15 Forecasting Time Series

15.1 Forecasting Stationary Time Series

We investigate the problem of predicting the values X_{n+h}, $h > 0$, of a stationary time series with known mean μ and autocovariance function γ in terms of $\{X_n, \ldots, X_1\}$. Our goal is to find the linear combination of $1, X_n, X_{n-1}, \ldots, X_1$, that forecasts X_{n+h} with minimum mean squared error. The best linear predictor will be denoted by $P_n X_{n+h}$ and clearly has the form

$$P_n X_{n+h} = a_0 + a_1 X_n + \ldots + a_n X_1.$$

It remains to determine the coefficients $a_0, \ldots, a_n$, by finding the values that minimize

$$S(a_0, \ldots, a_n) = \mathrm{E}(X_{n+h} - a_0 - a_1 X_n - \ldots - a_n X_1)^2.$$

Since S is a quadratic function of $a_0, \ldots, a_n$ and is bounded below by zero, it is clear that there is at least one value of $(a_0, \ldots, a_n)$ that minimizes S and that the minimum satisfies

$$\frac{\partial S(a_0, \ldots, a_n)}{\partial a_j} = 0, \quad j = 0, \ldots, n.$$

Evaluation of the derivatives gives the equivalent equations

$$\mathrm{E}\left[X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i}\right] = 0, \tag{15.1}$$

$$\mathrm{E}\left[\left(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i}\right) X_{n+1-j}\right] = 0, \quad j = 1, \ldots, n. \tag{15.2}$$

These equations can be written as

$$a_0 = \mu\left(1 - \sum_{i=1}^{n} a_i\right), \tag{15.3}$$

$$\boldsymbol{\Gamma}_n \boldsymbol{a}_n = \boldsymbol{\gamma}_n(h), \tag{15.4}$$

where $\boldsymbol{a}_n = (a_1, \ldots, a_n)'$, $\boldsymbol{\Gamma}_n := [\gamma(i-j)]_{i,j=1}^{n}$ and $\boldsymbol{\gamma}_n(h) := (\gamma(h), \ldots, \gamma(h+n-1))'$.

Proposition 15.1.1. *The the best linear predictor $P_n X_{n+h}$ is*

$$P_n X_{n+h} = \mu + \sum_{i=1}^{n} a_i (X_{n+1-i} - \mu), \tag{15.5}$$

where $\boldsymbol{a}_n$ satisfies (15.4). *From* (15.5) *the expected value of the prediction error $X_{n+h} - P_n X_{n+h}$ is zero, and the mean square prediction error is therefore*

$$\mathrm{E}(X_{n+h} - P_n X_{n+h})^2 = \gamma(0) - \boldsymbol{a_n}' \boldsymbol{\gamma}_n(h),$$

where $\boldsymbol{a}_n = (a_1, \ldots, a_n)'$ and $\boldsymbol{\gamma}_n(h) := (\gamma(h), \ldots, \gamma(h+n-1))'$.

Remark. If $\{Y_t\}$ is a stationary time series with mean μ and if $\{X_t\}$ is the zero-mean series defined by $X_t = Y_t - \mu$, then $P_n Y_{n+h} = \mu + P_n X_{n+h}$. So from now on, we can restrict attention to zero-mean stationary time series.

Example. Consider the time series

$$X_t = \phi X_{t-1} + Z_t, \quad t = 0, \pm 1, \ldots,$$

where $|\phi| < 1$ and $Z_t \sim \mathrm{WN}(0, \sigma^2)$. The best linear predictor of X_{n+1} in terms of $\{1, X_n, \ldots, X_1\}$ is

$$P_n X_{n+1} = \boldsymbol{a}_n' \boldsymbol{X}_n,$$

where $\boldsymbol{X}_n = (X_n, \ldots, X_1)'$ and

$$\begin{pmatrix} 1 & \phi & \ldots & \phi^{n-1} \\ \phi & 1 & \ldots & \phi^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \ldots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \phi \\ \phi^2 \\ \vdots \\ \phi^n \end{pmatrix}.$$

A solution is clearly

$$\boldsymbol{a}_n = (\phi, 0, \ldots, 0)',$$

and hence the best linear predictor of X_{n+1} in terms of $\{X_1, \ldots, X_n\}$ is

$$P_n X_{n+1} = \boldsymbol{a}_n' \boldsymbol{X}_n = \phi X_n,$$

with mean squared error

$$\mathrm{E}(X_{n+1} - P_n X_{n+1})^2 = \gamma(0) - \boldsymbol{a}_n' \boldsymbol{\gamma}_n(1) = \frac{\sigma^2}{1-\phi^2} - \phi\gamma(1) = \sigma^2.$$

Remark. If $\{X_t\}$ is a zero-mean stationary series with autocovariance function $\gamma(\cdot)$, then in principle determining the best linear predictor $P_n X_{n+h}$ in terms of $\{X_n, \ldots, X_1\}$ is possible. However, the direct approach requires the determination of a solution of a system of n linear equations, which for large n may be difficult and time-consuming. Therefore it would be helpful if the one-step predictor $P_n X_{n+1}$ based on the n previous observations could be used to simplify the calculation of $P_{n+1} X_{n+2}$, the one-step predictor based on $n+1$ previous observations. Prediction algorithms that utilize this idea are said to be recursive. The algorithms to be discussed in this chapter allow us to compute best predictors without having to perform any matrix inversions.

Remark. There are two important recursive prediction algorithms:

- Durbin-Levinson algorithm: Section 15.1.1 and Brockwell and Davis (2002) pp. 69-71.

- Innovations algorithm: Section 15.1.2, Section 16.1.3, Brockwell and Davis (2002) pp. 71-75 and pp. 150-156.

15.1.1 Durbin-Levinson Algorithm

From Proposition 15.1.1 we know that if the matrix $\mathbf{\Gamma}_n$ is nonsingular, then

$$P_n X_{n+1} = \boldsymbol{\phi}_n' \boldsymbol{X}_n = \phi_{n1} X_n + \ldots + \phi_{nn} X_1, \tag{15.6}$$

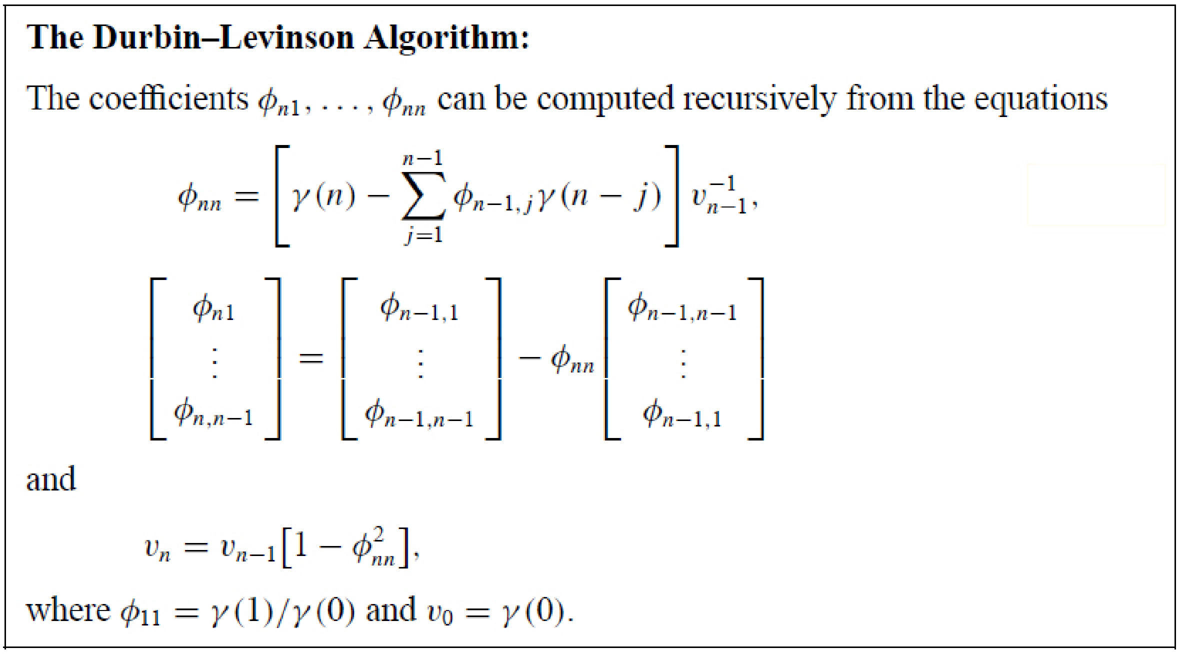
where

$$\boldsymbol{\phi}_n = \mathbf{\Gamma}_n^{-1} \boldsymbol{\gamma}_n, \quad \boldsymbol{\gamma}_n = (\gamma(1), \ldots, \gamma(n))'$$

and the corresponding mean squared error is

$$\nu_n := \mathrm{E}(X_{n+1} - P_n X_{n+1})^2 = \gamma(0) - \boldsymbol{\phi_n}' \boldsymbol{\gamma}_n.$$

A useful sufficient condition for nonsingularity of all the autocovariance matrices $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \ldots$ is $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$. The coefficients $\phi_{n1}, \ldots, \phi_{nn}$ can be computed recursively with the Durbin-Levinson algorithm (see Figure 15.1).



