cases}$$

which gives:

$$\begin{cases} \Gamma_p \underline{\phi} = \underline{\gamma}_p \\ \underline{\phi}'\underline{\gamma}_p = \gamma(0) - \sigma^2 \end{cases} \quad \Rightarrow \quad \sigma^2 = \gamma(0) - \underline{\phi}'\underline{\gamma}_p,$$

where

$$\Gamma_p = \begin{pmatrix} \gamma(0) & \ldots & \gamma(p-1) \\ \vdots & & \\ \gamma(p-1) & \ldots & \gamma(0) \end{pmatrix} \quad \text{and} \quad \underline{\gamma}_p = \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(p) \end{pmatrix}.$$

Given estimates of the autocovariance function $\widehat{\gamma}(\cdot)$, the Yule-Walker equations may be used to obtain moment method estimates of $\underline{\phi}$ and σ^2 by replacing Γ_p and $\underline{\gamma}_p$ with the estimates $\widehat{\Gamma}_p$ and $\widehat{\underline{\gamma}}_p$. These are known as the *Yule-Walker estimates*

$$\widehat{\Gamma}_p\widehat{\underline{\phi}} = \widehat{\underline{\gamma}}_p \quad \text{and} \quad \widehat{\sigma}^2 = \widehat{\gamma}(0) - \widehat{\underline{\phi}}\,'\widehat{\underline{\gamma}}_p,$$

where (of course)

$$\widehat{\Gamma}_p = \begin{pmatrix} \widehat{\gamma}(0) & \dots & \widehat{\gamma}(p-1) \\ \vdots & & \\ \widehat{\gamma}(p-1) & \dots & \widehat{\gamma}(0) \end{pmatrix} \quad \text{and} \quad \widehat{\underline{\gamma}}_p = \begin{pmatrix} \widehat{\gamma}(1) \\ \vdots \\ \widehat{\gamma}(p) \end{pmatrix}.$$

The estimator $\widehat{\rho}(\cdot)$ for the autocorrelation may be obtained simply by dividing the equations for the autocovariance by $\widehat{\gamma}(\cdot)$. This gives:

$$\widehat{R}_p\widehat{\underline{\phi}} = \widehat{\underline{\rho}}_p \quad \text{and} \quad \widehat{\sigma}^2 = \widehat{\gamma}(0)[1 - \widehat{\underline{\phi}}\,'\widehat{\underline{\rho}}_p],$$

where, of course, $\widehat{R}_p = \frac{1}{\widehat{\gamma}(0)}\widehat{\Gamma}_p$ and $\widehat{\underline{\rho}}_p = \frac{1}{\widehat{\gamma}(0)}\widehat{\underline{\gamma}}_p$.
This gives:

$$\widehat{\underline{\phi}} = \widehat{R}_p^{-1}\widehat{\underline{\rho}}_p \quad \text{and} \quad \widehat{\sigma}^2 = \widehat{\gamma}(0)[1 - \widehat{\underline{\rho}}_p\,'\widehat{R}_p^{-1}\widehat{\underline{\rho}}_p].$$

The following theorem is stated without proof; asymptotic normality of the sample mean has already been dealt with and the main techniques for the following result is similar.

Theorem 4.1. *Let $\{X_t\}$ be a causal AR(p) process where $\{\epsilon_t\} \sim IID(0, \sigma^2)$. Let $\widehat{\underline{\phi}}$ be the Yule-Walker estimate of $\underline{\phi}$, then*

$$\widehat{\underline{\phi}} \sim AN\left(\underline{\phi}, \frac{\sigma^2\Gamma_p^{-1}}{n}\right), \quad \textit{for large values of } n.$$

The estimator of σ^2 is asymptotically consistent:

$$\widehat{\sigma}^2 \longrightarrow_{(p)} \sigma^2.$$

Proof Omitted (we'll deal with this in the next lecture when we consider statistical properties of the ACVF and ACF estimators). □

Now consider an ARMA(p, q) process where that $q > 0$. If estimates of the autocovariance function are available, then the system of equations (4.1) may be used (replacing the autocovariance with the estimated autocovariance) to obtain moment method estimates of $\widehat{\underline{\phi}}$ and $\widehat{\underline{\theta}}$.

