5 Modeling and Forecasting 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 estimation of the mean, the coefficients $\{\phi_i, i = 1, \ldots, p\}$, $\{\theta_i, i = 1, \ldots, q\}$, and the white noise variance σ^2. Final selection of the model depends on a variety of goodness of fit tests, although it can be systematized to a large degree by use of criteria such as minimization of the AICC statistic as discussed in Section 5.5. (A useful option in the program ITSM is `Model>Estimation>Autofit`, which automatically minimizes the AICC statistic over all ARMA(p, q) processes with p and q in a specified range.)

This chapter is primarily devoted to the problem of estimating the parameters $\boldsymbol{\phi} = (\phi_i, \ldots, \phi_p)$, $\boldsymbol{\theta} = (\theta_i, \ldots, \theta_q)$, and σ^2 when p and q are assumed to be known, but the crucial issue of order selection is also considered. It will be assumed throughout (unless the mean is believed a priori to be zero) that the data have been "mean-corrected" by subtraction of the sample mean, so that it is appropriate to fit a zero-mean ARMA model to the adjusted data $x_1, \ldots, x_n$. If the model fitted to the mean-corrected data is

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

then the corresponding model for the original stationary series $\{Y_t\}$ is found on replacing X_t for each t by $Y_t - \bar{y}$, where $\bar{y} = n^{-1}\sum_{j=1}^{n} y_j$ is the sample mean of the original data, treated as a fixed constant.

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 $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$

P.J. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-29854-2_5

and σ^2. The estimators obtained by this procedure are known as maximum likelihood (or maximum Gaussian likelihood) estimators. Maximum likelihood estimation is discussed in Section 5.2 and can be carried out in practice using the ITSM option `Model>Estimation>Max likelihood`, after first specifying a preliminary model to initialize the maximization algorithm. Maximization of the likelihood and selection of the minimum AICC model over a specified range of p and q values can also be carried out using the option `Model>Estimation>Autofit`.

The maximization is nonlinear in the sense that the function to be maximized is not a quadratic function of the unknown parameters, so the estimators cannot be found by solving a system of linear equations. They are found instead by searching numerically for the maximum of the likelihood surface. The algorithm used in ITSM 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 in the option `Model>Estimation>Preliminary` of ITSM. They are described in Section 5.1. For pure autoregressive models the choice is between Yule-Walker and Burg estimation, while for models with $q > 0$ it is between the innovations and Hannan–Rissanen algorithms. It is also possible to begin the search with an arbitrary causal ARMA model by using the option `Model>Specify` and entering the desired parameter values. The initial values are chosen automatically in the option `Model>Estimation>Autofit`.

Calculation of the exact Gaussian likelihood for an ARMA model (and in fact for *any* second-order model) is greatly simplified by use of the innovations algorithm. In Section 5.2 we take advantage of this simplification in discussing maximum likelihood estimation and consider also the construction of confidence intervals for the estimated coefficients.

Section 5.3 deals with goodness of fit tests for the chosen model and Section 5.4 with the use of the fitted model for forecasting. In Section 5.5 we discuss the theoretical basis for some of the criteria used for order selection.

For an overview of the general strategy for model-fitting see Section 6.2.

5.1 Preliminary Estimation

In this section we shall consider four techniques for preliminary estimation of the parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$, $\boldsymbol{\theta} = (\theta_1, \ldots, \phi_p)'$, and σ^2 from observations $x_1, \ldots, x_n$ of the causal ARMA(p, q) process defined by

$$\phi(B)X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}\big(0, \sigma^2\big). \tag{5.1.1}$$

The Yule–Walker and Burg procedures apply to the fitting of pure autoregressive models. (Although the former can be adapted to models with $q > 0$, its performance is less efficient than when $q = 0$.) The innovation and Hannan–Rissanen algorithms are used in ITSM to provide preliminary estimates of the ARMA parameters when $q > 0$.

For pure autoregressive models Burg's algorithm usually gives higher likelihoods than the Yule–Walker equations. For pure moving-average models the innovations algorithm frequently gives slightly higher likelihoods than the Hannan–Rissanen algorithm (we use only the first two steps of the latter for preliminary estimation). For mixed models (i.e., those with $p > 0$ and $q > 0$) the Hannan–Rissanen algorithm is usually more successful in finding causal models (which are required for initialization of the likelihood maximization).

5.1.1 Yule–Walker Estimation

For a pure autoregressive model the moving-average polynomial $\theta(z)$ is identically 1, and the causality assumption in (5.1.1) allows us to write X_t in the form

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \tag{5.1.2}$$

where, from Section 3.1, $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = 1/\phi(z)$. Multiplying each side of (5.1.1) by X_{t-j}, $j = 0, 1, 2, \ldots, p$, taking expectations, and using (5.1.2) to evaluate the right-hand side of the first equation, we obtain the Yule–Walker equations

$$\Gamma_p \boldsymbol{\phi} = \boldsymbol{\gamma}_p \tag{5.1.3}$$

and

$$\sigma^2 = \gamma(0) - \boldsymbol{\phi}' \boldsymbol{\gamma}_p, \tag{5.1.4}$$

where Γ_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 $\boldsymbol{\phi}$.

On the other hand, if we replace the covariances $\gamma(j)$, $j = 0, \ldots, p$, appearing in (5.1.3) and (5.1.4) by the corresponding sample covariances $\hat{\gamma}(j)$, we obtain a set of equations for the so-called Yule–Walker estimators $\hat{\boldsymbol{\phi}}$ and $\hat{\sigma}^2$ of $\boldsymbol{\phi}$ and σ^2, namely,

$$\hat{\Gamma}_p \hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\gamma}}_p \tag{5.1.5}$$

and

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}' \hat{\boldsymbol{\gamma}}_p, \tag{5.1.6}$$

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

If $\hat{\gamma}(0) > 0$, then $\hat{\Gamma}_m$ is nonsingular for every $m = 1, 2, \ldots$ (see Brockwell and Davis (1991), Problem 7.11), so we can rewrite equations (5.1.5) and (5.1.6) in the following form:

Sample Yule–Walker Equations:

$$\hat{\boldsymbol{\phi}} = \left(\hat{\phi}_1, \ldots, \hat{\phi}_p\right)' = \hat{R}_p^{-1} \hat{\boldsymbol{\rho}}_p \tag{5.1.7}$$

and

$$\hat{\sigma}^2 = \hat{\gamma}(0)\left[1 - \hat{\boldsymbol{\rho}}_p' \hat{R}_p^{-1} \hat{\boldsymbol{\rho}}_p\right], \tag{5.1.8}$$

where $\hat{\boldsymbol{\rho}}_p = \left(\hat{\rho}(1), \ldots, \hat{\rho}(p)\right)' = \hat{\boldsymbol{\gamma}}_p / \hat{\gamma}(0)$.

With $\hat{\boldsymbol{\phi}}$ as defined by (5.1.7), it can be shown that $1 - \hat{\phi}_1 z - \cdots - \hat{\phi}_p z^p \neq 0$ for $|z| \leq 1$ (see Brockwell and Davis (1991), Problem 8.3). Hence the fitted model

$$X_t - \hat{\phi}_1 X_{t-1} - \cdots - \hat{\phi}_p X_{t-p} = Z_t, \quad \{Z_t\} \sim \mathrm{WN}\left(0, \hat{\sigma}^2\right)$$

is causal. The autocovariances $\gamma_F(h)$, $h = 0, \ldots, p$, of the fitted model therefore satisfy the $p + 1$ linear equations

$$\gamma_F(h) - \hat{\phi}_1 \gamma_F(h-1) - \cdots - \hat{\phi}_p \gamma_F(h-p) = \begin{cases} 0, & h = 1, \ldots, p, \\ \hat{\sigma}^2, & h = 0. \end{cases}$$

However, from (5.1.5) and (5.1.6) we see that the solution of these equations is $\gamma_F(h) = \hat{\gamma}(h)$, $h = 0, \ldots, p$, so that the autocovariances of the fitted model at lags $0, 1, \ldots, p$ coincide with the corresponding sample autocovariances.

The argument of the preceding paragraph shows that for *every* nonsingular covariance matrix of the form $\Gamma_{p+1} = [\gamma(i-j)]_{i,j=1}^{p+1}$ there is an AR(p) process whose autocovariances at lags $0, \ldots, p$ are $\gamma(0), \ldots, \gamma(p)$. (The required coefficients and white noise variance are found from (5.1.7) and (5.1.8) on replacing $\hat{\rho}(j)$ by $\gamma(j)/\gamma(0)$, $j = 0, \ldots, p$, and $\hat{\gamma}(0)$ by $\gamma(0)$.) There may not, however, be an MA(p) process with this property. For example, if $\gamma(0) = 1$ and $\gamma(1) = \gamma(-1) = \beta$, the matrix Γ_2 is a nonsingular covariance matrix for all $\beta \in (-1, 1)$. Consequently, there is an AR(1) process with autocovariances 1 and β at lags 0 and 1 for all $\beta \in (-1, 1)$. However, there is an MA(1) process with autocovariances 1 and β at lags 0 and 1 if and only if $|\beta| \leq \frac{1}{2}$. (See Example 2.1.1).

It is often the case that moment estimators, i.e., estimators that (like $\hat{\boldsymbol{\phi}}$) are obtained by equating theoretical and sample moments, have much higher variances than estimators obtained by alternative methods such as maximum likelihood. However, the Yule–Walker estimators of the coefficients $\phi_1, \ldots, \phi_p$ of an AR(p) process have approximately the same distribution for large samples as the corresponding maximum likelihood estimators. For a precise statement of this result see Brockwell and Davis (1991), Section 8.10. For our purposes it suffices to note the following:

Large-Sample Distribution of Yule–Walker Estimators:
For a large sample from an AR(p) process,

$$\hat{\boldsymbol{\phi}} \approx N\left(\boldsymbol{\phi}, n^{-1}\sigma^2\Gamma_p^{-1}\right).$$

If we replace σ^2 and Γ_p by their estimates $\hat{\sigma}^2$ and $\hat{\Gamma}_p$, we can use this result to find large-sample confidence regions for $\boldsymbol{\phi}$ and each of its components as in (5.1.12) and (5.1.13) below.

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. The second is more systematic and extends beyond the narrow class of pure autoregressive models.

- Some guidance in the choice of order is provided by a large-sample result (see Brockwell and Davis (1991), Section 8.10), which states that if $\{X_t\}$ is the causal AR(p) process defined by (5.1.1) with $\{Z_t\} \sim \text{iid}\left(0, \sigma^2\right)$ and if we fit a model with order $m > p$ using the Yule–Walker equations, i.e., if we fit a model with coefficient vector

  $$\hat{\boldsymbol{\phi}}_m = \hat{R}_m^{-1}\hat{\boldsymbol{\rho}}_m, \quad m > p,$$

  then the *last component*, $\hat{\phi}_{mm}$, of the vector $\hat{\boldsymbol{\phi}}_m$ 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 as defined in Section 3.2.3.

  Now, we already know from Example 3.2.6 that for an AR(p), process the partial autocorrelations ϕ_{mm}, $m > p$, are zero. By the result of the previous paragraph,

if an AR(p) model is appropriate for the data, then the values $\hat{\phi}_{kk}$, $k > p$, should be compatible with observations from the distribution 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 $k > m$.

The program ITSM plots the sample PACF $\{\hat{\phi}_{mm}, m = 1, 2, \ldots\}$ together with the bounds $\pm 1.96/\sqrt{n}$. From this graph it is easy to read off the preliminary estimator of p defined above.

- A more systematic approach to order selection is to find the values of p and $\boldsymbol{\phi}_p$ that minimize the AICC statistic (see Section 5.5.2 below)

$$\text{AICC} = -2 \ln L(\boldsymbol{\phi}_p, S(\boldsymbol{\phi}_p)/n) + 2(p+1)n/(n-p-2),$$

where L is the Gaussian likelihood defined in (5.2.9) and S is defined in (5.2.11). The `Preliminary Estimation` dialog box of ITSM (opened by pressing the blue PRE button) allows you to search for the minimum AICC Yule–Walker (or Burg) models by checking `Find AR model with min AICC`. This causes the program to fit autoregressions of orders $0, 1, \ldots, 27$ and to return the model with smallest AICC value.