The Durbin–Levinson Algorithm:

The coefficients $\phi_{n1}, \ldots, \phi_{nn}$ can be computed recursively from the equations

$$\phi_{nn} = \left[\gamma(n) - \sum_{j=1}^{n-1} \phi_{n-1,j} \gamma(n-j) \right] v_{n-1}^{-1},$$

$$\begin{bmatrix} \phi_{n1} \\ \vdots \\ \phi_{n,n-1} \end{bmatrix} = \begin{bmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{bmatrix} - \phi_{nn} \begin{bmatrix} \phi_{n-1,n-1} \\ \vdots \\ \phi_{n-1,1} \end{bmatrix}$$

and

$$v_n = v_{n-1}[1 - \phi_{nn}^2],$$

where $\phi_{11} = \gamma(1)/\gamma(0)$ and $v_0 = \gamma(0)$.

Figure 15.1: Algorithm for estimating the parameters $\boldsymbol{\phi}_n = (\phi_{n1}, \ldots, \phi_{nn})$ in a pure autoregressive model. Source: Brockwell and Davis (2002), p. 70.

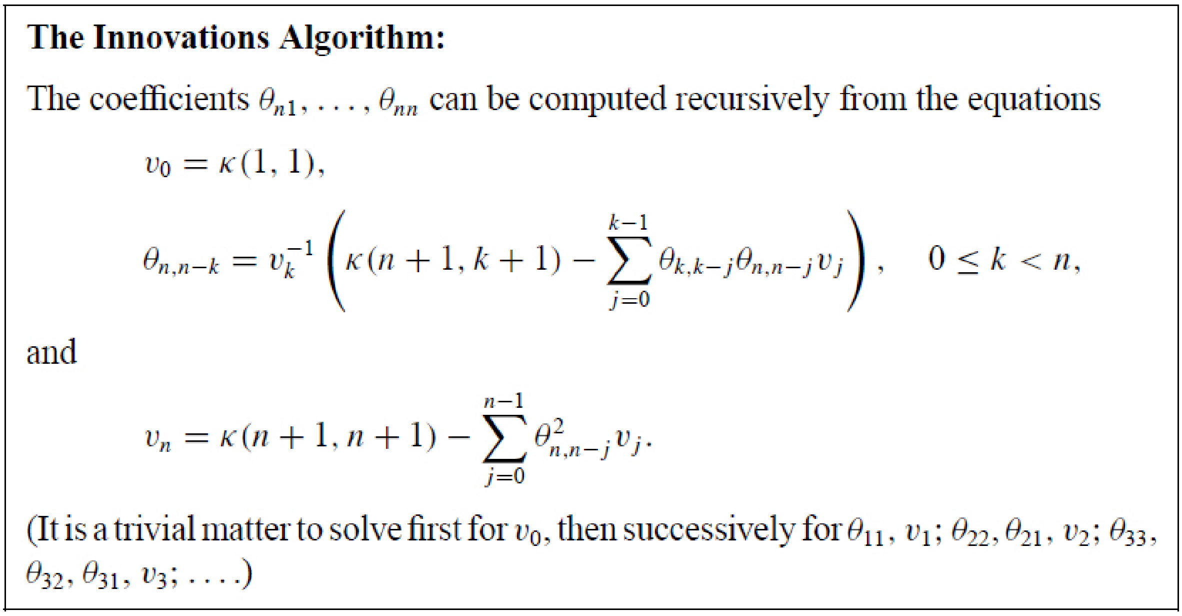
15.1.2 Innovations Algorithm

The recursive algorithm to be discussed in this section is applicable to all series with finite second moments, regardless of whether they are stationary or not. Its application, however, can be simplified in certain special cases.

Proposition 15.1.2. *Suppose that $\{X_t\}$ is a zero-mean series with $\mathrm{E}|X_t|^2 < \infty$ for each t and $\mathrm{E}(X_i X_j) = \kappa(i,j)$, where the matrix $[\kappa(i,j)]_{i,j=1}^{n}$ is non-singular for each $n = 1, 2, \ldots$, then the one-step predictor is given by*

$$\hat{X}_{n+1} = \begin{cases} 0, & n = 0, \\ \displaystyle\sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), & n \geq 1, \end{cases}$$

from which the one-step predictors $\hat{X}_1, \hat{X}_2, \ldots$ can be computed recursively once the coefficients θ_{ij} have been determined. The innovations algorithm (Figure 15.2) generates these coefficients and the mean squared errors $\nu_i = \mathrm{E}(X_{i+1} - \hat{X}_{i+1})^2$, starting from the covariances $\kappa(i,j)$.



The Innovations Algorithm:

The coefficients $\theta_{n1}, \ldots, \theta_{nn}$ can be computed recursively from the equations

$$v_0 = \kappa(1,1),$$

$$\theta_{n,n-k} = v_k^{-1}\left(\kappa(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right), \quad 0 \leq k < n,$$

and

$$v_n = \kappa(n+1,n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j.$$

(It is a trivial matter to solve first for v_0, then successively for θ_{11}, v_1; θ_{22}, θ_{21}, v_2; θ_{33}, θ_{32}, θ_{31}, v_3; ....)

Figure 15.2: Innovations algorithm for estimating the parameters $\boldsymbol{\theta}_n = (\theta_{n1}, \ldots, \theta_{nn})$. Source: Brockwell and Davis (2002), p. 73.

Example. If $\{X_t\}$ is the MA(1) process

$$X_t = Z_t + \theta Z_{t-1}, \qquad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2),$$

then $\kappa(i,j) = 0$ for $|i-j| > 1$, $\kappa(i,i) = \sigma^2(1+\theta^2)$ and $\kappa(i,i+1) = \theta\sigma^2$. Using the innovations algorithm (see Figure 15.2) we find

$$\begin{aligned} \nu_0 &= \sigma^2(1+\theta^2), \\ \theta_{n1} &= \sigma^2 \nu_{n-1}^{-1}\theta, \\ \theta_{nj} &= 0, \quad 2 \leq j \leq n, \end{aligned}$$

and

$$\nu_n = \sigma^2[1 + \theta^2 - \nu_{n-1}^{-1}\theta^2\sigma^2].$$

If we define $r_n = \nu_n/\sigma^2$, then we can write

$$\hat{X}_{n+1} = \frac{\theta(X_n - \hat{X}_n)}{r_{n-1}},$$

where $r_0 = 1 + \theta^2$ and $r_{n+1} = 1 + \theta^2 - \theta^2/r_n$.

15.2 Forecasting ARMA Processes

The innovations algorithm is a recursive method for forecasting second-order zero-mean processes that are not necessarily stationary. Proposition 15.1.2 can of course be applied directly to the prediction of the causal ARMA process,

$$\phi(B)X_t = \theta(B)Z_t, \qquad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

However a drastic simplification in the calculations can be achieved, if, instead of applying Proposition 15.1.2 to $\{X_t\}$, we apply it to the transformed process

$$W_t = \begin{cases} \sigma^{-1}X_t, & t = 1, \ldots, m, \\ \sigma^{-1}\phi(B)X_t, & t > m, \end{cases}$$

where $m = \max(p, q)$.