4.4 The Hannan-Rissanen algorithm

Yule-Walker estimation works well for AR processes. Suppose the underlying process is AR, but p is unknown. We can proceed as if $\{X_t\}$ is an AR(m) process for $m = 1, 2, \ldots$ until it looks as if $m \geq p$. For any fixed $m > p$, set

$$\underline{\phi}_m = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Theorem 4.1 holds with p replaced by m and $\underline{\phi}$ by $\underline{\phi}_m$.

Now consider the prediction problem. Recall that the best linear predictor $\widehat{X}_{n+1}$ of X_{n+1} in terms of $X_1, X_2, \ldots, X_n$ is

$$\widehat{X}_{n+1} = \sum_{i=1}^{n} a_{n,i} X_{n+1-i}, \quad n = 1, 2, \ldots,$$

where $\underline{a}_n = \Gamma_n^{-1} \underline{\gamma}_n$. It follows from the fact that the partial correlation is 0 for lags greater than p for a causal AR(p) process that $\underline{a}_n = \begin{pmatrix} \underline{\phi} \\ \underline{0} \end{pmatrix}$ when $n > p$ and therefore the parameters of an AR(p) process can be estimated from solving the prediction problem.

4.5 The Hannan-Rissanen algorithm

For a causal AR(p) model, with no further distributional assumptions on $\{\epsilon_t\} \sim \text{WN}(0, \sigma^2)$, the defining equation

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = \epsilon_t$$

of a causal zero-mean AR(p) can be written on the form

$$\underline{Y} = X\underline{\phi} + \underline{\epsilon} \quad \text{or} \quad \underline{\epsilon} = \underline{Y} - X\underline{\phi},$$

where

$$\underline{Y} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \quad \underline{x}_k = \begin{pmatrix} X_{-k} \\ \vdots \\ X_{n-1-k} \end{pmatrix} \quad \text{for } k = 0, \ldots, p-1,$$

$$X = (\underline{x}_0, \ldots, \underline{x}_{p-1}) = \begin{pmatrix} X_0 & X_{-1} & \ldots & X_{1-p} \\ \vdots & \vdots & & \\ X_{n-1} & X_{n-2} & \ldots & X_{n-p} \end{pmatrix} \quad \text{and} \quad \underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

The idea in least square estimation is to consider the X_k:s as fixed and to minimise $\widehat{\underline{\epsilon}}'\widehat{\underline{\epsilon}}$ with respect to $\underline{\phi}$. Assume that $X_{-p+1}, \ldots, X_n$ are observed. Let $\widehat{\underline{\phi}}$ denote the least square estimate, i.e. the value of $\underline{\phi}$ which minimises

$$S(\underline{\phi}) = \widehat{\underline{\epsilon}}'\widehat{\underline{\epsilon}} = \|\widehat{\underline{\epsilon}}\|^2 = \|\underline{Y} - X\underline{\phi}\|^2.$$

Consider the Hilbert spaces

$$\mathcal{H} = \overline{\text{spa}}\{\underline{Y}, \underline{x}_0, \ldots, \underline{x}_{p-1}\} \quad \text{and} \quad \mathcal{M} = \overline{\text{spa}}\{\underline{x}_0, \ldots, \underline{x}_{p-1}\}.$$

It follows from the projection theorem that $\widehat{\phi}$ satisfies:

$$P_{\mathcal{M}}\underline{Y} = X\widehat{\underline{\phi}}.$$

It follows from the projection theorem that:

$$\langle \underline{x}_k, X\widehat{\underline{\phi}} \rangle = \langle \underline{x}_k, \underline{Y} \rangle \text{ for } k = 0, \ldots, p-1$$

from which

$$X^t X\widehat{\underline{\phi}} = X'\underline{Y}$$

giving

$$\widehat{\underline{\phi}} = (X^t X)^{-1} X^t \underline{Y} \text{ provided } X^t X \text{ is non-singular.}$$

The estimator $\widehat{\underline{\phi}}$ has good statistical properties if $p << n$.