Definition 5.1.1

The **fitted Yule–Walker AR(m) model** is

$$X_t - \hat{\phi}_{m1} X_{t-1} - \cdots - \hat{\phi}_{mm} X_{t-m} = Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \hat{v}_m\right), \tag{5.1.9}$$

where

$$\hat{\boldsymbol{\phi}}_m = \left(\hat{\phi}_{m1}, \ldots, \hat{\phi}_{mm}\right)' = \hat{R}_m^{-1} \hat{\boldsymbol{\rho}}_m \tag{5.1.10}$$

and

$$\hat{v}_m = \hat{\gamma}(0)\left[1 - \hat{\boldsymbol{\rho}}_m' \hat{R}_m^{-1} \hat{\boldsymbol{\rho}}_m\right]. \tag{5.1.11}$$

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 in Section 2.5.3, where we derived a recursive scheme for solving the equations (5.1.3) and (5.1.4) with p successively taking the values $1, 2, \ldots$. Here we can use exactly the same scheme (the Durbin–Levinson algorithm) to solve the Yule–Walker equations (5.1.5) and (5.1.6), the only difference being that the covariances in (5.1.3) and (5.1.4) are replaced by their sample counterparts. This is the algorithm used by ITSM to perform the necessary calculations.

Confidence Regions for the Coefficients

Under 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, if $\chi^2_{1-\alpha}(p)$ denotes the $(1-\alpha)$ quantile of the chi-squared distribution with p degrees of freedom, then for large sample-size n the region

$$\left\{\boldsymbol{\phi} \in \mathbf{R}^p : \left(\hat{\boldsymbol{\phi}}_p - \boldsymbol{\phi}\right)' \hat{\Gamma}_p \left(\hat{\boldsymbol{\phi}}_p - \boldsymbol{\phi}\right) \le n^{-1} \hat{v}_p \chi^2_{1-\alpha}(p)\right\} \tag{5.1.12}$$

contains ϕ_p with probability close to $(1-\alpha)$. (This follows from Problem A.7 and the fact that $\sqrt{n}(\hat{\phi}_p - \phi_p)$ is approximately normally distributed with mean $\mathbf{0}$ and covariance matrix $\hat{v}_p\hat{\Gamma}_p^{-1}$.) Similarly, if $\Phi_{1-\alpha}$ denotes the $(1-\alpha)$ quantile of the standard normal distribution and $\hat{v}_{jj}$ is the jth diagonal element of $\hat{v}_p\hat{\Gamma}_p^{-1}$, then for large n the interval bounded by

$$\hat{\phi}_{pj} \pm \Phi_{1-\alpha/2} n^{-1/2} \hat{v}_{jj}^{1/2} \tag{5.1.13}$$

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

Example 5.1.1 The Dow Jones Utilities Index, Aug. 28–Dec. 18, 1972; DOWJ.TSM

The very slowly decaying positive sample ACF of the time series contained in the file DOWJ.TSM this time series suggests differencing at lag 1 before attempting to fit a stationary model. One application of the operator $(1-B)$ produces a new series $\{Y_t\}$ with no obvious deviations from stationarity. We shall therefore try fitting an AR process to this new series

$$Y_t = D_t - D_{t-1}$$

using the Yule–Walker equations. There are 77 values of Y_t, which we shall denote by $Y_1, \ldots, Y_{77}$. (We ignore the unequal spacing of the original data resulting from the five-day working week.) The sample autocovariances of the series $y_1, \ldots, y_{77}$ are $\hat{\gamma}(0) = 0.17992$, $\hat{\gamma}(1) = 0.07590$, $\hat{\gamma}(2) = 0.04885$, etc.

Applying the Durbin–Levinson algorithm to fit successively higher-order autoregressive processes to the data, we obtain

$$\hat{\phi}_{11} = \hat{\rho}(1) = 0.4219,$$

$$\hat{v}_1 = \hat{\gamma}(0)\left[1 - \hat{\rho}^2(1)\right] = 0.1479,$$

$$\hat{\phi}_{22} = \left[\hat{\gamma}(2) - \hat{\phi}_{11}\hat{\gamma}(1)\right]/\hat{v}_1 = 0.1138,$$

$$\hat{\phi}_{21} = \hat{\phi}_{11} - \hat{\phi}_{11}\hat{\phi}_{22} = 0.3739,$$

$$\hat{v}_2 = \hat{v}_1\left[1 - \hat{\phi}_{22}^2\right] = 0.1460.$$

The sample ACF and PACF of the data can be displayed by pressing the second yellow button at the top of the ITSM window. They are shown in Figures 5-1 and 5-2, respectively. Also plotted are the bounds $\pm 1.96/\sqrt{77}$. Since the PACF values at lags greater than 1 all lie between the bounds, the first order-selection criterion described above indicates that we should fit an AR(1) model to the data set $\{Y_t\}$. Unless we wish to assume that $\{Y_t\}$ is a zero-mean process, we should subtract the sample mean from the data before attempting to fit a (zero-mean) AR(1) model. When the blue PRE (preliminary estimation) button at the top of the ITSM window is pressed, you will be given the option of subtracting the mean from the data. In this case (as in most) click `Yes` to obtain the new series

$$X_t = Y_t - 0.1336.$$

You will then see the `Preliminary Estimation` dialog box. Enter 1 for the AR order, zero for the MA order, select `Yule-Walker`, and click `OK`. We have already computed $\hat{\phi}_{11}$ and $\hat{v}_1$ above using the Durbin–Levinson algorithm. The Yule–Walker AR(1) model obtained by ITSM for $\{X_t\}$ is therefore (not surprisingly)

$$X_t - 0.4219X_{t-1} = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 0.1479), \tag{5.1.14}$$

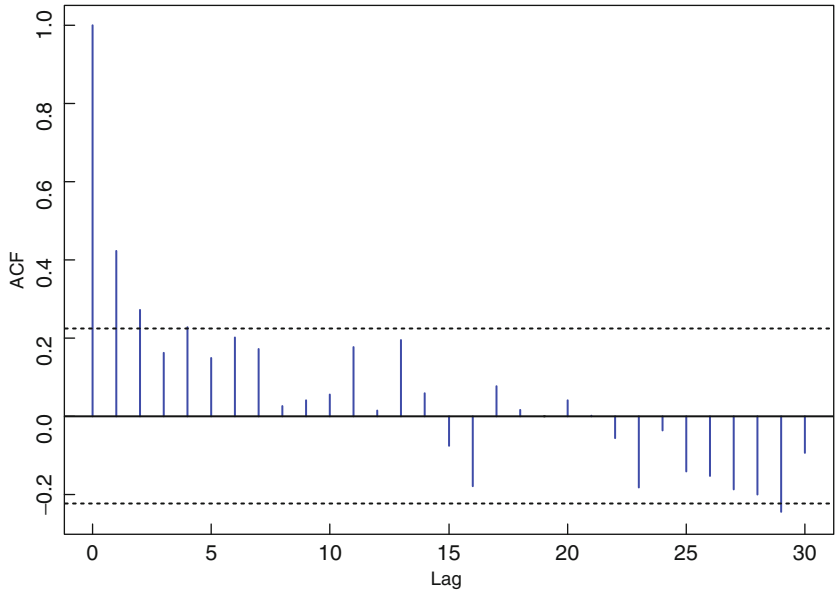

Figure 5-1
The sample ACF of the differenced series $\{Y_t\}$ in Example 5.1.1

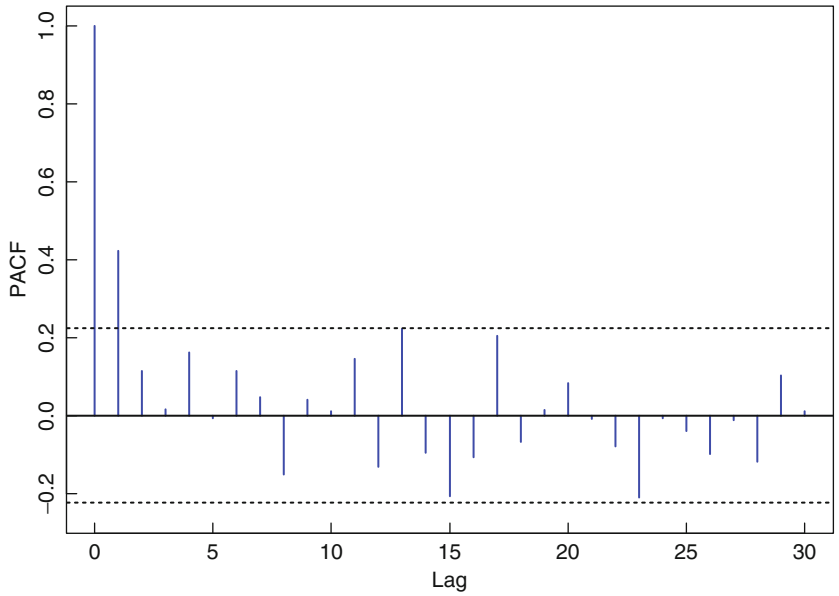

Figure 5-2
The sample PACF of the differenced series $\{Y_t\}$ in Example 5.1.1

and the corresponding model for $\{Y_t\}$ is

$$Y_t - 0.1336 - 0.4219(Y_{t-1} - 0.1336) = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 0.1479). \tag{5.1.15}$$

Assuming that our observed data really are generated by an AR process with $p = 1$, (5.1.13) gives us approximate 95 % confidence bounds for the autoregressive coefficient ϕ,

$$0.4219 \pm \frac{(1.96)(0.1479)^{1/2}}{(0.17992)^{1/2}\sqrt{77}} = (0.2194, 0.6244).$$

Besides estimating the autoregressive coefficients, ITSM computes and prints out the ratio of each coefficient to 1.96 times its estimated standard deviation. From these numbers large-sample 95 % confidence intervals for each of the coefficients are easily

obtained. In this particular example there is just one coefficient estimate, $\hat{\phi}_1 = 0.4219$, with ratio of coefficient to 1.96×standard error equal to 2.0832. Hence the required 95 % confidence bounds are $0.4219 \pm 0.4219/2.0832 = (0.2194, 0.6244)$, as found above.

A useful technique for preliminary autoregressive estimation that incorporates automatic model selection (i.e., choice of p) is to minimize the AICC [see equation (5.5.4)] over all fitted autoregressions of orders 0 through 27. This is achieved by selecting *both* `Yule-Walker` and `Find AR model with min AICC` in the `Preliminary Estimation` dialog box. (The MA order must be set to zero, but the AR order setting is immaterial.) Click `OK`, and the program will search through all the Yule–Walker AR(p) models, $p = 0, 1, \ldots, 27$, selecting the one with smallest AICC value. The minimum-AICC Yule–Walker AR model turns out to be the one defined by (5.1.14) with $p = 1$ and AICC value 74.541.

□

Yule–Walker Estimation with $q > 0$; Moment Estimators

The Yule–Walker estimates for the parameters in an AR(p) model are examples of moment estimators: The autocovariances at lags $0, 1, \ldots, p$ are replaced by the corresponding sample estimates in the Yule–Walker equations (5.1.3), which are then solved for the parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ and σ^2. The analogous procedure for ARMA(p, q) models with $q > 0$ is easily formulated, but the corresponding equations are nonlinear in the unknown coefficients, leading to possible nonexistence and nonuniqueness of solutions for the required estimators.

From (3.2.5), the equations to be solved for $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ and σ^2 are

$$\hat{\gamma}(k) - \phi_1\hat{\gamma}(k-1) - \cdots - \phi_p\hat{\gamma}(k-p) = \sigma^2 \sum_{j=k}^{q} \theta_j\psi_{j-k}, \quad 0 \le k \le p+q, \tag{5.1.16}$$

where ψ_j must first be expressed in terms of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ using the identity $\psi(z) = \theta(z)/\phi(z)$ ($\theta_0 := 1$ and $\theta_j = \psi_j = 0$ for $j < 0$).

Example 5.1.2 For the MA(1) model the equation (5.1.16) are equivalent to

$$\hat{\gamma}(0) = \hat{\sigma}^2\left(1 + \hat{\theta}_1^2\right), \tag{5.1.17}$$

$$\hat{\rho}(1) = \frac{\hat{\theta}_1}{1 + \hat{\theta}_1^2}. \tag{5.1.18}$$

If $\left|\hat{\rho}(1)\right| > 0.5$, there is no real solution, so we define $\hat{\theta}_1 = \hat{\rho}(1)/\left|\hat{\rho}(1)\right|$. If $\left|\hat{\rho}(1)\right| \le 0.5$, then the solution of (5.1.17)–(5.1.18) (with $|\hat{\theta}| \le 1$) is

$$\hat{\theta}_1 = \left(1 - \left(1 - 4\hat{\rho}^2(1)\right)^{1/2}\right) / \left(2\hat{\rho}(1)\right),$$

$$\hat{\sigma}^2 = \hat{\gamma}(0) / \left(1 + \hat{\theta}_1^2\right).$$

For the overshort data of Example 3.2.8, $\hat{\rho}(1) = -0.5035$ and $\hat{\gamma}(0) = 3416$, so the fitted MA(1) model has parameters $\hat{\theta}_1 = -1.0$ and $\hat{\sigma}^2 = 1708$.