For notational convenience we define $\theta_0 := 1$, $\theta_j := 0$ for $j > q$ and assume that $p \geq 1$ and $q \geq 1$. The autocovariances $\kappa(i,j) = \mathrm{E}(W_iW_j)$, $i, j \geq 1$, are found from

$$\kappa(i,j) = \begin{cases} \sigma^{-2}\gamma_X(i-j), & 1 \leq i,j \leq m, \\ \sigma^{-2}\left[\gamma_X(i-j) - \sum_{r=1}^{p}\phi_r\gamma_X(r - |i-j|)\right], & \min(i,j) \leq m < \max(i,j) \leq 2m, \\ \sum_{r=0}^{q}\theta_r\theta_{r+|i-j|}, & \min(i,j) > m, \\ 0, & \text{otherwise.} \end{cases} \tag{15.7}$$

Applying the innovations algorithm to the process $\{W_t\}$ and replacing $(W_j - \hat{W}_j)$ by $\sigma^{-1}(X_j - \hat{X}_j)$ we finally obtain

$$\hat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n}\theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), & 1 \leq n < m = \max(p,q), \\ \phi_1X_n + \ldots + \phi_pX_{n+1-p} + \sum_{j=1}^{q}\theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), & n \geq m, \end{cases} \tag{15.8}$$

and

$$\mathrm{E}(X_{n+1} - \hat{X}_{n+1})^2 = \sigma^2\,\mathrm{E}(W_{n+1} - \hat{W}_{n+1})^2 = \sigma^2 r_n,$$

where θ_{nj} and r_n are found from the innovations algorithm (see Figure 15.2) with κ as in (15.7). Equations (15.8) determine the one-step predictors $\hat{X}_2, \hat{X}_3, \ldots$ recursively.

Example. Prediction of ARMA$(1,1)$ processes. Let

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}, \qquad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2),$$

and $|\phi| < 1$, then equations (15.8) reduce to the single equation

$$\hat{X}_{n+1} = \phi X_n + \theta_{n1}(X_n - \hat{X}_n), \quad n \geq 1.$$

We know that

$$\gamma_X(0) = \sigma^2\,\frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2}.$$

Substituting in (15.7) then gives, for $i, j \geq 1$,

$$\kappa(i,j) = \begin{cases} \dfrac{1 + 2\theta\phi + \theta^2}{1 - \phi^2}, & i = j = 1, \\ 1 + \theta^2, & i = j \geq 2, \\ \theta, & |i - j| = 1, i \geq 1, \\ 0, & \text{otherwise.} \end{cases}$$

With these values of $\kappa(i,j)$, the recursions of the innovations algorithm reduce to

$$r_0 = \frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2}, \qquad \theta_{n1} = \frac{\theta}{r_{n-1}}, \qquad r_n = 1 + \theta^2 - \frac{\theta^2}{r_{n-1}},$$

which can be solved quite explicitly.

16 Modeling with ARMA Processes

The determination of an appropriate ARMA(p, q) model to represent an observed stationary time series involves a number of interrelated problems. These include the choice of p and q (order selection) and the estimation of the mean, the coefficients $\{\phi_i, i = 1, \dots, p\}$, $\{\theta_i, i = 1, \dots, q\}$ and the white noise variance σ^2. Final selection of the model depends on a variety of goodness of fit tests[1], although it can be systematized to a large degree by use of criteria such as minimization of the AICC statistic as discussed in Section 16.2.2.

When p and q are known, good estimators of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can be found by imagining the data to be observations of a stationary Gaussian time series and maximizing the likelihood with respect to the $p + q + 1$ parameters $\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q$ and σ^2. The estimators obtained by this procedure are known as maximum likelihood estimators (Section 16.2). The algorithm used to determine the maximum likelihood estimators requires the specification of initial parameter values with which to begin the search. The closer the preliminary estimates are to the maximum likelihood estimates, the faster the search will generally be. To provide these initial values, a number of preliminary estimation algorithms are available (Section 16.1).

Section 16.3 deals with goodness of fit tests for the chosen model and Chapter 15 with the use of the fitted model for forecasting. In Section 16.2.2 we discuss the theoretical basis for some of the criteria used for order selection.

Remark. A good overview for determining an adequate ARMA model to a time series is given in Chapter 6 in Schlittgen and Streitberg (2001) (in German).

16.1 Preliminary Estimation

We shall consider different techniques for preliminary estimation of the parameters

$$\boldsymbol{\phi} = (\phi_1, \dots, \phi_p), \qquad \boldsymbol{\theta} = (\theta_1, \dots, \theta_q),$$

and σ^2 from observations $x_1, \dots, x_n$ of the causal ARMA(p, q) process defined by

$$\phi(B)X_t = \theta(B)Z_t, \qquad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2). \tag{16.1}$$

1. Pure autoregressive models:
    - Yule-Walker procedure (Section 16.1.1)
    - Burg's procedure (Section 16.1.2)

[1] Goodness of fit tests are used to judge the adequacy of a given statistical model.

For pure autoregressive models Burg's algorithm usually gives higher likelihoods than the Yule-Walker equations.

2. ARMA(p,q), $p,q > 0$:

   - Innovations algorithm (Section 16.1.3 and Brockwell and Davis (2002), pp. 154–156)
   - Hannan-Rissanen algorithm (Section 16.1.4)

   For pure moving-average models the innovations algorithm frequently gives slightly higher likelihoods than than the Hannan-Rissanen algorithm. For mixed models the Hannan-Rissanen algorithm is usually more successful in finding causal models which are required for initialization of the likelihood maximization.

16.1.1 Yule-Walker Equations[2]

For a pure autoregressive model the causality assumption allows us to write X_t in the form

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \quad \text{where} \quad \psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{1}{\phi(z)}. \tag{16.2}$$

Multiplying each side of (16.1) by X_{t-j}, $j = 0, \ldots, p$, taking expectations, and using (16.2) to evaluate the right-hand side of the first equation, we obtain the Yule-Walker equations

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

where $\boldsymbol{\Gamma}_p$ is the covariance matrix $[\gamma(i-j)]_{i,j=1}^{p}$ and $\boldsymbol{\gamma}_p = (\gamma(1), \ldots, \gamma(p))'$. These equations can be used to determine $\gamma(0), \ldots, \gamma(p)$ from σ^2 and $\boldsymbol{\phi}$.

If we replace the covariances $\gamma(j)$, $j = 0, \ldots, p$ by the corresponding sample covariances $\hat{\gamma}(j)$, we obtain a set of equations for the so-called sample Yule-Walker estimators $\hat{\boldsymbol{\phi}}$ and $\hat{\sigma}^2$ of $\boldsymbol{\phi}$ and σ^2, namely,

$$\widehat{\boldsymbol{\Gamma}}_p \hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\gamma}}_p \tag{16.3}$$

and

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}' \hat{\boldsymbol{\gamma}}_p \tag{16.4}$$

where

$$\widehat{\boldsymbol{\Gamma}}_p = [\hat{\gamma}(i-j)]_{i,j=1}^{p} \quad \text{and} \quad \hat{\boldsymbol{\gamma}}_p = (\hat{\gamma}(1), \ldots, \hat{\gamma}(p))'.$$

If $\hat{\gamma}(0) > 0$, then $\widehat{\boldsymbol{\Gamma}}_m$ is nonsingular for every $m = 1, 2, \ldots$, so we can rewrite equations (16.3) and (16.4) in the following form:

[2] Gibert Walker 1868-1958, George Udny Yule 1871-1951.