Now let $\{X_t\}$ be a general ARMA(p, q) process with $q > 0$:

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}, \quad \{Z_t\} \sim \text{IID}(0, \sigma^2).$$

The problem is that X_t is regressed not only onto $X_{t-1}, \ldots, X_{t-p}$ but also on the unobserved quantities $\epsilon_{t-1}, \ldots, \epsilon_{t-q}$. The main idea in the *Hannan–Rissanen algorithm* is to first replace $\epsilon_{t-1}, \ldots, \epsilon_{t-q}$ with their estimates $\widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-q}$ and then to estimate

$$\underline{\beta} := \begin{pmatrix} \underline{\phi} \\ \underline{\theta} \end{pmatrix}$$

by regressing X_t onto $X_{t-1}, \ldots, X_{t-p}, \widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-q}$.
In more detail:

Step 1 A high order AR(m) model (with $m > \max(p,q)$) is fitted to the data by Yule-Walker estimation. If $\widehat{\phi}_{m1}, \ldots, \widehat{\phi}_{mm}$ are the estimated coefficients, then ϵ_t is estimated by

$$\widehat{\epsilon}_t = X_t - \widehat{\phi}_{m1} X_{t-1} - \ldots - \widehat{\phi}_{mm} X_{t-m}, \quad t = m+1, \ldots, n.$$

Step 2 The vector $\underline{\beta}$ is estimated by least square regression of X_t onto

$$X_{t-1}, \ldots, X_{t-p}, \widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-q} :$$

that is, by minimising

$$S(\underline{\beta}) = \sum_{t=m+1}^{n} (X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} - \theta_1 \widehat{\epsilon}_{t-1} - \ldots - \theta_q \widehat{\epsilon}_{t-q})^2$$

with respect to $\underline{\beta}$. This gives the Hannan Rissanen estimator

$$\widehat{\underline{\beta}} = (Z'Z)^{-1} Z' \underline{X}_n \quad \text{provided } Z'Z \text{ is non-singular,}$$

where

$$\underline{X}_n = \begin{pmatrix} X_{m+1} \\ \vdots \\ X_n \end{pmatrix}$$

and

$$Z = \begin{pmatrix} X_m & X_{m-1} & \ldots & X_{m-p+1} & \widehat{\epsilon}_m & \widehat{\epsilon}_{m-1} & \ldots & \widehat{\epsilon}_{m-q+1} \\ \vdots & \vdots & & & & & & \\ X_{n-1} & X_{n-2} & \ldots & X_{n-p} & \widehat{\epsilon}_{n-1} & \widehat{\epsilon}_{n-2} & \ldots & \widehat{\epsilon}_{n-q} \end{pmatrix}.$$

The Hannan Rissanen estimate of the white noise variance σ^2 is:

$$\widehat{\sigma}^2_{\text{HR}} = \frac{S(\widehat{\underline{\beta}})}{n-m}.$$

An appropriate ARMA(p,q) model

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}, \quad \{\epsilon_t\} \sim \text{IID}(0, \sigma^2),$$

requires an *order selection*, that is a choice of p and q and, having chosen p and q, estimates of the unknown parameters. Firstly, the mean is estimated, then it is removed, and then, having removed the mean, estimate:

$$\underline{\phi} = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix}, \quad \underline{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_q \end{pmatrix} \quad \text{and} \quad \sigma^2.$$

As usual, assume that $X_1, \ldots, X_n$ are observed. The assumption that $\{\epsilon_t\} \sim \text{IIDN}(0, \sigma^2)$ (rather than $\text{WN}(0, \sigma^2)$) allows for greater precision in the estimates.

4.6 Maximum Likelihood and Least Square estimation

If the innovations are assumed to be IID $N(0, \sigma^2)$ (Gaussian), then it is possible to obtain estimates using the method of maximum likelihood. Least squares may also be used. The least squares estimator is obtained by minimising:

$$S(\underline{\phi}, \underline{\theta}) = \sum_{j=1}^{n} \frac{(X_j - \widehat{X}_j)^2}{r_{j-1}},$$

where $r_{j-1} = v_{j-1}/\sigma^2$, with respect to $\underline{\phi}$ and $\underline{\theta}$. This is straightforward and will be outlined below. The estimates are obtained by recursive methods. The least square estimate of σ^2 is

$$\widehat{\sigma}^2_{\mathrm{LS}} = \frac{S(\widehat{\underline{\phi}}_{\mathrm{LS}}, \widehat{\underline{\theta}}_{\mathrm{LS}})}{n - p - q}$$

where, of course, $(\widehat{\underline{\phi}}_{\mathrm{LS}}, \widehat{\underline{\theta}}_{\mathrm{LS}})$ is the estimate obtained by minimizing $S(\underline{\phi}, \underline{\theta})$.