□

Relative Efficiency of Estimators

The performance of two competing estimators is often measured by computing their asymptotic relative efficiency. In a general statistics estimation problem, suppose $\hat{\theta}_n^{(1)}$ and $\hat{\theta}_n^{(2)}$ are two estimates of the parameter θ in the parameter space Θ based on the observations $X_1, \ldots, X_n$. If $\hat{\theta}_n^{(i)}$ is approximately $\mathrm{N}\big(\theta, \sigma_i^2(\theta)\big)$ for large n, $i = 1, 2$, then the **asymptotic efficiency** of $\hat{\theta}_n^{(1)}$ relative to $\hat{\theta}_n^{(2)}$ is defined to be

$$e\left(\theta, \hat{\theta}^{(1)}, \hat{\theta}^{(2)}\right) = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}.$$

If $e\big(\theta, \hat{\theta}^{(1)}, \hat{\theta}^{(2)}\big) \le 1$ for all $\theta \in \Theta$, then we say that $\hat{\theta}_n^{(2)}$ is a more efficient estimator of θ than $\hat{\theta}_n^{(1)}$ (strictly more efficient if in addition, $e\big(\theta, \hat{\theta}^{(1)}, \hat{\theta}^{(2)}\big) < 1$ for some $\theta \in \Theta$). For the MA(1) process the moment estimator $\theta_n^{(1)}$ discussed in Example 5.1.2 is approximately $\mathrm{N}\big(\theta_1, \sigma_1^2(\theta_1)/n\big)$ with

$$\sigma_1^2\big(\theta_1\big) = \big(1 + \theta_1^2 + 4\theta_1^4 + \theta_1^6 + \theta_1^8\big)/\big(1 - \theta_1^2\big)^2$$

(see Brockwell and Davis (1991), p. 254). On the other hand, the innovations estimator $\hat{\theta}_n^{(2)}$ discussed in the next section is distributed approximately as $\mathrm{N}\big(\theta_1, n^{-1}\big)$. Thus, $e\big(\theta_1, \hat{\theta}^{(1)}, \hat{\theta}^{(2)}\big) = \sigma_1^{-2}(\theta_1) \le 1$ for all $|\theta_1| < 1$, with strict inequality when $\theta \neq 1$. In particular,

$$e\left(\theta_1, \hat{\theta}^{(1)}, \hat{\theta}^{(2)}\right) = \begin{cases} 0.82, & \theta_1 = 0.25, \\ 0.37, & \theta_1 = 0.50, \\ 0.06, & \theta_1 = 0.75, \end{cases}$$

demonstrating the superiority, at least in terms of asymptotic relative efficiency, of $\hat{\theta}_n^{(2)}$ over $\hat{\theta}_n^{(1)}$. On the other hand (Section 5.2), the maximum likelihood estimator $\hat{\theta}_n^{(3)}$ of θ_1 is approximately $\mathrm{N}(\theta_1, (1 - \theta_1^2)/n)$. Hence,

$$e\left(\theta_1, \hat{\theta}^{(2)}, \hat{\theta}^{(3)}\right) = \begin{cases} 0.94, & \theta_1 = 0.25, \\ 0.75, & \theta_1 = 0.50, \\ 0.44, & \theta_1 = 0.75. \end{cases}$$

While $\hat{\theta}_n^{(3)}$ is more efficient, $\hat{\theta}_n^{(2)}$ has reasonably good efficiency, except when $|\theta_1|$ is close to 1, and can serve as initial value for the nonlinear optimization procedure in computing the maximum likelihood estimator.

While the method of moments is an effective procedure for fitting autoregressive models, it does not perform as well for ARMA models with $q > 0$. From a computational point of view, it requires as much computing time as the more efficient estimators based on either the innovations algorithm or the Hannan–Rissanen procedure and is therefore rarely used except when $q = 0$.

5.1.2 Burg's Algorithm

The Yule–Walker coefficients $\hat{\phi}_{p1}, \ldots, \hat{\phi}_{pp}$ 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 ACF of $\{X_t\}$ coincides with the sample ACF at lags $1, \ldots, p$.

Burg's algorithm estimates the PACF $\{\phi_{11}, \phi_{22}, \ldots\}$ 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 $u_i(t)$, $t = i + 1, \ldots, n$, $0 \le i < n$, 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 $v_i(t)$, $t = i+1, \ldots, n$, $0 \le i < n$, 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 (see Problem 5.6) that the **forward and backward prediction errors** $\{u_i(t)\}$ and $\{v_i(t)\}$ satisfy the recursions

$$u_0(t) = v_0(t) = x_{n+1-t},$$

$$u_i(t) = u_{i-1}(t-1) - \phi_{ii} v_{i-1}(t), \tag{5.1.19}$$

and

$$v_i(t) = v_{i-1}(t) - \phi_{ii} u_{i-1}(t-1). \tag{5.1.20}$$

Burg's estimate $\phi_{11}^{(B)}$ of ϕ_{11} is found by minimizing

$$\sigma_1^2 := \frac{1}{2(n-1)} \sum_{t=2}^{n} \left[u_1^2(t) + v_1^2(t) \right]$$

with respect to ϕ_{11}. This gives corresponding numerical values for $u_1(t)$ and $v_1(t)$ and σ_1^2 that can then be substituted into (5.1.19) and (5.1.20) with $i = 2$. Then we minimize

$$\sigma_2^2 := \frac{1}{2(n-2)} \sum_{t=3}^{n} \left[u_2^2(t) + v_2^2(t) \right]$$

with respect to ϕ_{22} to obtain the Burg estimate $\phi_{22}^{(B)}$ of ϕ_{22} and corresponding values of $u_2(t)$, $v_2(t)$, and σ_2^2. This process can clearly be continued to obtain estimates $\phi_{pp}^{(B)}$ and corresponding minimum values, $\sigma_p^{(B)2}$, $p \le n-1$. Estimates of the coefficients ϕ_{pj}, $1 \le j \le p-1$, in the best linear predictor

$$P_p X_{p+1} = \phi_{p1} X_p + \cdots + \phi_{pp} X_1$$

are then found by substituting the estimates $\phi_{ii}^{(B)}$, $i = 1, \ldots, p$, for ϕ_{ii} in the recursions (2.5.20)–(2.5.22). The resulting estimates of ϕ_{pj}, $j = 1, \ldots, p$, are the coefficient estimates of the Burg AR(p) model for the data $\{x_1, \ldots, x_n\}$. The Burg estimate of the white noise variance is the minimum value $\sigma_p^{(B)2}$ found in the determination of $\phi_{pp}^{(B)}$. The calculation of the estimates of ϕ_{pp} and σ_p^2 described above is equivalent (Problem 5.7) to solving the following recursions:

Burg's Algorithm:

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

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

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

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

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, $\mathrm{N}\big(\boldsymbol{\phi}, n^{-1}\sigma^2\Gamma_p^{-1}\big)$. Approximate large-sample confidence intervals for the coefficients can be found as in Section 5.1.1 by substituting estimated values for σ^2 and Γ_p.

Example 5.1.3 The Dow Jones Utilities Index

The fitting of AR models using Burg's algorithm in the program ITSM is completely analogous to the use of the Yule–Walker equations. Applying the same transformations as in Example 5.1.1 to the Dow Jones Utilities Index and selecting `Burg` instead of `Yule-Walker` in the `Preliminary Estimation` dialog box, we obtain the minimum AICC Burg model

$$X_t - 0.4371X_{t-1} = Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, 0.1423), \tag{5.1.21}$$

with AICC = 74.492. This is slightly different from the Yule–Walker AR(1) model fitted in Example 5.1.1, and it has a larger likelihood L, i.e., a smaller value of $-2\ln L$ (see Section 5.2). 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. From the ratio of the estimated coefficient to (1.96× standard error) displayed by ITSM, we obtain the 95 % confidence bounds for ϕ: $0.4371 \pm 0.4371/2.1668 = (0.2354, 0.6388)$.

□

Example 5.1.4 The Lake Data

This series $\{Y_t, t = 1, \ldots, 98\}$ has already been studied in Example 1.3.5. In this example we shall consider the problem of fitting an AR process directly to the data without first removing any trend component. A graph of the data was displayed in Figure 1-9. The sample ACF and PACF are shown in Figures 5-3 and 5-4, respectively.

The sample PACF shown in Figure 5-4 strongly suggests fitting an AR(2) model to the mean-corrected data $X_t = Y_t - 9.0041$. After clicking on the blue preliminary estimation button of ITSM select `Yes` to subtract the sample mean from $\{Y_t\}$. Then

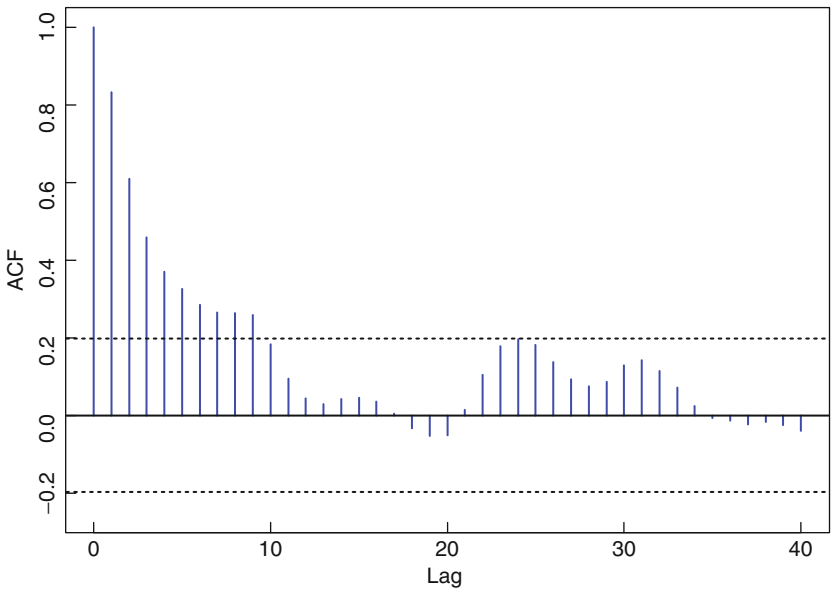

Figure 5-3
The sample ACF of the lake data in Example 5.1.4

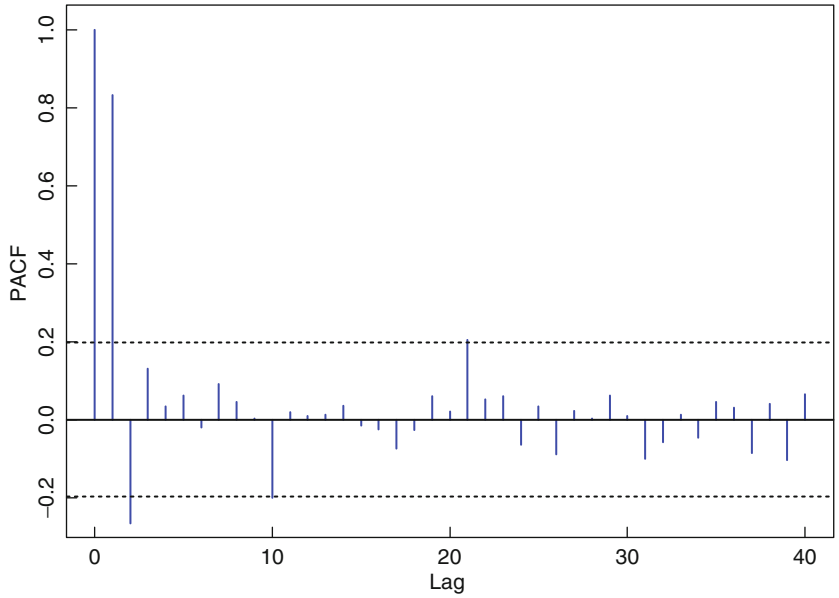

Figure 5-4
The sample PACF of the lake data in Example 5.1.4

specify 2 for the AR order, 0 for the MA order, and `Burg` for estimation. Click OK to obtain the model

$$X_t - 1.0449X_{t-1} + 0.2456X_{t-2} = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 0.4706),$$

with AICC value 213.55 and 95 % confidence bounds

$$\phi_1 : 1.0449 \pm 1.0449/5.5295 = (0.8559, 1.2339),$$

$$\phi_2 : -0.2456 \pm 0.2456/1.2997 = (-0.4346, -0.0566).$$

Selecting the Yule–Walker method for estimation, we obtain the model

$$X_t - 1.0538X_{t-1} + 0.2668X_{t-2} = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 0.4920),$$

with AICC value 213.57 and 95 % confidence bounds

$$\phi_1 : 1.0538 \pm 1.0538/5.5227 = (0.8630, 1.2446),$$

$$\phi_2 : -0.2668 \pm 0.2668/1.3980 = (-0.4576, -0.0760).$$

We notice, as in Example 5.1.3, that the Burg model again has smaller white noise variance and larger Gaussian likelihood than the Yule–Walker model.