Definition 16.1.1 (Sample Yule-Walker equations)**.**

$$\hat{\boldsymbol{\phi}} = \left(\hat{\phi}_1, \dots, \hat{\phi}_p\right)' = \widehat{\mathbf{R}}_p^{-1} \hat{\boldsymbol{\rho}}_p$$

and

$$\hat{\sigma}^2 = \hat{\gamma}(0) \left(1 - \hat{\boldsymbol{\rho}}_p' \widehat{\mathbf{R}}_p^{-1} \hat{\boldsymbol{\rho}}_p\right),$$

where

$$\hat{\boldsymbol{\rho}}_p = (\hat{\rho}(1), \dots, \hat{\rho}(p))' = \hat{\boldsymbol{\gamma}}_p / \hat{\gamma}(0) \quad \text{and} \quad \widehat{\mathbf{R}}_p = \widehat{\boldsymbol{\Gamma}}_p / \hat{\gamma}(0).$$

Proposition 16.1.2 (Large-sample distribution of Yule-Walker equations)**.** *For a large sample from an* AR(p) *process*

$$n^{1/2}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \sim \mathrm{N}_p(\mathbf{0}, \sigma^2 \boldsymbol{\Gamma}_p^{-1}).$$

Order selection In practice we do not know the true order of the model generating the data. In fact, it will usually be the case that there is no true AR model, in which case our goal is simply to find one that represents the data optimally in some sense. Two useful techniques for selecting an appropriate AR model are given below:

- Some guidance in the choice of order is provided by a large-sample result which states that if $\{X_t\}$ is A causal AR(p) process with $\{Z_t\} \sim \mathrm{IID}(0, \sigma^2)$ and if we fit a model with order $m > p$ using the Yule-Walker equations, then the last component, $\hat{\phi}_{mm}$, of the vector $\hat{\phi}_m$ is approximately normally distributed with mean 0 and variance $1/n$. Notice that $\hat{\phi}_{mm}$ is exactly the sample partial autocorrelation at lag m.

  Now, we already know that for an AR(p) process the partial autocorrelation function ϕ_{mm}, $m > p$, are zero. Therefore, if an AR(p) model is appropriate for the data, then the values $\hat{\phi}_{kk}$, $k > p$, should be compatible with observations from $N(0, 1/n)$. In particular, for $k > p$, $\hat{\phi}_{kk}$ will fall between the bounds $\pm 1.96 n^{-1/2}$ with probability close to 0.95. This suggests using as a preliminary estimator of p the smallest value m such that $|\hat{\phi}_{kk}| < 1.96 n^{-1/2}$ for $k > m$.

- A more systematic approach to order selection is to find the values of p and ϕ_p that minimize the AICC statistic

  $$\mathrm{AICC} = -2 \ln L\left(\boldsymbol{\phi}_p, \frac{S(\boldsymbol{\phi}_p)}{n}\right) + \underbrace{\frac{2(p+1)n}{n-p-2}}_{\text{penalty factor}},$$

  where L is the Gaussian likelihood defined in (16.7) and S is defined in (16.8).

Proposition 16.1.3. *The fitted Yule-Walker* AR(m) *model is*

$$X_t - \hat{\phi}_{m1} X_{t-1} - \ldots - \hat{\phi}_{mm} X_{t-m} = Z_t, \qquad \{Z_t\} \sim \mathrm{WN}(0, \hat{\nu}_m),$$

where

$$\hat{\boldsymbol{\phi}}_m = \left(\hat{\phi}_{m1}, \ldots, \hat{\phi}_{mm}\right)' = \widehat{\mathbf{R}}_m^{-1} \hat{\boldsymbol{\rho}}_m$$

and

$$\hat{\nu}_m = \hat{\gamma}(0) \left(1 - \hat{\boldsymbol{\rho}}_m' \widehat{\mathbf{R}}_m^{-1} \hat{\boldsymbol{\rho}}_m\right).$$

Remark. For both approaches to order selection we need to fit AR models of gradually increasing order to our given data. The problem of solving the Yule-Walker equations with gradually increasing orders has already been encountered in a slightly different context (see Section 15.1.1. Here we can use exactly the same scheme, the Durbin-Levinson algorithm, to solve the Yule-Walker equations (16.3) and (16.4).

U nder the assumption that the order p of the fitted model is the correct value, we can use the asymptotic distribution of $\hat{\boldsymbol{\phi}}_p$ to derive approximate large-sample confidence regions for the true coefficient vector $\boldsymbol{\phi}_p$ and for its individual components ϕ_{pj}. Thus, for large sample-size n the region

$$\left\{\boldsymbol{\phi} \in \mathbb{R}^p : \left(\hat{\boldsymbol{\phi}}_p - \boldsymbol{\phi}\right)' \widehat{\boldsymbol{\Gamma}}_p \left(\hat{\boldsymbol{\phi}}_p - \boldsymbol{\phi}\right) \leq n^{-1} \hat{\nu}_p \chi^2_{1-\alpha}(p)\right\}$$

contains $\boldsymbol{\phi}_p$ with probability close to $(1-\alpha)$.

Similarly, if $\hat{\nu}_{jj}$ is the jth diagonal element of $\hat{\nu}_p \widehat{\boldsymbol{\Gamma}}_p^{-1}$, then for large n the interval bounded by

$$\hat{\phi}_{pj} \pm \Phi_{1-\alpha/2}\, n^{-1/2}\, \hat{\nu}_{jj}^{1/2}$$

contains ϕ_{pj} with probability close to $(1-\alpha)$.

16.1.2 Burg's algorithm (~ 1967)

The Yule-Walker coefficients $\hat{\phi}_{pi}, \ldots, \hat{\phi}_{pp}$ are precisely the coefficients of the best linear predictor of X_{p+1} in terms of $\{X_p, \ldots, X_1\}$ under the assumption that the autocorrelation function of $\{X_t\}$ coincides with the sample autocorrelation function at lags $1, \ldots, p$.

Burg's algorithm estimates the partial autocorrelation function $\{\phi_{11}, \phi_{22} \ldots\}$ by successively minimizing sums of squares of forward and backward one-step prediction errors with respect to the coefficients ϕ_{ii}. Given observations $\{x_1, \ldots, x_n\}$ of a stationary zero-mean time series $\{X_t\}$ we define $u_i(t)$, $t = i+1, \ldots, n$, $0 \leq i < n$, to be the difference between $x_{n+1+i-t}$ and the best linear estimate of $x_{n+1+i-t}$ in terms of the preceding i observations. Similarly, we define $\nu_i(t)$, $t = i+1, \ldots, n$, $0 \leq i < n$, to be the difference between x_{n+1-t} and the best linear estimate of x_{n+1-t} in terms of the subsequent

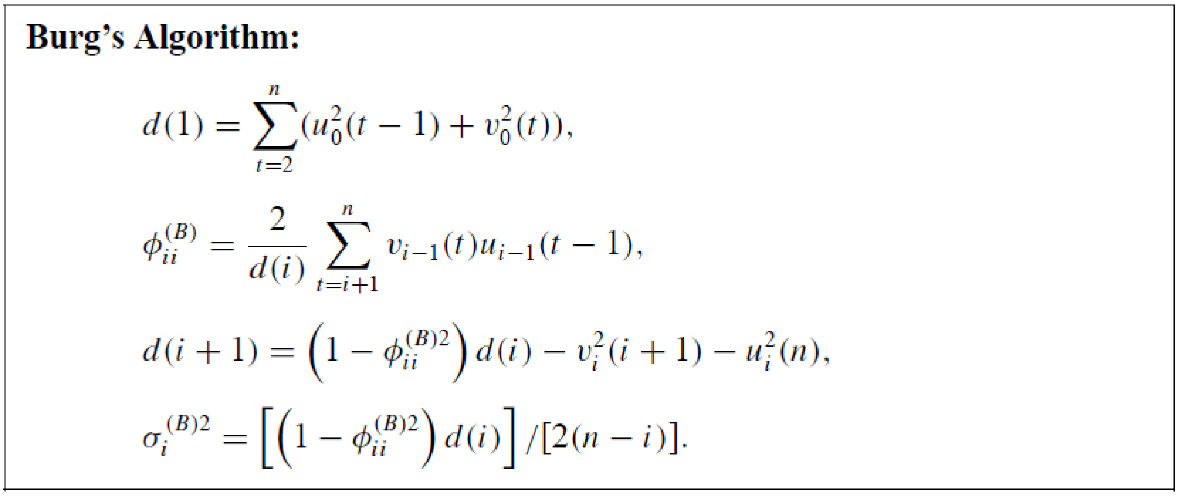
i observations. Then it can be shown that the forward and backward prediction errors $\{u_i(t)\}$ and $\{\nu_i(t)\}$ satisfy

$$u_0(t) = \nu_0(t) = x_{n+1-t},$$
$$u_i(t) = u_{i-1}(t-1) - \phi_{ii}\nu_{i-1}(t)$$

and

$$\nu_i(t) = \nu_{i-1}(t) - \phi_{ii}u_{i-1}(t-1).$$

The calculation of the estimates of ϕ_{pp} described above and σ_p^2 is equivalent to solving the following recursions ($i = 1, \ldots, p$):



Burg's Algorithm:

$$d(1) = \sum_{t=2}^{n} (u_0^2(t-1) + v_0^2(t)),$$

$$\phi_{ii}^{(B)} = \frac{2}{d(i)} \sum_{t=i+1}^{n} v_{i-1}(t)u_{i-1}(t-1),$$

$$d(i+1) = \left(1 - \phi_{ii}^{(B)2}\right) d(i) - v_i^2(i+1) - u_i^2(n),$$

$$\sigma_i^{(B)2} = \left[\left(1 - \phi_{ii}^{(B)2}\right) d(i)\right] / [2(n-i)].$$

Figure 16.1: Algorithm for estimating the parameters $\boldsymbol{\phi}$ in a pure autoregressive model. Source: Brockwell and Davis (2002), p. 148.