Example 4.3 (MA(1) process)**.**

As usual, for an MA(1) process,

$$\begin{aligned} X_1 &= \epsilon_1 + \theta\epsilon_0 \quad \text{or} \quad \epsilon_1 = X_1 - \theta\epsilon_0 \\ X_2 &= \epsilon_2 + \theta\epsilon_1 \quad \text{or} \quad \epsilon_2 = X_2 - \theta\epsilon_1 \\ &\vdots \\ X_n &= \epsilon_n + \theta\epsilon_{n-1} \quad \text{or} \quad \epsilon_n = X_n - \theta\epsilon_{n-1} \end{aligned}$$

If $\epsilon_0 = 0$, then $\epsilon_1, \ldots, \epsilon_n$ can be calculated for a given θ. Since $\widehat{X}_k = \theta\epsilon_{k-1}$, it follows that $v_j = \sigma^2$ and hence $r_j = 1$ for all j and $\sum_{j=1}^{n} \epsilon_j^2$ can be minimised numerically with respect to θ. Let $\widehat{\theta}_n^{(2)}$ denote the estimate. Then it can be shown that

$$\widehat{\theta}_n^{(2)} \sim \mathrm{AN}\left(\theta, \frac{(1-\theta^2)}{n}\right), \quad \text{for large values of } n.$$

For the general ARMA process, the $\widehat{X}_j$s may be computed recursively by the innovations algorithm. Recall (from earlier that $X_1 - \widehat{X}_1, X_2 - \widehat{X}_2, \ldots, X_n - \widehat{X}_n$ are *orthogonal*. This means that they are *uncorrelated*. Under the assumption that the process is Gaussian, this implies that they are *independent* and the Mean Squared Prediction Error gives their respective variances.

It follows that, for any fixed values of $\underline{\phi}$, $\underline{\theta}$, and σ^2, the innovations $X_1 - \widehat{X}_1, \ldots, X_n - \widehat{X}_n$ are independent and normally distributed with zero means and variances $v_0 = \sigma^2 r_0 = \gamma_X(0)$, $v_1 = \sigma^2 r_1, \ldots, v_{n-1} = \sigma^2 r_{n-1}$. Thus the density of $X_j - \widehat{X}_j$ is

$$f_{X_j - \widehat{X}_j}(x) = \frac{1}{\sqrt{2\pi\sigma^2 r_{j-1}}} \exp\left\{-\frac{x^2}{2\sigma^2 r_{j-1}}\right\}.$$

The likelihood function is therefore:

$$\begin{aligned} L(\underline{\phi}, \underline{\theta}, \sigma^2; \underline{x}) &= \prod_{j=1}^{n} f_{X_j - \widehat{X}_j}(x_j - \widehat{x}_j) \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n r_0 \cdots r_{n-1}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} \frac{(x_j - \widehat{x}_j)^2}{r_{j-1}} \right\} \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n r_0 \cdots r_{n-1}}} \exp\left\{ -\frac{S(\underline{\phi}, \underline{\theta})}{2\sigma^2} \right\}. \end{aligned}$$

Proceeding in the usual way,

$$\ln L(\underline{\phi}, \underline{\theta}, \sigma^2) = -\frac{1}{2} \ln((2\pi\sigma^2)^n r_0 \cdots r_{n-1}) - \frac{S(\underline{\phi}, \underline{\theta})}{2\sigma^2}.$$

Clearly, $r_0, \ldots, r_{n-1}$ depend on $\underline{\phi}$ and $\underline{\theta}$, but not on σ^2. For fixed values of $\underline{\phi}$ and $\underline{\theta}$,

$$\frac{\partial \ln L(\underline{\phi}, \underline{\theta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{S(\underline{\phi}, \underline{\theta})}{2(\sigma^2)^2},$$

The log likelihood function $\ln L(\underline{\phi}, \underline{\theta}, \sigma^2)$ is maximized by $\sigma^2 = n^{-1} S(\underline{\phi}, \underline{\theta})$, from which:

$$\begin{aligned} \ln L(\underline{\phi}, \underline{\theta}, n^{-1} S(\underline{\phi}, \underline{\theta})) &= -\frac{1}{2} \ln((2\pi n^{-1} S(\underline{\phi}, \underline{\theta}))^n r_0 \cdots r_{n-1}) - \frac{n}{2} \\ &= -\frac{1}{2} \left(n \ln(2\pi) + n \ln(n^{-1} S(\underline{\phi}, \underline{\theta})) + \ln r_0 + \ldots + \ln r_{n-1} \right) - \frac{n}{2} \\ &= -\frac{n}{2} \left(\ln(n^{-1} S(\underline{\phi}, \underline{\theta})) + n^{-1} \sum_{j=1}^{n} \ln r_{j-1} \right) + \text{constant}. \end{aligned}$$