If we determine the minimum AICC Yule–Walker and Burg models, we find that they are both of order 2. Thus the order suggested by the sample PACF coincides again with the order obtained by AICC minimization.

□

5.1.3 The Innovations Algorithm

Just as we can fit autoregressive models of orders $1, 2, \ldots$ 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

$$X_t = Z_t + \hat{\theta}_{m1}Z_{t-1} + \cdots + \hat{\theta}_{mm}Z_{t-m}, \quad \{Z_t\} \sim \text{WN}\left(0, \hat{v}_m\right) \tag{5.1.22}$$

of orders $m = 1, 2, \ldots$ by means of the innovations algorithm (Section 2.5.4). The estimated coefficient vectors $\hat{\boldsymbol{\theta}}_m := \left(\hat{\theta}_{m1}, \ldots, \hat{\theta}_{mm}\right)'$ and white noise variances $\hat{v}_m$,

$m = 1, 2, \ldots$, are specified in the following definition. (The justification for using estimators defined in this way is contained in Remark 1 following the definition.)

Definition 5.1.2

The **fitted innovations MA(m) model** is

$$X_t = Z_t + \hat{\theta}_{m1} Z_{t-1} + \cdots + \hat{\theta}_{mm} Z_{t-m}, \quad \{Z_t\} \sim \text{WN}(0, \hat{v}_m),$$

where $\hat{\boldsymbol{\theta}}_m$ and $\hat{v}_m$ are obtained from the innovations algorithm with the ACVF replaced by the sample ACVF.

Remark 1. It can be shown (see Brockwell and Davis 1988) that if $\{X_t\}$ is an invertible MA(q) process

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \quad \{Z_t\} \sim \text{IID}\left(0, \sigma^2\right),$$

with $EZ_t^4 < \infty$, and if we define $\theta_0 = 1$ and $\theta_j = 0$ for $j > q$, then the innovation estimates have the following large-sample properties. If $n \to \infty$ and $m(n)$ is any sequence of positive integers such that $m(n) \to \infty$ but $n^{-1/3} m(n) \to 0$, then for each positive integer k the joint distribution function of

$$n^{1/2}\left(\hat{\theta}_{m1} - \theta_1, \hat{\theta}_{m2} - \theta_2, \ldots, \hat{\theta}_{mk} - \theta_k\right)'$$

converges to that of the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $A = [a_{ij}]_{i,j=1}^{k}$, where

$$a_{ij} = \sum_{r=1}^{\min(i,j)} \theta_{i-r}\theta_{j-r}. \tag{5.1.23}$$

This result enables us to find approximate large-sample confidence intervals for the moving-average coefficients from the innovation estimates as described in the examples below. Moreover, the estimator $\hat{v}_m$ is **consistent** for σ^2 in the sense that for every $\epsilon > 0$, $P\left(\left|\hat{v}_m - \sigma^2\right| > \epsilon\right) \to 0$ as $m \to \infty$. □

Remark 2. Although the recursive fitting of moving-average models using the innovations algorithm is closely analogous to the recursive fitting of autoregressive models using the Durbin–Levinson algorithm, there is one important distinction. For an AR(p) process the Yule–Walker and Burg estimators $\hat{\boldsymbol{\phi}}_p$ are consistent estimators of $(\phi_1, \ldots, \phi_p)'$ as the sample size $n \to \infty$. However, for an MA(q) process the estimator $\hat{\boldsymbol{\theta}}_q = (\theta_{q1}, \ldots, \theta_{qq})'$ is not consistent for $(\theta_1, \ldots, \theta_q)'$. For consistency it is necessary to use the estimators $(\theta_{m1}, \ldots, \theta_{mq})'$ with $m(n)$ satisfying the conditions of Remark 1. The choice of m for any fixed sample size can be made by increasing m until the vector $(\theta_{m1}, \ldots, \theta_{mq})'$ stabilizes. It is found in practice that there is a large range of values of m for which the fluctuations in θ_{mj} are small compared with the estimated asymptotic standard deviation $n^{-1/2}\left(\sum_{i=0}^{j-1} \hat{\theta}_{mi}^2\right)^{1/2}$ as found from (5.1.23) when the coefficients θ_j are replaced by their estimated values $\hat{\theta}_{mj}$. □

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 from Section 3.2.2 that for an MA(q) process the autocorrelations $\rho(m)$, $m > q$, are zero. Moreover, we know from Bartlett's formula (Section 2.4) that the sample autocorrelation $\hat{\rho}(m)$, $m > q$, is approximately normally distributed with mean $\rho(m) = 0$ and variance $n^{-1}\big[1 + 2\rho^2(1) + \cdots + 2\rho^2(q)\big]$. 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 "significantly different from zero" is often interpreted as "larger than $1.96/\sqrt{n}$ in absolute value" (cf. the corresponding approach to order selection for AR models based on the sample PACF and described in Section 5.1.1).
- If in addition to examining $\hat{\rho}(m)$, $m = 1, 2, \ldots$, we examine the coefficient vectors $\hat{\theta}_m$, $m = 1, 2, \ldots$, we are able not only to assess the appropriateness of a moving-average model and estimate its order q, but at the same time 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 times its approximate standard deviation $\sigma_j = n^{-1/2}\big[\sum_{i=0}^{j-1} \hat{\theta}_{mi}^2\big]^{1/2}$, we can see which of the coefficient estimates are most significantly different from zero, estimate the order of the model to be fitted as the largest lag j for which the ratio is larger than 1 in absolute value, and at the same time read off estimated values for each of the coefficients. A default value of m is set by the program, but it may be altered manually. As m is increased the values $\hat{\theta}_{m1}, \ldots, \hat{\theta}_{mm}$ stabilize in the sense that the fluctuations in each component are of order $n^{-1/2}$, the asymptotic standard deviation of θ_{m1}.
- As for autoregressive models, a more systematic approach to order selection for moving-average models is to find the values of q and $\hat{\boldsymbol{\theta}}_q = \big(\hat{\theta}_{m1}, \ldots, \hat{\theta}_{mq}\big)'$ that minimize the AICC statistic

$$\text{AICC} = -2 \ln L(\boldsymbol{\theta}_q, S(\boldsymbol{\theta}_q)/n) + 2(q+1)n/(n-q-2),$$

  where L is the Gaussian likelihood defined in (5.2.9) and S is defined in (5.2.11). (See Section 5.5 for further details.)

Confidence Regions for the Coefficients

Asymptotic confidence regions for the coefficient vector $\boldsymbol{\theta}_q$ and for its individual components can be found with the aid of the large-sample distribution specified in Remark 1. For example, approximate 95 % confidence bounds for θ_j are given by

$$\hat{\theta}_{mj} \pm 1.96 n^{-1/2} \left(\sum_{i=0}^{j-1} \hat{\theta}_{mi}^2 \right)^{1/2}. \tag{5.1.24}$$

Example 5.1.5 The Dow Jones Utilities Index

In Example 5.1.1 we fitted an AR(1) model to the differenced Dow Jones Utilities Index. The sample ACF of the differenced data shown in Figure 5-1 suggests that an MA(2) model might also provide a good fit to the data. To apply the innovation technique for preliminary estimation, we proceed as in Example 5.1.1 to difference the series DOWJ.TSM to obtain observations of the differenced series $\{Y_t\}$. We then select preliminary estimation by clicking on the blue PRE button and subtract the mean of the differences to obtain observations of the differenced and mean-corrected series $\{X_t\}$. In the `Preliminary Estimation` dialog box enter 0 for the AR order and

2 for the MA order, and select `Innovations` as the estimation method. We must then specify a value of m, which is set by default in this case to 17. If we accept the default value, the program will compute $\hat{\theta}_{17,1}, \ldots, \hat{\theta}_{17,17}$ and print out the first two values as the estimates of θ_1 and θ_2, together with the ratios of the estimated values to their estimated standard deviations. These are

MA COEFFICIENT

0.4269 0.2704

COEFFICIENT/(1.96*STANDARD ERROR)

1.9114 1.1133

The remaining parameter in the model is the white noise variance, for which two estimates are given:

WN VARIANCE ESTIMATE = (RESID SS)/N

0.1470

INNOVATION WN VARIANCE ESTIMATE

0.1122

The first of these is the average of the squares of the rescaled one-step prediction errors under the fitted MA(2) model, i.e., $\frac{1}{77}\sum_{j=1}^{77}\left(X_j - \hat{X}_j\right)^2/r_{j-1}$. The second value is the innovation estimate, $\hat{v}_{17}$. (By default ITSM retains the first value. If you wish instead to use the innovation estimate, you must change the white noise variance by selecting `Model>Specify` and setting the white noise value to the desired value.) The fitted model for $X_t (= Y_t - 0.1336)$ is thus

$$X_t = Z_t + 0.4269Z_{t-1} + 0.2704Z_{t-2}, \quad \{Z_t\} \sim \text{WN}(0, 0.1470),$$

with AICC = 77.467.

To see all 17 estimated coefficients $\hat{\theta}_{17,j}, j = 1, \ldots, 17$, we repeat the preliminary estimation, this time fitting an MA(17) model with $m = 17$. The coefficients and ratios for the resulting model are found to be as follows:

MA COEFFICIENT

0.4269	0.2704	0.1183	0.1589	0.1355	0.1568	0.1284	−0.0060
0.0148	−0.0017	0.1974	−0.0463	0.2023	0.1285	−0.0213	−0.2575
0.0760							

COEFFICIENT/(1.96*STANDARD ERROR)

1.9114	1.1133	0.4727	0.6314	0.5331	0.6127	0.4969	−0.0231
0.0568	−0.0064	0.7594	−0.1757	0.7667	0.4801	−0.0792	−0.9563
0.2760							

The ratios indicate that the estimated coefficients most significantly different from zero are the first and second, reinforcing our original intention of fitting an MA(2) model to the data. Estimated coefficients $\hat{\theta}_{mj}$ for other values of m can be examined in the same way, and it is found that the values obtained for $m > 17$ change only slightly from the values tabulated above.

By fitting MA(q) models of orders 0, 1, 2, ..., 26 using the innovations algorithm with the default settings for m, we find that the minimum AICC model is the one with $q = 2$ found above. Thus the model suggested by the sample ACF again coincides with the more systematically chosen minimum AICC model.

□

Innovations Algorithm Estimates when $p > 0$ and $q > 0$
The causality assumption (Section 3.1) ensures that

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

where the coefficients ψ_j satisfy

$$\psi_j = \theta_j + \sum_{i=1}^{\min(j,\, p)} \phi_i \psi_{j-i}, \quad j = 0, 1, \ldots, \tag{5.1.25}$$

and we define $\theta_0 := 1$ and $\theta_j := 0$ for $j > q$. To estimate $\psi_1, \ldots, \psi_{p+q}$ we can use the innovation estimates $\hat{\theta}_{m1}, \ldots, \hat{\theta}_{m,p+q}$, whose large-sample behavior is specified in Remark 1. Replacing ψ_j by $\hat{\theta}_{mj}$ in (5.1.25) and solving the resulting equations

$$\hat{\theta}_{mj} = \theta_j + \sum_{i=1}^{\min(j,p)} \phi_i \hat{\theta}_{m,j-i}, \quad j = 1, \ldots, p+q, \tag{5.1.26}$$

for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, we obtain initial parameter estimates $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$. To solve (5.1.26) we first find $\boldsymbol{\phi}$ from the last q equations:

$$\begin{bmatrix} \hat{\theta}_{m,q+1} \\ \hat{\theta}_{m,q+2} \\ \vdots \\ \hat{\theta}_{m,q+p} \end{bmatrix} = \begin{bmatrix} \hat{\theta}_{mq} & \hat{\theta}_{m,q-1} & \cdots & \hat{\theta}_{m,q+1-p} \\ \hat{\theta}_{m,q+1} & \hat{\theta}_{m,q} & \cdots & \hat{\theta}_{m,q+2-p} \\ \vdots & \vdots & & \vdots \\ \hat{\theta}_{m,q+p-1} & \hat{\theta}_{m,q+p-2} & \cdots & \hat{\theta}_{m,q} \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix}. \tag{5.1.27}$$

Having solved (5.1.27) for $\hat{\boldsymbol{\phi}}$ (which may not be causal), we can easily determine the estimate of $\boldsymbol{\theta}$ from

$$\hat{\theta}_j = \hat{\theta}_{mj} - \sum_{i=1}^{\min(j,\, p)} \hat{\phi}_i \hat{\theta}_{m,\, j-i}, \quad j = 1, \ldots, q.$$

Finally, the white noise variance σ^2 is estimated by

$$\hat{\sigma}^2 = n^{-1} \sum_{t=1}^{n} \left(X_t - \hat{X}_t \right)^2 / r_{t-1},$$

where $\hat{X}_t$ is the one-step predictor of X_t computed from the fitted coefficient vectors $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$, and r_{t-1} is defined in (3.3.8).