Remark. The large-sample distribution of the estimated coefficients for the Burg estimators of the coefficients of an AR(p) process is the same as for the Yule-Walker estimators, namely $\mathrm{N}(\boldsymbol{\phi}, n^{-1}\sigma^2\boldsymbol{\Gamma}_p^{-1}$. Although the two methods give estimators with the same large-sample distributions, for finite sample sizes the Burg model usually has smaller estimated white noise variance and larger Gaussian likelihood.

For further details and examples see Brockwell and Davis (2002) pp. 147–150.

16.1.3 Innovations algorithm

Just as we can fit autoregressive models of order $1, 2, \ldots$ to the data $\{x_1, \ldots, x_n\}$ by applying the Durbin-Levinson algorithm to the sample autocovariances, we can also fit moving average models

$$X_t = Z_t + \hat{\theta}_{m1}Z_{t-1} + \ldots + \hat{\theta}_{mm}Z_{t-m}, \qquad \{Z_t\} \sim \mathrm{WN}(0, \hat{\nu}_m)$$

of orders $m = 1, 2 \ldots$ by means of the innovations algorithm (see Figure 15.2, p. 15-4).

Proposition 16.1.4. *The fitted innovations* MA(m) *model is*

$$X_t = Z_t + \hat{\theta}_{m1} Z_{t-1} + \ldots + \hat{\theta}_{mm} Z_{t-m}, \qquad \{Z_t\} \sim \mathrm{WN}(0, \hat{\nu}_m),$$

where $\hat{\boldsymbol{\theta}}_m = (\hat{\theta}_{m1}, \ldots, \hat{\theta}_{mm})$ *and* $\hat{\nu}_m$ *are obtained from the innovations algorithm with the autocovariance function replaced by the sample autocovariance function.*

Order selection Three useful techniques for selecting an appropriate MA model are given below. The third is more systematic and extends beyond the narrow class of pure moving-average models.

- We know that for an MA(q) process the autocorrelations $\rho(m)$, $m > q$, are zero. Moreover, the sample autocorrelation $\hat{\rho}_m$, $m > q$, is approximately normally distributed with mean $\rho(m) = 0$ and variance

$$n^{-1}\left[1 + 2\rho^2(1) + \ldots 2\rho^2(q)\right].$$

  This result enables us to use the graph of $\hat{\rho}(m)$, $m = 1, 2, \ldots$, both to decide whether or not a given data set can be plausibly modeled by a moving-average process and also to obtain a preliminary estimate of the order q as the smallest value of m such that $\hat{\rho}(k)$ is not significantly different from zero for all $k > m$. For practical purposes $\hat{\rho}(k)$ is compared to $1.96n^{-1/2}$ in absolute value.

- We examine the coefficient vectors $\hat{\boldsymbol{\theta}}_m$, $m = 1, 2, \ldots$ to be able not only to assess the appropriateness of a moving-average model and to estimate its order q, but also to obtain preliminary estimates $\hat{\theta}_{m1}, \ldots, \hat{\theta}_{mq}$ of the coefficients. By inspecting the estimated coefficients $\hat{\theta}_{m1}, \ldots, \hat{\theta}_{mm}$ for $m = 1, 2, \ldots$ and the ratio of each coefficient estimate $\hat{\theta}_{mj}$ to

$$1.96\,\sigma_j = 1.96\,n^{-1/2}\left(\sum_{i=0}^{j-1} \hat{\theta}_{mi}^2\right)^{1/2},$$

  we can see which of the coefficient estimates are most significantly different from zero, estimate for order of the model to be fitted as the largest j for which the ratio is larger than 1 in absolute value.

- As for autoregressive models, a more systematic approach to order selection for moving-average models is to find the values of q and $\hat{\boldsymbol{\theta}}_q = (\hat{\theta}_{m1}, \ldots, \hat{\theta}_{mq})'$ that minimize the AICC statistic

$$\mathrm{AICC} = -2\ln L\left(\boldsymbol{\theta}_q, \frac{S(\boldsymbol{\theta}_q)}{n}\right) + \frac{2(q+1)n}{n-q-2},$$

  where L is the Gaussian likelihood defined in (16.7) and S is defined in (16.8).

16.1.4 Hannan-Rissanen algorithm (1982)

The defining equations for a causal AR(p) model have the form of a linear regression model with coefficient vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$. This suggests the use of simple least squares regression for obtaining preliminary parameter estimates when $q = 0$. Application of this technique when $q > 0$ is complicated by the fact that in the general ARMA(p, q) equations $\{X_t\}$ is regressed not only on $X_{t-1}, \ldots, X_{t-p}$, but also on the unobserved quantities $Z_{t-1}, \ldots, Z_{t-q}$. Nevertheless, it is still possible to apply least squares regression to the estimation of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ by first replacing the unobserved quantities $Z_{t-1}, \ldots, Z_{t-q}$ by estimated values $\hat{Z}_{t-1}, \ldots, \hat{Z}_{t-q}$. The parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are then estimated by regressing X_t onto $X_{t-1}, \ldots, X_{t-p}, \hat{Z}_{t-1}, \ldots, \hat{Z}_{t-q}$. These are the main steps in the Hannan-Rissanen estimation procedure, which we now describe in more detail.

Step 1: A high-order AR(m) model (with $m > \max(p, q)$) is fitted to the data using the Yule-Walker estimates of Section 16.1.1. If $(\hat{\phi}_{m1}, \ldots, \hat{\phi}_{mm})'$ is the vector of estimated coefficients, then the estimated residuals are computed from the equations

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

Step 2: Once the estimated residuals $\hat{Z}_t$, $t = m+1, \ldots, n$, have been computed as in Step 1, the vector of parameters $\boldsymbol{\beta} = (\boldsymbol{\phi}', \boldsymbol{\theta}')$ is estimated by least squares linear regression of X_t onto $(X_{t-1}, \ldots, X_{t-p}, \hat{Z}_{t-1}, \ldots, \hat{Z}_{t-q})$, $t = m+1+q, \ldots, n$, i.e., by minimizing

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

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

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{X}_n,$$

where $\boldsymbol{X}_n = (X_{m+1+q}, \ldots, X_n)'$ and $\mathbf{Z}$ is the $(n-m-q) \times (p+q)$ matrix

$$\mathbf{Z} = \begin{pmatrix} X_{m+q} & X_{m+q-1} & \cdots & X_{m+q+1-p} & \hat{Z}_{m+q} & \cdots & \hat{Z}_{m+1} \\ X_{m+q+1} & X_{m+q} & \cdots & X_{m+q+2-p} & \hat{Z}_{m+q+1} & \cdots & \hat{Z}_{m+2} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ X_{n-1} & X_{n-2} & \cdots & X_{n-p} & \hat{Z}_{n-1} & \cdots & \hat{Z}_{n-q} \end{pmatrix}.$$

The Hannan-Rissanen estimate of the white noise variance is

$$\hat{\sigma}^2_{HR} = \frac{S(\hat{\boldsymbol{\beta}})}{n-m-q}.$$

16.2 Maximum Likelihood Estimation

Suppose that $\{X_t\}$ is a Gaussian time series with mean zero and autocovariance function $\kappa(i,j) = \mathrm{E}(X_i X_j)$. Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)'$ and let $\hat{\boldsymbol{X}}_n = (\hat{X}_1, \ldots, \hat{X}_n)'$, where $\hat{X}_1 = 0$ and $\hat{X}_j = \mathrm{E}(X_j | X_1, \ldots, X_{j-1}) = P_{j-1}X_j$, $j \geq 2$. Let $\boldsymbol{\Gamma}_n$ denote the covariance matrix $\boldsymbol{\Gamma}_n = \mathrm{E}(\boldsymbol{X}_n \boldsymbol{X}_n')$, and assume that $\boldsymbol{\Gamma}_n$ is nonsingular.