It follows that the problem of maximising $\ln L(\underline{\phi}, \underline{\theta}, \sigma^2)$ is the same as the problem of minimising

$$\ell(\underline{\phi}, \underline{\theta}) = \ln(n^{-1} S(\underline{\phi}, \underline{\theta})) + n^{-1} \sum_{j=1}^{n} \ln r_{j-1}.$$

Numerical methods are required for this.

For a process that is both causal and invertible, $r_n \to 1$ and therefore $n^{-1} \sum_{j=1}^{n} \ln r_{j-1}$ is asymptotically negligible compared with $\ln S(\underline{\phi}, \underline{\theta})$. It follows that both the least square and the maximum likelihood methods give asymptotically the same result for causal invertible processes.

4.7 Order selection

Now assume that an ARMA(p, q) process gives a good model for the time series. To begin with, suppose that $q = 0$; it is known that an AR(p) process provides a good model, but the value of p is unknown. A natural approach would be to try fitting AR(m) models for increasing values of m. For

each m, a quantity that indicates the validity of the model, for example $S(\widehat{\underline{\phi}})$ or $L(\widehat{\underline{\phi}}, \widehat{\sigma}^2)$ is calculated. If $m \leq p$, then $S(\widehat{\underline{\phi}})$ should decrease with m; it should remain constant for $m \geq p$. Similarly $L(\widehat{\underline{\phi}}, \widehat{\sigma}^2)$ should increase for $m \leq p$.

The problem with these measures is that they continue to show improvement (reduction in sum of squares, increase in the log likelihood), even if the parameter being estimated is 0; a model with n observations and n fitted parameters will have zero residual sum of squares.

Therefore, when fitting an ARMA(p, q) process to data (that is estimating p, q, $(\underline{\phi}, \underline{\theta})$ and σ^2) a penalty has to be introduced when adding parameters. Consider maximum likelihood estimation; maximising $L(\underline{\phi}, \underline{\theta}, \sigma^2)$ or, equivalently, minimising $-2 \ln L(\underline{\phi}, \underline{\theta}, \sigma^2)$, where L is regarded as a function also of p and q. The likelihood will be maximised when the total number of parameters $p + q$ is equal to the number of observations. There are two accepted approaches found in the literature to penalise additional parameters; the *Akaike Information Criterion* (AIC) and the *Bayesian Information Criterion* (BIC). The Bayesian Information Criterion is almost the same as the *Minimum Description Length* (MDL). The form of the AIC commonly used is the AICC criterion, where the additional C means 'bias corrected'. The criterion is: choose p, q, and $(\underline{\phi}_p, \underline{\theta}_q)$, to minimise

$$\begin{cases} \text{AIC} = -2 \ln L(\underline{\phi}_p, \underline{\theta}_q, S(\underline{\phi}_p, \underline{\theta}_q)/n) + 2(p + q + 1) \\ \text{AICC} = -2 \ln L(\underline{\phi}_p, \underline{\theta}_q, S(\underline{\phi}_p, \underline{\theta}_q)/n) + 2(p + q + 1) \frac{n}{n-p-q-2}, \end{cases}$$

where there is the additional requirement (over the AIC) that the minimisation is restricted to the class of unbiased estimators.

The resulting estimates of the number of parameters $\widehat{p}$ and $\widehat{q}$ are not consistent; they do not satisfy

$$\widehat{p} \longrightarrow_{(p)} p \text{ and } \widehat{q} \longrightarrow_{(p)} q \text{ as } n \to \infty.$$

The BIC, minimises

$$\text{BIC} = -2 \ln L(\underline{\phi}_p, \underline{\theta}_q, S(\underline{\phi}_p, \underline{\theta}_q)/n) + 2(p + q + 1) \ln n,$$

and $\widehat{p}, \widehat{q}$ are consistent with BIC.