The above calculations can all be carried out by selecting the ITSM option `Model>Estimation>Preliminary`. This option also computes, if $p = q$, the ratio of each estimated coefficient to 1.96 times its estimated standard deviation. Approximate 95 % confidence intervals can therefore easily be obtained in this case. If the fitted model is noncausal, it cannot be used to initialize the search for the maximum likelihood estimators, and so the autoregressive coefficients should be set to some causal values (e.g., all equal to 0.001) using the `Model>Specify` option. If both the innovation and Hannan–Rissanen algorithms give noncausal models, it is an indication (but not a conclusive one) that the assumed values of p and q may not be appropriate for the data.

Order Selection for Mixed Models

For models with $p > 0$ and $q > 0$, the sample ACF and PACF are difficult to recognize and are of far less value in order selection than in the special cases where $p = 0$ or $q = 0$. A systematic approach, however, is still available through minimization of the AICC statistic

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

which is discussed in more detail in Section 5.5. For fixed p and q it is clear from the definition that the AICC value is minimized by the parameter values that maximize the likelihood. Hence, final decisions regarding the orders p and q that minimize AICC must be based on maximum likelihood estimation as described in Section 5.2.

Example 5.1.6 The Lake Data

In Example 5.1.4 we fitted AR(2) models to the mean-corrected lake data using the Yule–Walker equations and Burg's algorithm. If instead we fit an ARMA(1,1) model using the innovations method in the option `Model>Estimation>Preliminary` of ITSM (with the default value $m = 17$), we obtain the model

$$X_t - 0.7234X_{t-1} = Z_t + 0.3596Z_{t-1}, \quad \{Z_t\} \sim \text{WN}(0, 0.4757),$$

for the mean-corrected series $X_t = Y_t - 9.0041$. The ratio of the two coefficient estimates $\hat{\phi}$ and $\hat{\theta}$ to 1.96 times their estimated standard deviations are given by ITSM as 3.2064 and 1.8513, respectively. The corresponding 95 % confidence intervals are therefore

$$\phi : 0.7234 \pm 0.7234/3.2064 = (0.4978, 0.9490),$$

$$\theta : 0.3596 \pm 0.3596/1.8513 = (0.1654, 0.5538).$$

It is interesting to note that the value of AICC for this model is 212.89, which is smaller than the corresponding values for the Burg and Yule–Walker AR(2) models in Example 5.1.4. This suggests that an ARMA(1,1) model may be superior to a pure autoregressive model for these data. Preliminary estimation of a variety of ARMA(p, q) models shows that the minimum AICC value does in fact occur when $p = q = 1$. (Before committing ourselves to this model, however, we need to compare AICC values for the corresponding *maximum likelihood* models. We shall do this in Section 5.2.)

□

5.1.4 The Hannan–Rissanen Algorithm

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 $Z_{t-1}, \ldots, Z_{t-q}$ in (5.1.1) 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 5.1.1. If $\left(\hat{\phi}_{m1}, \ldots, \hat{\phi}_{mm}\right)'$ is the vector of estimated coefficients, then the estimated residuals are computed from the equations

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

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, $\boldsymbol{\beta} = \left(\boldsymbol{\phi}', \boldsymbol{\theta}'\right)'$ is estimated by least squares linear regression of X_t onto $\left(X_{t-1}, \ldots, X_{t-p}, \hat{Z}_{t-1}, \ldots, \hat{Z}_{t-q}\right)$, $t = m+1+q, \ldots, n$, i.e., by minimizing the sum of squares

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

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

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

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

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

(If $p = 0$, Z contains only the last q columns.) The Hannan–Rissanen estimate of the white noise variance is

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

Example 5.1.7 The Lake Data

In Example 5.1.6 an ARMA(1,1) model was fitted to the mean corrected lake data using the innovations algorithm. We can fit an ARMA(1,1) model to these data using the Hannan–Rissanen estimates by selecting `Hannan-Rissanen` in the *Preliminary Estimation* dialog box of ITSM. The fitted model is

$$X_t - 0.6961X_{t-1} = Z_t + 0.3788Z_{t-1}, \quad \{Z_t\} \sim \mathrm{WN}(0, 0.4774),$$

for the mean-corrected series $X_t = Y_t - 9.0041$. (Two estimates of the white noise variance are computed in ITSM for the Hannan–Rissanen procedure, $\hat{\sigma}^2_{\mathrm{HR}}$ and $\sum_{j=1}^{n}(X_t - \hat{X}_{t-1})^2/n$. The latter is the one retained by the program.) The ratios of the two coefficient estimates to 1.96 times their standard deviation are 4.5289 and 1.3120, respectively. The corresponding 95 % confidence bounds for ϕ and θ are

$$\phi : 0.6961 \pm 0.6961/4.5289 = (0.5424, 0.8498),$$

$$\theta : 0.3788 \pm 0.3788/1.3120 = (0.0901, 0.6675).$$

Clearly, there is little difference between this model and the one fitted using the innovations method in Example 5.1.6. (The AICC values are 213.18 for the current model and 212.89 for the model fitted in Example 5.1.6.)

□

Hannan and Rissanen include a third step in their procedure to improve the estimates.

Step 3. Using the estimate $\hat{\boldsymbol{\beta}} = \left(\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q\right)'$ from Step 2, set

$$\tilde{Z}_t = \begin{cases} 0, & \text{if } t \le \max(p, q), \\ X_t - \sum_{j=1}^{p} \hat{\phi}_j X_{t-j} - \sum_{j=1}^{q} \hat{\theta}_j \tilde{Z}_{t-j}, & \text{if } t > \max(p, q). \end{cases}$$

Now for $t = 1, \ldots, n$ put

$$V_t = \begin{cases} 0, & \text{if } t \le \max(p, q), \\ \sum_{j=1}^{p} \hat{\phi}_j V_{t-j} + \tilde{Z}_t, & \text{if } t > \max(p, q), \end{cases}$$

and

$$W_t = \begin{cases} 0, & \text{if } t \le \max(p, q), \\ -\sum_{j=1}^{p} \hat{\theta}_j W_{t-j} + \tilde{Z}_t, & \text{if } t > \max(p, q). \end{cases}$$

(Observe that both V_t and W_t satisfy the AR recursions $\hat{\phi}(B)V_t = \tilde{Z}_t$ and $\hat{\theta}(B)W_t = \tilde{Z}_t$ for $t = 1, \ldots, n$.) If $\hat{\boldsymbol{\beta}}^\dagger$ is the regression estimate of $\boldsymbol{\beta}$ found by regressing $\tilde{Z}_t$ on $(V_{t-1}, \ldots, V_{t-p}, W_{t-1}, \ldots, W_{t-q})$, i.e., if $\hat{\boldsymbol{\beta}}^\dagger$ minimizes

$$S^\dagger(\beta) = \sum_{t=\max(p,q)+1}^{n} \left(\tilde{Z}_t - \sum_{j=1}^{p} \beta_j V_{t-j} - \sum_{k=1}^{q} \beta_{k+p} W_{t-k} \right)^2,$$

then the improved estimate of $\boldsymbol{\beta}$ is $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^\dagger + \hat{\boldsymbol{\beta}}$. The new estimator $\tilde{\boldsymbol{\beta}}$ then has the same asymptotic efficiency as the maximum likelihood estimator. In ITSM, however, we eliminate Step 3, using the model produced by Step 2 as the initial model for the calculation (by numerical maximization) of the maximum likelihood estimator itself.

5.2 Maximum Likelihood Estimation

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

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

$$L(\Gamma_n) = (2\pi)^{-n/2} (\det \Gamma_n)^{-1/2} \exp\left(-\frac{1}{2} \mathbf{X}_n' \Gamma_n^{-1} \mathbf{X}_n \right). \tag{5.2.1}$$

As we shall now show, the direct calculation of $\det \Gamma_n$ and Γ_n^{-1} can be avoided by expressing this in terms of the one-step prediction errors $X_j - \hat{X}_j$ and their variances $v_{j-1}, j = 1, \ldots, n$, both of which are easily calculated recursively from the innovations algorithm (Section 2.5.4).

Let $\theta_{ij}, j = 1, \ldots, i; i = 1, 2, \ldots$, denote the coefficients obtained when the innovations algorithm is applied to the autocovariance function κ of $\{X_t\}$, and let C_n be the $n \times n$ lower triangular matrix defined in Section 2.5.4. From (2.5.27) we have the identity

$$\mathbf{X}_n = C_n\left(\mathbf{X}_n - \hat{\mathbf{X}}_n\right). \tag{5.2.2}$$

We also know from Remark 5 of Section 2.5.4 that the components of $\mathbf{X}_n - \hat{\mathbf{X}}_n$ are uncorrelated. Consequently, by the definition of v_j, $\mathbf{X}_n - \hat{\mathbf{X}}_n$ has the diagonal covariance matrix

$$D_n = \operatorname{diag}\{v_0, \ldots, v_{n-1}\}.$$

From (5.2.2) and (A.2.5) we conclude that

$$\Gamma_n = C_n D_n C_n'. \tag{5.2.3}$$

From (5.2.2) and (5.2.3) we see that

$$\mathbf{X}_n' \Gamma_n^{-1} \mathbf{X}_n = \left(\mathbf{X}_n - \hat{\mathbf{X}}_n\right)' D_n^{-1}\left(\mathbf{X}_n - \hat{\mathbf{X}}_n\right) = \sum_{j=1}^{n}\left(X_j - \hat{X}_j\right)^2 / v_{j-1} \tag{5.2.4}$$

and

$$\det \Gamma_n = (\det C_n)^2 (\det D_n) = v_0 v_1 \cdots v_{n-1}. \tag{5.2.5}$$

The likelihood (5.2.1) of the vector $\mathbf{X}_n$ therefore reduces to

$$L(\Gamma_n) = \frac{1}{\sqrt{(2\pi)^n v_0 \cdots v_{n-1}}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\left(X_j - \hat{X}_j\right)^2 / v_{j-1}\right\}. \tag{5.2.6}$$

If Γ_n is expressible in terms of a finite number of unknown parameters $\beta_1, \ldots, \beta_r$ (as is the case when $\{X_t\}$ is an ARMA(p, q) process), the **maximum likelihood estimators** of the parameters are those values that maximize L for the given data set. When $X_1, X_2, \ldots, X_n$ are iid, it is known, under mild assumptions and for n large, that maximum likelihood estimators are approximately normally distributed with variances that are at least as small as those of other asymptotically normally distributed estimators (see, e.g., Lehmann 1983).

Even if $\{X_t\}$ is not Gaussian, it still makes sense to regard (5.2.6) as a measure of goodness of fit of the model to the data, and to choose the parameters $\beta_1, \ldots, \beta_r$ in such a way as to maximize (5.2.6). We shall always refer to the estimators $\hat{\beta}_1, \ldots, \hat{\beta}_r$ so obtained as "maximum likelihood" estimators, even when $\{X_t\}$ is not Gaussian. Regardless of the joint distribution of $X_1, \ldots, X_n$, we shall refer to (5.2.1) and its algebraic equivalent (5.2.6) as the "likelihood" (or "Gaussian likelihood") of $X_1, \ldots, X_n$. A justification for using maximum Gaussian likelihood estimators of ARMA coefficients is that the large-sample distribution of the estimators is the same for $\{Z_t\} \sim \text{IID}(0, \sigma^2)$, regardless of whether or not $\{Z_t\}$ is Gaussian (see Brockwell and Davis (1991), Section 10.8).