The likelihood of $\boldsymbol{X}_n$ is

$$L(\boldsymbol{\Gamma}_n) = (2\pi)^{-n/2} (\det \boldsymbol{\Gamma}_n)^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{X}_n' \boldsymbol{\Gamma}_n^{-1} \boldsymbol{X}_n\right). \tag{16.5}$$

The direct calculation of $\det \boldsymbol{\Gamma}_n$ and $\boldsymbol{\Gamma}_n^{-1}$ can be avoided by expressing this in terms of the one-step predictors $\hat{X}_j$, and their mean squared errors ν_{j-1}, $j = 1, \ldots, n$, both of which are calculated recursively from the innovations algorithm.

It can be shown, that

$$\boldsymbol{X}_n' \boldsymbol{\Gamma}_n^{-1} \boldsymbol{X}_n = \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{\nu_{j-1}},$$

and

$$\det(\boldsymbol{\Gamma}_n) = \nu_0 \cdot \ldots \cdot \nu_{n-1}.$$

The likelihood (16.5) of the vector $\boldsymbol{X}_n$ therefore reduces to

$$L(\boldsymbol{\Gamma}_n) = (2\pi)^{-n/2} (\nu_0 \cdot \ldots \cdot \nu_{n-1})^{-1/2} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{\nu_{j-1}}\right). \tag{16.6}$$

Remark. Even if $\{X_t\}$ is not Gaussian, it still makes sense to estimate the unknown parameters $\boldsymbol{\beta} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots \theta_q)'$ in such a way as to maximize (16.6).

A justification for using maximum Gaussian likelihood estimators of ARMA coefficients is that the large-sample distribution of the estimators is the same for $\{Z_t\} \sim \mathrm{IID}(0, \sigma^2)$, regardless of whether or not $\{Z_t\}$ is Gaussian.

16.2.1 Estimation for ARMA processes

Suppose now that $\{X_t\}$ is a causal ARMA(p,q) process. Applying the innovations algorithm ((15.8)) we find the one-step predictors $\hat{X}_{i+1}$ and the mean squared errors $\mathrm{E}(X_{i+1} - \hat{X}_{i+1}) = \sigma^2 r_i$. Substituting in the general expression (16.6), we find the Gaussian likelihood of the vector of observations $\boldsymbol{X}_n = (X_1, \ldots, X_n)'$.

Proposition 16.2.1. *The Gaussian likelihood for an* ARMA(p,q) *process:*

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n r_0 \cdots r_{n-1}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{r_{j-1}}\right) \tag{16.7}$$

with $r_j = \mathrm{E}(X_{j+1} - \hat{X}_{j+1})^2/\sigma^2$.

Differentiating $\ln L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$ *partially with respect to* σ^2 *and noting that* $\hat{X}_j$ *and* r_j *are independent of* σ^2, *we find that the maximum likelihood estimators* $\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}$, *and* $\hat{\sigma}^2$ *satisfy*

$$\hat{\sigma}^2 = n^{-1} S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$$

where

$$S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{r_{j-1}} \tag{16.8}$$

and $\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}$ *are the values of* $\boldsymbol{\phi}, \boldsymbol{\theta}$ *that minimize*

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}) = \ln(n^{-1} S(\boldsymbol{\phi}, \boldsymbol{\theta})) + \frac{1}{n} \sum_{j=1}^{n} \ln r_{j-1}. \tag{16.9}$$

Remark. Minimization of $l(\boldsymbol{\phi}, \boldsymbol{\theta})$ must be done numerically. The search procedure may be greatly accelerated if we begin with parameter values $\boldsymbol{\phi}_0, \boldsymbol{\theta}_0$ which are close to the minimum of l. It is for this reason that simple, reasonably good preliminary estimates of $\boldsymbol{\phi}, \boldsymbol{\theta}$ are important (see Section 16.1). It is essential to begin the search with a causal parameter vector $\boldsymbol{\phi}_0$ since causality is assumed in the computation of $l(\boldsymbol{\phi}, \boldsymbol{\theta})$.

16.2.2 Order selection

Let $\{X_t\}$ denote the mean-corrected transformed series. The problem now is to find the most satisfactory ARMA(p, q) model to represent $\{X_t\}$. If p and q were known in advance this would be a straightforward application of the estimation techniques. However this is usually not the case, so that it becomes necessary also to identify appropriate values for p and q.

It might appear at first sight that the higher the values of p and q chosen, the better the fitted model will be. However we must beware of the danger of overfitting, i.e. of tailoring the fit too closely to the particular numbers observed.

Criteria have been developed, which attempt to prevent overfitting by effectively assigning a cost to the introduction of each additional parameter.

We choose a biased-corrected form of the Akaike's AIC criterion, defined for an ARMA(p, q) model with coefficients $\boldsymbol{\phi}_p$ and $\boldsymbol{\theta}_q$, by

$$\mathrm{AICC} = -2 \ln L\left(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q, \frac{S(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q)}{n}\right) + \frac{2(p+q+1)n}{n-p-q-2},$$

where $L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$ is the likelihood of the data under the Gaussian ARMA model (16.7) with parameters $(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$ and $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ is the residual sum of squares (16.8). We select the model that minimizes the value of AICC. Intuitively one can think of $2(p + q + 1)n/(n-p-q-2)$ as a penalty term to discourage over-parameterization. Once a model has been found which minimizes the AICC value, it must then be checked for goodness of fit (essentially by checking that the residuals are like white noise) as discussed in Section 16.3.

Identification of Mixed Models

The identification of a pure autoregressive or moving average process is reasonably straightforward using the sample autocorrelation and partial autocorrelation functions, the preliminary estimators $\hat{\boldsymbol{\phi}}_m$ and $\hat{\boldsymbol{\theta}}_m$ and the AICC. On the other hand, for ARMA(p,q) processes with p and q both non-zero, the sample ACF and PACF are much more difficult to interpret. We therefore search directly for values of p and q such that the AICC is minimum. The search can be carried out in a variety of ways, e.g. by trying all (p,q) values such that $p+q=1$, then $p+q=2$, etc., or alternatively by using the following steps:

i) Use maximum likelihood estimation to fit ARMA processes of orders $(1,1), (2,2), \ldots$, to the data, selecting the model which gives the smallest value of the AICC.

ii) Starting from the minimum-AICC ARMA(p,p) model, eliminate one or more coefficients (guided by the standard errors of the estimated coefficients), maximize the likelihood for each reduced model and compute the AICC value.

iii) Select the model with smallest AICC value.

16.3 Diagnostic Checking

Typically the goodness of fit of a statistical model to a set of data is judged by comparing the observed values with the corresponding predicted values obtained from the fitted model. If the fitted model is appropriate, then the residuals should behave in a manner that is consistent with the model.

When we fit an ARMA(p,q) model to a given series we determine the maximum likelihood estimators $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\sigma}^2$ of the parameters $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and σ^2. In the course of this procedure the predicted values $\hat{X}_t(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}})$ of X_t based on $X_1,\ldots,X_{t-1}$ are computed for the fitted model. The residuals are then defined, in the notation of Section 15.2, by

$$\hat{W}_t = \frac{(X_t - \hat{X}_t(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}}))}{\left(r_{t-1}(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}})\right)^{1/2}}, \qquad t = 1,\ldots,n.$$

The properties of the residuals $\{\hat{W}_t\}$ should be similar to those of the white noise sequence

$$W_t = \frac{(X_t - \hat{X}_t(\boldsymbol{\phi},\boldsymbol{\theta}))}{(r_{t-1}(\boldsymbol{\phi},\boldsymbol{\theta}))^{1/2}}, \qquad t = 1,\ldots,n.$$

Moreover, $\mathrm{E}(W_t(\boldsymbol{\phi},\boldsymbol{\theta}) - Z_t)^2$ is small for large t, so that properties of the residuals $\{\hat{W}_t\}$ should reflect those of the white noise sequence $\{Z_t\}$ generating the underlying ARMA(p,q) process. In particular the sequence $\{\hat{W}_t\}$ should be approximately

- uncorrelated if $\{Z_t\} \sim \mathrm{WN}(0,\sigma^2)$,

- independent if $\{Z_t\} \sim \mathrm{IID}(0, \sigma^2)$, and
- normally distributed if $\{Z_t\} \sim \mathrm{N}(0, \sigma^2)$.

The following diagnostic checks are all based on the expected properties of the residuals under the assumption that the fitted model is correct and that $\{Z_t\} \sim \mathrm{IID}(0, \sigma^2)$. They are the same tests introduced in Section 12.6.

a) Graph of $\{\hat{W}_t\}$: If the fitted model is appropriate, then the graph of the residuals $\{\hat{W}_t,\ t = 1, \ldots, n\}$ should resemble that of a white noise sequence.

b) The sample autocorrelation function of the residuals $\{\hat{W}_t\}$: We know from Section 12.6 that for large n the sample autocorrelations of an iid sequence $Y_1, \ldots, Y_n$ with finite variance are approximately iid with distribution $\mathrm{N}(0, 1/n)$. We can therefore test whether or not the observed residuals are consistent with iid noise by examining the sample autocorrelations of the residuals and rejecting the iid noise hypothesis if more than two or three out of 40 fall outside the bounds $\pm 1.96/\sqrt{n}$ or if one falls far outside the bounds.

c) Tests for randomness of the residuals: see Section 12.6.

d) Check for normality: A rough check for normality is provided by visual inspection of the histogram of the residuals or by a Gaussian-QQ-Plot of the residuals. The Jarque-Bera statistic

$$n\left(\frac{m_3^2}{6m_2^3} + \frac{\left(\frac{m_4}{m_2^3} - 3\right)^2}{24}\right),$$

where

$$m_r = \frac{1}{n}\sum_{j=1}^{n}(Y_i - \overline{Y})^r,$$

is distributed asymptotically as $\chi^2(2)$ if $\{Y_t\} \sim \mathrm{IID}\,\mathrm{N}(\mu, \sigma^2)$. This hypothesis is rejected if the statistic is sufficiently large.

Definition: $\{X_t\}$ is an ARMA(p,q) process if $\{X_t\}$ is stationary and if for every t,

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q},$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$ and the polynomials $\phi(z)$ and $\theta(z)$ have no common factors.
If $\phi(z) \equiv 1$, then the process is said to be a moving-average process of order q or MA(q).
If $\theta(z) \equiv 1$ then the process is said to be an autoregressive process of order p or AR(p).

	AR(p)	MA(q)	ARMA(p,q)
Model	$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t$ $\phi(B)X_t = Z_t$	$X_t = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}$ $X_t = \theta(B) Z_t$	$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}$ $\phi(B)X_t = \theta(B)Z_t$
Stationarity	AR(1): $\lvert\phi\rvert < 1$	per definition	if and only if $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p \neq 0$ for all $\lvert z\rvert = 1$
Causality			
Definition			there exist constants $\{\psi_j\}$ such that $\sum_{j=0}^{\infty} \lvert\psi_j\rvert < \infty$ and $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ for all t
Proposition		per definition	if and only if $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p \neq 0$ for all $\lvert z\rvert \leq 1$
Coefficients			The coefficients $\{\psi_j\}$ are determined by the relation $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z), \qquad \lvert z\rvert \leq 1$
Invertibility			
Definition			there exist constants $\{\pi_j\}$ such that $\sum_{j=0}^{\infty} \lvert\pi_j\rvert < \infty$ and $Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$ for all t
Proposition	per definition		if and only if $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q \neq 0$ for all $\lvert z\rvert \leq 1$
Coefficients			The coefficients $\{\pi_j\}$ are determined by the relation $\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \phi(z)/\theta(z), \qquad \lvert z\rvert \leq 1$
ACF	AR(1): $\rho(h) = \phi^{\lvert h\rvert}$ AR(p): decays infinite	$\rho(h) = \begin{cases} \frac{\sum_{j=0}^{q-\lvert h\rvert} \theta_j \theta_{j+\lvert h\rvert}}{\sum_{j=0}^{q} \theta_j^2} & \text{if } \lvert h\rvert \leq q \\ 0 & \text{if } \lvert h\rvert > q \end{cases}$ finite	infinite (damped exponentials and/or sine waves after first $q-p$ lags)
PACF	$\alpha(h) = \begin{cases} \phi_p & \text{if } h = p \\ 0 & \text{if } h > p \end{cases}$ finite	MA$(1): \alpha(h) = -\frac{(-\theta)^h(1-\theta^2)}{1-\theta^{2(h+1)}}$ MA(q): decays infinite	infinite (damped exponentials and/or sine waves after first $p-q$ lages)

Figure 16.2: Summary of the ARMA time series models

17 Nonstationary and Seasonal Time Series Models

If the data (a) exhibit no apparent deviations from stationarity and (b) have a rapidly decreasing autocovariance function, we attempt to fit an ARMA model to the mean-corrected data using the techniques developed in Chapter 16. Otherwise, we look first for a transformation of the data that generates a new series with the properties (a) and (b). This can frequently be achieved by differencing, leading us to consider the class of ARIMA (autoregressive integrated moving-average) models.

17.1 ARIMA Models

Definition 17.1.1. If d is a non-negative integer, then $\{X_t\}$ is an ARIMA(p, d, q) process if

$$Y_t := (1 - B)^d X_t$$

is a causal ARMA(p, q) process.

Remark. This definition means that $\{X_t\}$ satisfies a difference equation of the form

$$\phi^\star(B)X_t \equiv \phi(B)(1 - B)^d X_t = \theta(B)Z_t, \qquad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2), \tag{17.1}$$

where $\phi(z)$ and $\theta(z)$ are polynomials of degrees p and q, respectively, and $\phi(z) \neq 0$ for $|z| \leq 1$. The polynomial $\phi^\star(z)$ has a zero of order d at $z = 1$. The process $\{X_t\}$ is stationary if and only if $d = 0$, in which case it reduces to an ARMA(p, q) process.

Remark. Notice that if $d \geq 1$, we can add an arbitrary polynomial trend of degree $(d-1)$ to $\{X_t\}$ without violating the difference equation (17.1). ARIMA models are therefore useful for representing data with trend.

17.2 SARIMA Models

We have already seen how differencing the series $\{X_t\}$ at lag s is a convenient way of eliminating a seasonal component of period s. If we fit an ARMA(p, q) model $\phi(B)Y_t = \theta(B)Z_t$ to the differenced series $Y_t = (1 - B^s)X_t$, then the model for the original series is $\phi(B)(1 - B^s)X_t = \theta(B)Z_t$. This is a special case of the general seasonal ARIMA (SARIMA) model defined as follows.

The background idea of the application of SARIMA models can be described in three steps (see Schlittgen and Streitberg (2001)).

1. The existence of seasonal effects means, that an observation of a specific month depends on the observations of the same month for past years (e.g. January 2010 depends on January 2009, January 2008 etc.). Considering the observations of the

same month (data with lag $s = 12$), these dependency can be modeled with an ARIMA process: $\Phi(B^s)(1 - B^s)^D X_t = \Theta(B^s)U_t$, where $\{X_t\}$ is the original time series, e.g. the Basel monthly mean temperature time series from 1900 to 2010.

2. U_t is not a White-noise process since there is also a dependency of the temperatures between successive months (e.g. January 2010 depends on December 2009, November 2009 etc.). Therefore these non-seasonal patterns are also modeled with an ARIMA process: $\phi(B)(1 - B)^d U_t = \theta(B)Z_t$, where Z_t is a white-noise process.