The likelihood for data from an ARMA(p, q) process is easily computed from the innovations form of the likelihood (5.2.6) by evaluating the one-step predictors $\hat{X}_{i+1}$ and the corresponding mean squared errors v_i. These can be found from the recursions (Section 3.3)

$$\hat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{nj}\left(X_{n+1-j} - \hat{X}_{n+1-j}\right), & 1 \le n < m, \\ \phi_1 X_n + \cdots + \phi_p X_{n+1-p} + \sum_{j=1}^{q} \theta_{nj}\left(X_{n+1-j} - \hat{X}_{n+1-j}\right), & n \ge m, \end{cases} \tag{5.2.7}$$

and

$$E\left(X_{n+1}-\hat{X}_{n+1}\right)^2=\sigma^2E\left(W_{n+1}-\hat{W}_{n+1}\right)^2=\sigma^2r_n, \tag{5.2.8}$$

where θ_{nj} and r_n are determined by the innovations algorithm with κ as in (3.3.3) and $m=\max(p,q)$. Substituting in the general expression (5.2.6), we obtain the following:

The Gaussian Likelihood for an ARMA Process:

$$L\left(\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2\right)=\frac{1}{\sqrt{\left(2\pi\sigma^2\right)^n r_0\cdots r_{n-1}}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\frac{\left(X_j-\hat{X}_j\right)^2}{r_{j-1}}\right\}. \tag{5.2.9}$$

Differentiating $\ln L\left(\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2\right)$ 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 the following equations (Problem 5.8):

Maximum Likelihood Estimators:

$$\hat{\sigma}^2=n^{-1}S\left(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}}\right), \tag{5.2.10}$$

where

$$S\left(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}}\right)=\sum_{j=1}^{n}\left(X_j-\hat{X}_j\right)^2/r_{j-1}, \tag{5.2.11}$$

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

$$\ell(\boldsymbol{\phi},\boldsymbol{\theta})=\ln\left(n^{-1}S(\boldsymbol{\phi},\boldsymbol{\theta})\right)+n^{-1}\sum_{j=1}^{n}\ln r_{j-1}. \tag{5.2.12}$$

Minimization of $\ell(\boldsymbol{\phi},\boldsymbol{\theta})$ must be done numerically. Initial values for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can be obtained from ITSM using the methods described in Section 5.1. The program then searches systematically for the values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ that minimize the reduced likelihood (5.2.12) and computes the corresponding maximum likelihood estimate of σ^2 from (5.2.10).

Least Squares Estimation for Mixed Models

The least squares estimates $\tilde{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are obtained by minimizing the function S as defined in (5.2.11) rather than ℓ as defined in (5.2.12), subject to the constraints that the model be causal and invertible. The least squares estimate of σ^2 is

$$\tilde{\sigma}^2=\frac{S\left(\tilde{\phi},\tilde{\theta}\right)}{n-p-q}.$$

Order Selection

In Section 5.1 we introduced minimization of the AICC value as a major criterion for the selection of the orders p and q. This criterion is applied as follows:

AICC Criterion:
Choose p, q, $\boldsymbol{\phi}_p$, and $\boldsymbol{\theta}_q$ to minimize

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

For any fixed p and q it is clear that the AICC is minimized when $\boldsymbol{\phi}_p$ and $\boldsymbol{\theta}_q$ are the vectors that minimize $-2\ln L(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q, S(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q)/n)$, i.e., the maximum likelihood estimators. Final decisions with respect to order selection should therefore be made on the basis of maximum likelihood estimators (rather than the preliminary estimators of Section 5.1, which serve primarily as a guide). The AICC statistic and its justification are discussed in detail in Section 5.5.

One of the options in the program ITSM is `Model>Estimation>Autofit`. Selection of this option allows you to specify a range of values for both p and q, after which the program will automatically fit maximum likelihood ARMA(p, q) values for all p and q in the specified range, and select from these the model with smallest AICC value. This may be slow if a large range is selected (the maximum range is from 0 through 27 for both p and q), and once the model has been determined, it should be checked by preliminary estimation followed by maximum likelihood estimation to minimize the risk of the fitted model corresponding to a local rather than a global maximum of the likelihood. (For more details see Section E.3.1.)

Confidence Regions for the Coefficients

For large sample size the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta} := (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$ is approximately normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix $\left[n^{-1}V(\boldsymbol{\beta})\right]$ which can be approximated by $2H^{-1}(\boldsymbol{\beta})$, where H is the Hessian matrix $\left[\partial^2\ell(\boldsymbol{\beta})/\partial\beta_i\partial\beta_j\right]_{i,j=1}^{p+q}$. ITSM prints out the approximate standard deviations and correlations of the coefficient estimators based on the Hessian matrix evaluated numerically at $\hat{\boldsymbol{\beta}}$ unless this matrix is not positive definite, in which case ITSM instead computes the theoretical asymptotic covariance matrix in Section 9.8 of Brockwell and Davis (1991). The resulting covariances can be used to compute confidence bounds for the parameters.

Large-Sample Distribution of Maximum Likelihood Estimators:

For a large sample from an ARMA(p, q) process,

$$\hat{\boldsymbol{\beta}} \approx \text{N}\left(\boldsymbol{\beta}, n^{-1}V(\boldsymbol{\beta})\right).$$

The general form of $V(\boldsymbol{\beta})$ can be found in Brockwell and Davis (1991), Section 9.8. The following are several special cases.

Example 5.2.1 An AR(p) Model

The asymptotic covariance matrix in this case is the same as that for the Yule–Walker estimates given by

$$V(\boldsymbol{\phi}) = \sigma^2\Gamma_p^{-1}.$$

In the special cases $p = 1$ and $p = 2$, we have

$$\text{AR}(1): V(\phi) = \left(1 - \phi_1^2\right),$$

$$\text{AR}(2): V(\phi) = \begin{bmatrix} 1 - \phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1 - \phi_2^2 \end{bmatrix}.$$

□

Example 5.2.2 An MA(q) Model

Let Γ_q^* be the covariance matrix of $Y_1, \ldots, Y_q$, where $\{Y_t\}$ is the autoregressive process with autoregressive polynomial $\theta(z)$, i.e.,

$$Y_t + \theta_1 Y_{t-1} + \cdots + \theta_q Y_{t-q} = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 1).$$

Then it can be shown that

$$V(\boldsymbol{\theta}) = \Gamma_q^{*-1}.$$

Inspection of the results of Example 5.2.1 and replacement of ϕ_i by $-\theta_i$ yields

$$\text{MA}(1): V(\boldsymbol{\theta}) = \left(1 - \theta_1^2\right),$$

$$\text{MA}(2): V(\boldsymbol{\theta}) = \begin{bmatrix} 1 - \theta_2^2 & \theta_1(1-\theta_2) \\ \theta_1(1-\theta_2) & 1 - \theta_2^2 \end{bmatrix}.$$

□

Example 5.2.3 An ARMA(1, 1) Model

For a causal and invertible ARMA(1,1) process with coefficients ϕ and θ.

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

□

Example 5.2.4 The Dow Jones Utilities Index

For the Burg and Yule–Walker AR(1) models derived for the differenced and mean-corrected series in Examples 5.1.1 and 5.1.3, the `Model>Estimation>Preliminary` option of ITSM gives $-2\ln(L) = 70.330$ for the Burg model and $-2\ln(L) = 70.378$ for the Yule–Walker model. Since maximum likelihood estimation attempts to minimize $-2\ln L$, the Burg estimate appears to be a slightly better initial estimate of ϕ. We therefore retain the Burg AR(1) model and then select `Model>Estimation>Max Likelihood` and click `OK`. The Burg coefficient estimates provide initial parameter values to start the search for the minimizing values. The model found on completion of the minimization is

$$Y_t - 0.4471 Y_{t-1} = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 0.02117). \tag{5.2.13}$$

This model is different again from the Burg and Yule–Walker models. It has $-2\ln(L) = 70.321$, corresponding to a slightly higher likelihood. The standard error (or estimated standard deviation) of the estimator $\hat{\phi}$ is found from the program to be 0.1050. This is close to the estimated standard deviation $\sqrt{(1-(0.4471)^2)/77} = 0.1019$, based on the large-sample approximation given in Example 5.2.1. Using the value computed from ITSM, approximate 95 % confidence bounds for ϕ are $0.4471 \pm 1.96 \times 0.1050 = (0.2413, 0.6529)$. These are quite close to the bounds based on the Yule–Walker and Burg estimates found in Examples 5.1.1 and 5.1.3. To find the minimum-AICC model for the series $\{Y_t\}$ using ITSM, choose the option `Model>Estimation>Autofit`. Using the default range for both p and

q, and clicking on `Start`, we quickly find that the minimum AICC ARMA(p, q) model with $p \leq 5$ and $q \leq 5$ is the AR(1) model defined by (5.2.13). The corresponding AICC value is 74.483. If we increase the upper limits for p and q, we obtain the same result.

□

Example 5.2.5 The Lake Data

Using the option `Model>Estimation>Autofit` as in the previous example, we find that the minimum-AICC ARMA(p, q) model for the mean-corrected lake data, $X_t = Y_t - 9.0041$, of Examples 5.1.6 and 5.1.7 is the ARMA(1,1) model

$$X_t - 0.7446X_{t-1} = Z_t + 0.3213Z_{t-1}, \quad \{Z_t\} \sim \text{WN}(0, 0.4750). \tag{5.2.14}$$

The estimated standard deviations of the two coefficient estimates $\hat{\phi}$ and $\hat{\theta}$ are found from ITSM to be 0.0773 and 0.1123, respectively. (The respective estimated standard deviations based on the large-sample approximation given in Example 5.2.3 are 0.0788 and 0.1119.) The corresponding 95 % confidence bounds are therefore

$$\phi : 0.7446 \pm 1.96 \times 0.0773 = (0.5941, 0.8961),$$

$$\theta : 0.3208 \pm 1.96 \times 0.1123 = (0.1007, 0.5409).$$

The value of AICC for this model is 212.77, improving on the values for the preliminary models of Examples 5.1.4, 5.1.6, and 5.1.7.

□

5.3 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 3.3, by

$$\hat{W}_t = \left(X_t - \hat{X}_t\left(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}\right)\right) / \left(r_{t-1}\left(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}\right)\right)^{1/2}, \quad t = 1, \ldots, n. \tag{5.3.1}$$

If we were to assume that the maximum likelihood ARMA(p, q) model is the *true* process generating $\{X_t\}$, then we could say that $\left\{\hat{W}_t\right\} \sim \text{WN}\left(0, \hat{\sigma}^2\right)$. However, to check the appropriateness of an ARMA(p, q) model for the data we should assume only that $X_1, \ldots, X_n$ are generated by an ARMA(p, q) process with unknown parameters $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and σ^2, whose maximum likelihood *estimators* are $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\sigma}^2$, respectively. Then it is not true that $\left\{\hat{W}_t\right\}$ is white noise. Nonetheless $\hat{W}_t$, $t = 1, \ldots, n$, should have properties that are similar to those of the white noise sequence

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

Moreover, $W_t(\boldsymbol{\phi}, \boldsymbol{\theta})$ approximates the white noise term in the defining equation (5.1.1) in the sense that $E(W_t(\boldsymbol{\phi}, \boldsymbol{\theta}) - Z_t)^2 \to 0$ as $t \to \infty$ (Brockwell and Davis (1991), Section 8.11). Consequently, the properties of the residuals $\left\{\hat{W}_t\right\}$ 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 (1) uncorrelated if $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, (2) independent if $\{Z_t\} \sim \text{IID}(0, \sigma^2)$, and (3) normally distributed if $Z_t \sim \text{N}(0, \sigma^2)$.

The **rescaled residuals** $\hat{R}_t, t = 1, \ldots, n$, are obtained by dividing the residuals $\hat{W}_t, t = 1, \ldots, n$, by the estimate $\hat{\sigma} = \sqrt{\left(\sum_{t=1}^{n} \hat{W}_t^2\right)/n}$ of the white noise standard deviation. Thus,

$$\hat{R}_t = \hat{W}_t/\hat{\sigma}. \tag{5.3.2}$$

If the fitted model is appropriate, the rescaled residuals should have properties similar to those of a WN(0, 1) sequence or of an iid(0,1) sequence if we make the stronger assumption that the white noise $\{Z_t\}$ driving the ARMA process is independent white noise.

The following diagnostic checks are all based on the expected properties of the residuals or rescaled residuals under the assumption that the fitted model is correct and that $\{Z_t\} \sim \text{IID}(0, \sigma^2)$. They are the same tests introduced in Section 1.6.

5.3.1 The Graph of $\{\hat{R}_t, t = 1, \ldots, n\}$

If the fitted model is appropriate, then the graph of the rescaled residuals $\{\hat{R}_t, t = 1, \ldots, n\}$ should resemble that of a white noise sequence with variance one. While it is difficult to identify the correlation structure of $\{\hat{R}_t\}$ (or any time series for that matter) from its graph, deviations of the mean from zero are sometimes clearly indicated by a trend or cyclic component and nonconstancy of the variance by fluctuations in $\hat{R}_t$, whose magnitude depends strongly on t.

The rescaled residuals obtained from the ARMA(1,1) model fitted to the mean-corrected lake data in Example 5.2.5 are displayed in Figure 5-5. The graph gives no indication of a nonzero mean or nonconstant variance, so on this basis there is no reason to doubt the compatibility of $\hat{R}_1, \ldots, \hat{R}_n$ with unit-variance white noise.