3. Multiplying the first equation with $\phi(B)(1-B)^d$ and applying the second equation we get

$$\phi(B)(1 - B)^d \Phi(B^s)(1 - B^s)^D X_t = \phi(B)(1 - B)^d \Theta(B^s)U_t = \theta(B)\Theta(B^s)Z_t.$$

Definition 17.2.1. If d and D are non-negative integers then $\{X_t\}$ is a

$$\text{seasonal ARIMA}(p, d, q) \times (P, D, Q)_s \text{ process}$$

with period s if the differenced series

$$Y_t := (1 - B)^d (1 - B^s)^D X_t$$

is a causal ARMA(p, q) process defined by

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t, \qquad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2), \tag{17.2}$$

where

$$\begin{aligned}
\phi(z) &= 1 - \phi_1 z - \ldots - \phi_p z^p,\\
\Phi(z) &= 1 - \Phi_1 z - \ldots - \Phi_P z^P,\\
\theta(z) &= 1 + \theta_1 z + \ldots + \theta_q z^q,\\
\Theta(z) &= 1 + \Theta_1 z + \ldots + \Theta_Q z^Q.
\end{aligned}$$

Example (Basel, p. 1-6)**.** Since the autocorrelation function of Figure 12.10, p. 12-19, showed among other things a negative correlation at lag 12 after differencing the Basel monthly mean temperature time series with $(1-B)(1-B^{12})$, we will now use a SARIMA process to find an adequate model. Applying an ARIMA$(1, 1, 1) \times (1, 1, 1)_{12}$ process results in Figure 17.1 which shows that the residuals are white noise and therefore the model is appropriate. Finally Table 17.1 shows the results of the parameter estimation.

Remark. Note that the process $\{Y_t\}$ is causal if and only if $\phi(z) \neq 0$ and $\Theta(z) \neq 0$ for $|z| \leq 1$. In applications D is rarely more than one, and P and Q are typically less than three.

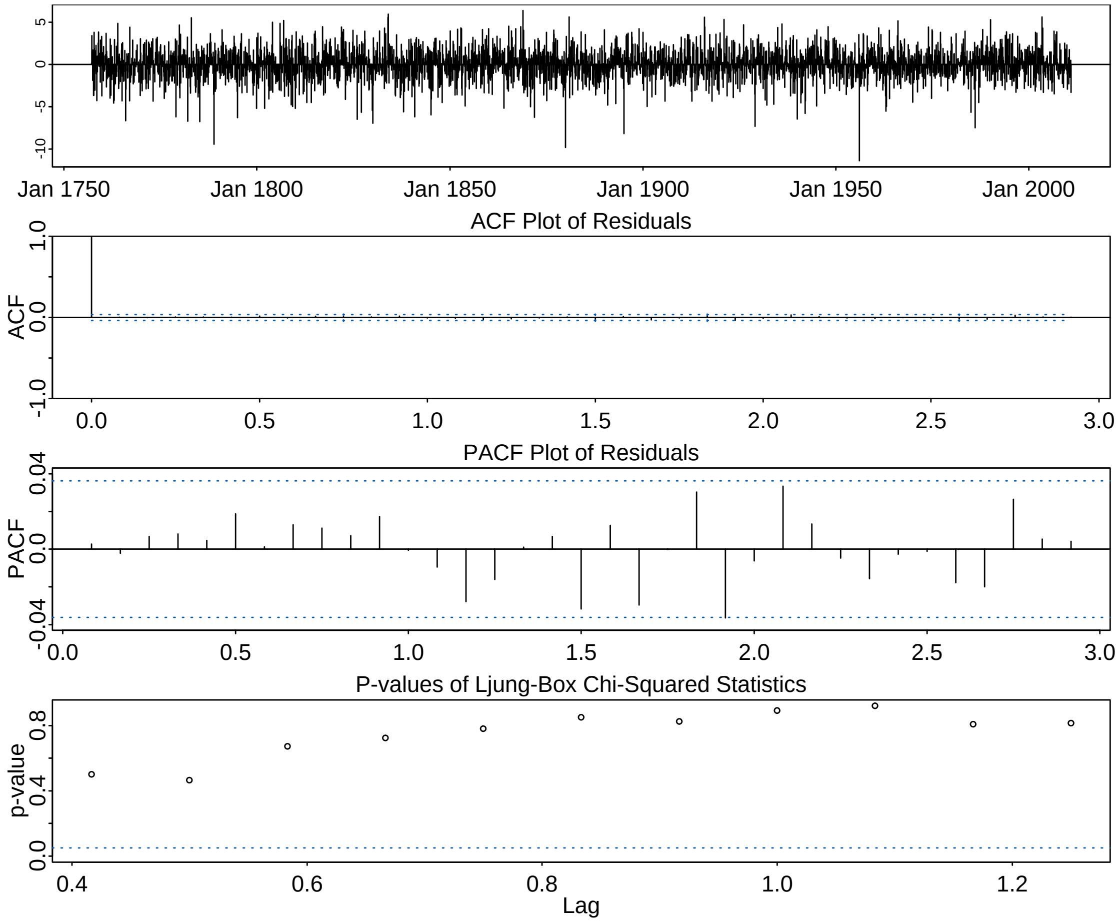


Figure 17.1: Residuals of the Basel monthly mean temperature time series from 1900 to 2010 after applying a seasonal ARIMA$(1, 1, 1) \times (1, 1, 1)_{12}$ process.

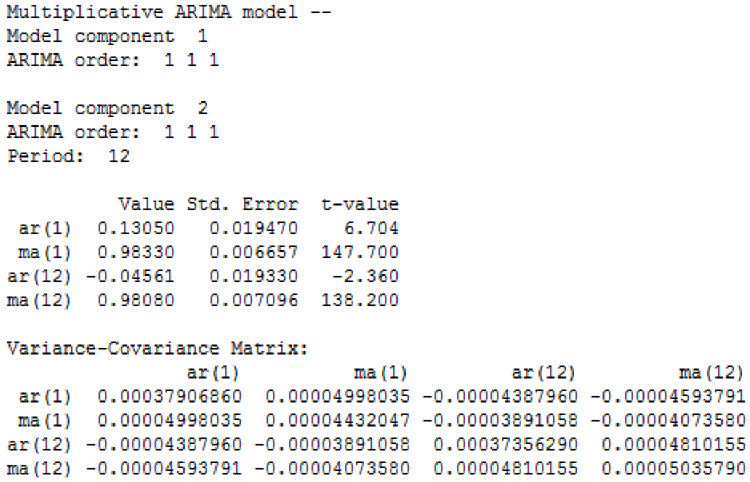
Remark. The equation (17.2) can be rewritten in the equivalent form

$$\phi^{\star}(B)Y_t = \theta^{\star}(B)Z_t,$$

where $\phi^{\star}(\cdot)$ and $\theta^{\star}(\cdot)$ are polynomials of degree $p+sP$ and $q+sQ$, respectively, whose coefficients can all be expressed in terms of $\phi_1, \ldots, \phi_p, \Phi_1, \ldots, \Phi_P, \theta_1, \ldots, \theta_q$, and $\Theta_1, \ldots, \Theta_Q$.

Remark. In Section 12.5 we discussed the classical decomposition model incorporating trend, seasonality, and random noise. In modeling real data it might not be reasonable to assume, as in the classical decomposition model, that the seasonal component s_t repeats itself precisely in the same way cycle after cycle. SARIMA models allow for randomness in the seasonal pattern from one cycle to the next.

Table 17.1: Parameter estimation of the seasonal ARIMA$(1,1,1) \times (1,1,1)_{12}$ process.



```
Multiplicative ARIMA model --
Model component  1
ARIMA order:  1 1 1

Model component  2
ARIMA order:  1 1 1
Period:  12

          Value Std. Error  t-value
 ar(1)  0.13050   0.019470    6.704
 ma(1)  0.98330   0.006657  147.700
ar(12) -0.04561   0.019330   -2.360
ma(12)  0.98080   0.007096  138.200

Variance-Covariance Matrix:
                ar(1)          ma(1)         ar(12)         ma(12)
 ar(1)  0.00037906860  0.00004998035 -0.00004387960 -0.00004593791
 ma(1)  0.00004998035  0.00004432047 -0.00003891058 -0.00004073580
ar(12) -0.00004387960 -0.00003891058  0.00037356290  0.00004810155
ma(12) -0.00004593791 -0.00004073580  0.00004810155  0.00005035790
```