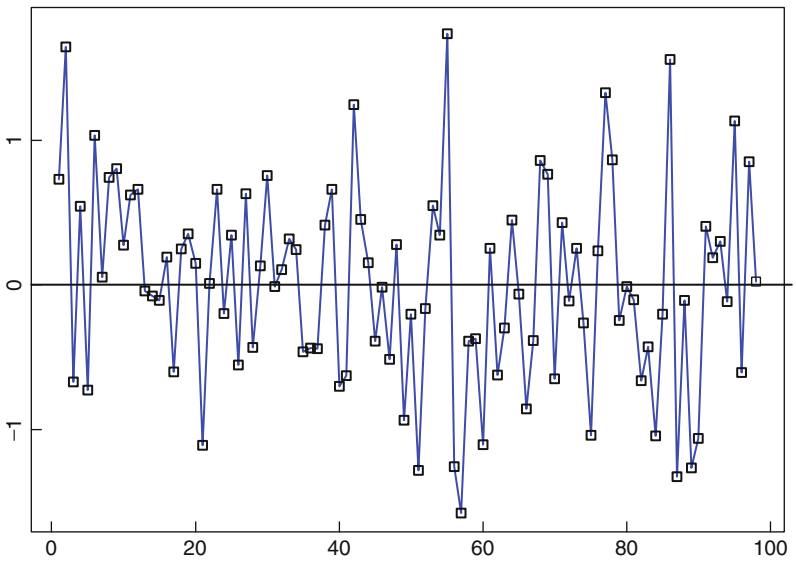

Figure 5-5
The rescaled residuals after fitting the ARMA(1,1) model of Example 5.2.5 to the lake data

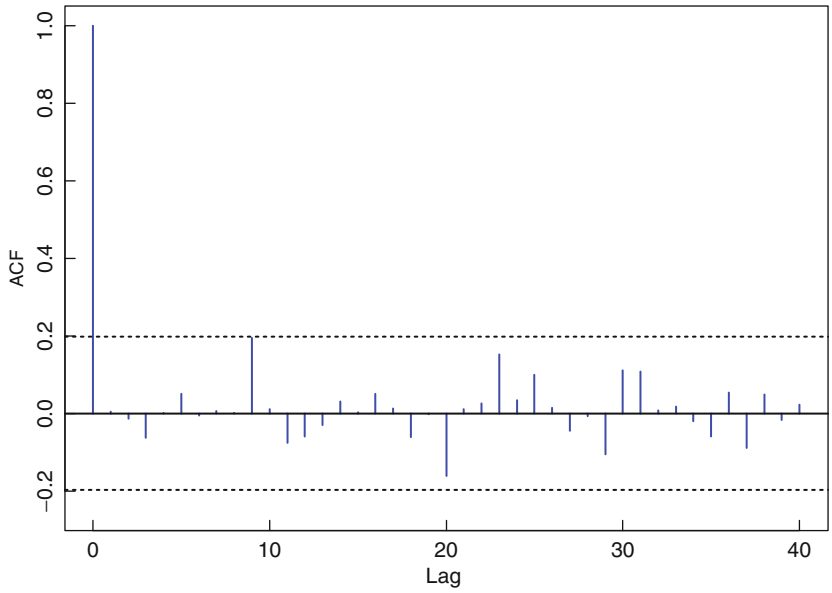

Figure 5-6
The sample ACF of the residuals after fitting the ARMA(1,1) model of Example 5.2.5 to the lake data

The next step is to check that the sample autocorrelation function of $\{\hat{W}_t\}$ (or equivalently of $\{\hat{R}_t\}$) behaves as it should under the assumption that the fitted model is appropriate.

5.3.2 The Sample ACF of the Residuals

We know from Section 1.6 that for large n the sample autocorrelations of an iid sequence $Y_1, \ldots, Y_n$ with finite variance are approximately iid with distribution $N(0, 1/n)$. 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. (As indicated above, our *estimated* residuals will not be precisely iid even if the true model generating the data is as assumed. To correct for this the bounds $\pm 1.96/\sqrt{n}$ should be modified to give a more precise test as in Box and Pierce (1970) and Brockwell and Davis (1991), Section 9.4.) The sample ACF and PACF of the residuals and the bounds $\pm 1.96/\sqrt{n}$ can be viewed by pressing the second green button (`Plot ACF/PACF of residuals`) at the top of the ITSM window. Figure 5-6 shows the sample ACF of the residuals after fitting the ARMA(1,1) of Example 5.2.5 to the lake data. As can be seen from the graph, there is no cause to reject the fitted model on the basis of these autocorrelations.

5.3.3 Tests for Randomness of the Residuals

The tests (b), (c), (d), (e), and (f) of Section 1.6 can be carried out using the program ITSM by selecting `Statistics>Residual Analysis>Tests of Randomness`.
Applying these tests to the residuals from the ARMA(1,1) model for the mean-corrected lake data (Example 5.2.5), and using the default value $h = 22$ suggested for the portmanteau tests, we obtain the following results:

RANDOMNESS TEST STATISTICS
LJUNG-BOX PORTM. = 10.23 CHISQUR(20) p=0.964
MCLEOD-LI PORTM. = 16.55 CHISQUR(22) p=0.788
TURNING POINTS = 69 ANORMAL(64.0, 4.14**2) p=0.227
DIFFERENCE-SIGN = 50 ANORMAL(48.5, 2.87**2) p=0.602
RANK TEST = 2083 ANORMAL(2376, 488.7**2) p=0.072
JARQUE-BERA=0.285 CHISQUR(2) p=0.867
ORDER OF MIN AICC YW MODEL FOR RESIDUALS = 0

This table shows the observed values of the statistics defined in Section 1.6, with each followed by its large-sample distribution under the null hypothesis of iid residuals, and the corresponding p-values. The observed values can thus be checked easily for compatibility with their distributions under the null hypothesis. Since all of the p-values are greater than 0.05, none of the test statistics leads us to reject the null hypothesis at this level. The order of the minimum AICC autoregressive model for the residuals also suggests the compatibility of the residuals with white noise.

A rough check for normality is provided by visual inspection of the histogram of the rescaled residuals, obtained by selecting the third green button at the top of the ITSM window. A Gaussian qq-plot of the residuals can also be plotted by selecting `Statistics > Residual Analysis > QQ-Plot (normal)`. No obvious deviation from normality is apparent in either the histogram or the qq-plot. The Jarque-Bera statistic, $n[m_3^2/(6m_2^3)+(m_4/m_2^3-3)^2/24]$, where $m_r = \sum_{j=1}^{n}(Y_j-\bar{Y})^r/n$, is distributed asymptotically as $\chi^2(2)$ if $\{Y_t\} \sim$ IID $N(\mu, \sigma^2)$. This hypothesis is rejected if the statistic is sufficiently large (at level α if the p-value of the test is less than α). In this case the large p-value computed by ITSM provides no evidence for rejecting the normality hypothesis.

5.4 Forecasting

Once a model has been fitted to the data, forecasting future values of the time series can be carried out using the method described in Section 3.3. We illustrate this method with one of the examples from Section 3.2.

Example 5.4.1 For the overshort data $\{X_t\}$ of Example 3.2.8, selection of the options `Model> Estimation >Preliminary`, the innovations algorithm, and then `Model> Estimation>Max likelihood`, gives the maximum likelihood MA(1) model for $\{X_t\}$,

$$X_t + 4.035 = Z_t - 0.818Z_{t-1}, \quad \{Z_t\} \sim \text{WN}(0, 2040.75). \tag{5.4.1}$$

To predict the next 7 days of overshorts, we treat (5.4.1) as the true model for the data, and use the results of Example 3.3.3 with $\phi = 0$. From (3.3.11), the predictors are given by

$$\begin{aligned} P_{57}X_{57+h} &= -4.035 + \sum_{j=h}^{1} \theta_{57+h-1,j}\left(X_{57+h-j} - \hat{X}_{57+h-j}\right) \\ &= \begin{cases} -4.035 + \theta_{57,1}\left(X_{57} - \hat{X}_{57}\right), & \text{if } h = 1, \\ -4.035, & \text{if } h > 1, \end{cases} \end{aligned}$$

Table 5.1 Forecasts of the next seven observations of the overshort data of Example 3.2.8 using model (5.4.1)

#	XHAT	SQRT(MSE)	XHAT + MEAN
58	1.0097	45.1753	−3.0254
59	0.0000	58.3602	−4.0351
60	0.0000	58.3602	−4.0351
61	0.0000	58.3602	−4.0351
62	0.0000	58.3602	−4.0351
63	0.0000	58.3602	−4.0351
64	0.0000	58.3602	−4.0351

with mean squared error

$$E(X_{57+h} - P_{57}X_{57+h})^2 = \begin{cases} 2040.75 r_{57}, & \text{if } h = 1, \\ 2040.75(1 + (-0.818)^2), & \text{if } h > 1, \end{cases}$$

where $\theta_{57,1}$ and r_{57} are computed recursively from (3.3.9) with $\theta = -0.818$.

These calculations are performed with ITSM by fitting the maximum likelihood model (5.4.1), selecting `Forecasting>ARMA`, and specifying the number of forecasts required. The 1-step, 2-step, ..., and 7-step forecasts of X_t are shown in Table 5.1. Notice that the predictor of X_t for $t \geq 59$ is equal to the sample mean, since under the MA(1) model $\{X_t, t \geq 59\}$ is uncorrelated with $\{X_t, t \leq 57\}$.

Assuming that the innovations $\{Z_t\}$ are normally distributed, an approximate 95 % prediction interval for X_{64} is given by

$$-4.0351 \pm 1.96 \times 58.3602 = (-118.42, 110.35).$$

□

The mean squared errors of prediction, as computed in Section 3.3 and the example above, are based on the assumption that the fitted model is in fact the true model for the data. As a result, they do not reflect the variability in the estimation of the model parameters. To illustrate this point, suppose the data $X_1, \ldots, X_n$ are generated from the causal AR(1) model

$$X_t = \phi X_{t-1} + Z_t, \quad \{Z_t\} \sim \text{iid}\left(0, \sigma^2\right).$$

If $\hat{\phi}$ is the maximum likelihood estimate of ϕ, based on $X_1, \ldots, X_n$, then the one-step ahead forecast of X_{n+1} is $\hat{\phi}X_n$, which has mean squared error

$$E\left(X_{n+1} - \hat{\phi}X_n\right)^2 = E\left(\left(\phi - \hat{\phi}\right)X_n + Z_{n+1}\right)^2 = E((\phi - \hat{\phi})X_n)^2 + \sigma^2. \tag{5.4.2}$$

The second equality follows from the independence of Z_{n+1} and $\left(\hat{\phi}, X_n\right)'$. To evaluate the first term in (5.4.2), first condition on X_n and then use the approximations

$$E\left(\left(\phi - \hat{\phi}\right)^2 | X_n\right) \approx E\left(\phi - \hat{\phi}\right)^2 \approx \left(1 - \phi^2\right)/n,$$

where the second relation comes from the formula for the asymptotic variance of $\hat{\phi}$ given by $\sigma^2\Gamma_1^{-1} = \left(1 - \phi^2\right)$ (see Example 5.2.1). The one-step mean squared error is then approximated by

$$E\left(\phi-\hat{\phi}\right)^2 EX_n^2+\sigma^2 \approx n^{-1}\left(1-\phi^2\right)\left(1-\phi^2\right)^{-1}\sigma^2+\sigma^2=\frac{n+1}{n}\sigma^2.$$

Thus, the error in parameter estimation contributes the term σ^2/n to the mean squared error of prediction. If the sample size is large, this factor is negligible, and so for the purpose of mean squared error computation, the estimated parameters can be treated as the true model parameters. On the other hand, for small sample sizes, ignoring parameter variability can lead to a severe underestimate of the actual mean squared error of the forecast.

5.5 Order Selection

Once the data have been transformed (e.g., by some combination of Box–Cox and differencing transformations or by removal of trend and seasonal components) to the point where the transformed series $\{X_t\}$ can potentially be fitted by a zero-mean ARMA model, we are faced with the problem of selecting appropriate values for the orders p and q.

It is not advantageous from a forecasting point of view to choose p and q arbitrarily large. Fitting a very high order model will generally result in a small estimated white noise variance, but when the fitted model is used for forecasting, the mean squared error of the forecasts will depend not only on the white noise variance of the fitted model but also on errors arising from estimation of the parameters of the model (see the paragraphs following Example 5.4.1). These will be larger for higher-order models. For this reason we need to introduce a "penalty factor" to discourage the fitting of models with too many parameters.

Many criteria based on such penalty factors have been proposed in the literature, since the problem of model selection arises frequently in statistics, particularly in regression analysis. We shall restrict attention here to a brief discussion of the FPE, AIC, and BIC criteria of Akaike and a bias-corrected version of the AIC known as the AICC.

5.5.1 The FPE Criterion

The FPE criterion was developed by Akaike (1969) to select the appropriate order of an AR process to fit to a time series $\{X_1,\ldots,X_n\}$. Instead of trying to choose the order p to make the estimated white noise variance as small as possible, the idea is to choose the model for $\{X_t\}$ in such a way as to minimize the one-step mean squared error when the model fitted to $\{X_t\}$ is used to predict an independent realization $\{Y_t\}$ of the same process that generated $\{X_t\}$.

Suppose then that $\{X_1,\ldots,X_n\}$ is a realization of an AR(p) process with coefficients $\phi_1,\ldots,\phi_p$, $p<n$, and that $\{Y_1,\ldots,Y_n\}$ is an independent realization of the same process. If $\hat{\phi}_1,\ldots,\hat{\phi}_p$, are the maximum likelihood estimators of the coefficients based on $\{X_1,\ldots,X_n\}$ and if we use these to compute the one-step predictor $\hat{\phi}_1 Y_n+\cdots+\hat{\phi}_p Y_{n+1-p}$ of Y_{n+1}, then the mean square prediction error is

$$\begin{aligned}
&E\left(Y_{n+1}-\hat{\phi}_1 Y_n-\cdots-\hat{\phi}_p Y_{n+1-p}\right)^2\\
&\quad=E\left[Y_{n+1}-\phi_1 Y_n-\cdots-\phi_p Y_{n+1-p}-\left(\hat{\phi}_1-\phi_1\right)Y_n-\cdots-\left(\hat{\phi}_p-\phi_p\right)Y_{n+1-p}\right]^2\\
&\quad=\sigma^2+E\left[\left(\hat{\phi}_p-\phi_p\right)'\left[Y_{n+1-i}Y_{n+1-j}\right]_{i,j=1}^{p}\left(\hat{\phi}_p-\phi\right)\right],
\end{aligned}$$

Table 5.2 $\hat{\sigma}_p^2$ and FPE_p for AR(p) models fitted to the lake data

p	σ_p^2	FPE_p
0	1.7203	1.7203
1	0.5097	0.5202
2	0.4790	0.4989
3	0.4728	0.5027
4	0.4708	0.5109
5	0.4705	0.5211
6	0.4705	0.5318
7	0.4679	0.5399
8	0.4664	0.5493
9	0.4664	0.5607
10	0.4453	0.5465

where $\phi_p' = (\phi_1, \ldots, \phi_p)'$, $\hat{\phi}_p' = \left(\hat{\phi}_1, \ldots, \hat{\phi}_p\right)'$, and σ^2 is the white noise variance of the AR(p) model. Writing the last term in the preceding equation as the expectation of the conditional expectation given $X_1, \ldots, X_n$, and using the independence of $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$, we obtain

$$E\left(Y_{n+1} - \hat{\phi}_1 Y_n - \cdots - \hat{\phi}_p Y_{n+1-p}\right)^2 = \sigma^2 + E\left[\left(\hat{\phi}_p - \phi_p\right)' \Gamma_p \left(\hat{\phi}_p - \phi\right)\right],$$

where $\Gamma_p = E[Y_i Y_j]_{i,j=1}^p$. We can approximate the last term by assuming that the random variable $n^{-1/2}\left(\hat{\phi}_p - \phi_p\right)$ has its large-sample distribution $\text{N}\left(\mathbf{0}, \sigma^2 \Gamma_p^{-1}\right)$ as given in Example 5.21. Using Problem 5.13, we then find that

$$E\left(Y_{n+1} - \hat{\phi}_1 Y_n - \cdots - \hat{\phi}_p Y_{n+1-p}\right)^2 \approx \sigma^2\left(1 + \frac{p}{n}\right). \tag{5.5.1}$$

If $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2, then for large n, $n\hat{\sigma}^2/\sigma^2$ is distributed approximately as chi-squared with $(n - p)$ degrees of freedom (see Brockwell and Davis (1991), Section 8.9). We therefore replace σ^2 in (5.5.1) by the estimator $n\hat{\sigma}^2/(n - p)$ to get the estimated mean square prediction error of Y_{n+1},

$$\text{FPE}_p = \hat{\sigma}^2 \frac{n + p}{n - p}. \tag{5.5.2}$$

To apply the FPE criterion for autoregressive order selection we therefore choose the value of p that minimizes FPE_p as defined in (5.5.2).

Example 5.5.1 FPE-Based Selection of an AR Model for the Lake Data

In Example 5.1.4 we fitted AR(2) models to the mean-corrected lake data, the order 2 being suggested by the sample PACF shown in Figure 5-4. To use the FPE criterion to select p, we have shown in Table 5.2 the values of FPE for values of p from 0 to 10. These values were found using ITSM by fitting maximum likelihood AR models with the option `Model>Estimation>Max likelihood`. Also shown in the table are the values of the maximum likelihood estimates of σ^2 for the same values of p. Whereas $\hat{\sigma}_p^2$ decreases steadily with p, the values of FPE_p have a clear minimum at $p = 2$, confirming our earlier choice of $p = 2$ as the most appropriate for this data set. □

5.5.2 The AICC Criterion

A more generally applicable criterion for model selection than the FPE is the information criterion of Akaike (1973), known as the AIC. This was designed to be an approximately unbiased estimate of the Kullback–Leibler index of the fitted model relative to the true model (defined below). Here we use a bias-corrected version of the AIC, referred to as the AICC, suggested by Hurvich and Tsai (1989).

If $\mathbf{X}$ is an n-dimensional random vector whose probability density belongs to the family $\{f(\cdot;\psi), \psi \in \Psi\}$, the Kullback–Leibler discrepancy between $f(\cdot;\psi)$ and $f(\cdot;\theta)$ is defined as

$$d(\psi|\theta) = \Delta(\psi|\theta) - \Delta(\theta|\theta),$$

where

$$\Delta(\psi|\theta) = E_\theta(-2\ln f(\mathbf{X};\psi)) = \int_{\mathbb{R}^n} -2\ln(f(\mathbf{x};\psi))f(\mathbf{x};\theta)\,d\mathbf{x}$$

is the Kullback–Leibler index of $f(\cdot;\psi)$ relative to $f(\cdot;\theta)$. (Note that in general, $\Delta(\psi|\theta) \neq \Delta(\theta|\psi)$.) By Jensen's inequality (see, e.g., Mood et al., 1974),

$$\begin{aligned} d(\psi|\theta) &= \int_{\mathbb{R}^n} -2\ln\left(\frac{f(\mathbf{x};\psi)}{f(\mathbf{x};\theta)}\right) f(\mathbf{x};\theta)\,d\mathbf{x} \\ &\geq -2\ln\left(\int_{\mathbb{R}^n} \frac{f(\mathbf{x};\psi)}{f(\mathbf{x};\theta)} f(\mathbf{x};\theta)\,d\mathbf{x}\right) \\ &= -2\ln\left(\int_{\mathbb{R}^n} f(\mathbf{x};\psi)\,d\mathbf{x}\right) \\ &= 0, \end{aligned}$$

with equality holding if and only if $f(\mathbf{x};\psi) = f(\mathbf{x};\theta)$.

Given observations $X_1, \ldots, X_n$ of an ARMA process with unknown parameters $\theta = (\boldsymbol{\beta}, \sigma^2)$, the true model could be identified if it were possible to compute the Kullback–Leibler discrepancy between all candidate models and the true model. Since this is not possible, we *estimate* the Kullback–Leibler discrepancies and choose the model whose estimated discrepancy (or index) is minimum. In order to do this, we assume that the true model and the alternatives are all Gaussian. Then for any given $\theta = (\boldsymbol{\beta}, \sigma^2)$, $f(\cdot;\theta)$ is the probability density of $(Y_1, \ldots, Y_n)'$, where $\{Y_t\}$ is a Gaussian ARMA(p, q) process with coefficient vector $\boldsymbol{\beta}$ and white noise variance σ^2. (The dependence of θ on p and q is through the dimension of the autoregressive and moving-average coefficients in $\boldsymbol{\beta}$.)

Suppose, therefore, that our observations $X_1, \ldots, X_n$ are from a Gaussian ARMA process with parameter vector $\theta = (\boldsymbol{\beta}, \sigma^2)$ and assume for the moment that the true order is (p, q). Let $\hat{\theta} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ be the maximum likelihood estimator of θ based on $X_1, \ldots, X_n$ and let $Y_1, \ldots, Y_n$ be an independent realization of the true process (with parameter θ). Then

$$-2\ln L_Y\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) = -2\ln L_X\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) + \hat{\sigma}^{-2} S_Y\left(\hat{\boldsymbol{\beta}}\right) - n,$$

where L_X, L_Y, S_X, and S_Y are defined as in (5.2.9) and (5.2.11). Hence,

$$E_\theta(\Delta(\hat{\theta}|\theta)) = E_{\boldsymbol{\beta},\sigma^2}\left(-2\ln L_Y\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right)\right)$$

$$= E_{\beta,\sigma^2}\left(-2\ln L_X\left(\hat{\boldsymbol{\beta}},\hat{\sigma}^2\right)\right) + E_{\beta,\sigma^2}\left(\frac{S_Y\left(\hat{\boldsymbol{\beta}}\right)}{\hat{\sigma}^2}\right) - n. \tag{5.5.3}$$

It can be shown using large-sample approximations (see Brockwell and Davis (1991), Section 10.3 for details) that

$$E_{\beta,\sigma^2}\left(\frac{S_Y\left(\hat{\boldsymbol{\beta}}\right)}{\hat{\sigma}^2}\right) \approx \frac{2(p+q+1)n}{n-p-q-2},$$

from which we see that $-2\ln L_X\left(\hat{\boldsymbol{\beta}},\hat{\sigma}^2\right) + 2(p+q+1)n/(n-p-q-2)$ is an approximately unbiased estimator of the expected Kullback–Leibler index $E_\theta\left(\Delta\left(\hat{\theta}|\theta\right)\right)$ in (5.5.3). Since the preceding calculations (and the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$) are based on the assumption that the true order is (p, q), we therefore select the values of p and q for our fitted model to be those that minimize $\text{AICC}\left(\hat{\boldsymbol{\beta}}\right)$, where

$$\text{AICC}(\boldsymbol{\beta}) := -2\ln L_X(\boldsymbol{\beta}, S_X(\boldsymbol{\beta})/n) + 2(p+q+1)n/(n-p-q-2). \tag{5.5.4}$$

The AIC statistic, defined as

$$\text{AIC}(\boldsymbol{\beta}) := -2\ln L_X(\boldsymbol{\beta}, S_X(\boldsymbol{\beta})/n) + 2(p+q+1),$$

can be used in the same way. Both $\text{AICC}\left(\boldsymbol{\beta},\sigma^2\right)$ and $\text{AIC}\left(\boldsymbol{\beta},\sigma^2\right)$ can be defined for arbitrary σ^2 by replacing $S_X(\boldsymbol{\beta})/n$ in the preceding definitions by σ^2. The value $S_X(\boldsymbol{\beta})/n$ is used in (5.5.4), since $\text{AICC}\left(\boldsymbol{\beta},\sigma^2\right)$ (like $\text{AIC}\left(\boldsymbol{\beta},\sigma^2\right)$) is minimized for any given $\boldsymbol{\beta}$ by setting $\sigma^2 = S_X(\beta)/n$.

For fitting autoregressive models, Monte Carlo studies (Jones 1975; Shibata 1976) suggest that the AIC has a tendency to overestimate p. The penalty factors $2(p+q+1)n/(n-p-q-2)$ and $2(p+q+1)$ for the AICC and AIC statistics are asymptotically equivalent as $n \to \infty$. The AICC statistic, however, has a more extreme penalty for large-order models, which counteracts the overfitting tendency of the AIC. The BIC is another criterion that attempts to correct the overfitting nature of the AIC. For a zero-mean causal invertible ARMA(p, q) process, it is defined (Akaike 1978) to be

$$\text{BIC} = (n-p-q)\ln\left[n\hat{\sigma}^2/(n-p-q)\right] + n\left(1+\ln\sqrt{2\pi}\right)$$

$$+(p+q)\ln\left[\left(\sum_{t=1}^{n} X_t^2 - n\hat{\sigma}^2\right)/(p+q)\right], \tag{5.5.5}$$

where $\hat{\sigma}^2$ is the maximum likelihood estimate of the white noise variance.

The BIC is a consistent order-selection criterion in the sense that if the data $\{X_1, \ldots, X_n\}$ *are* in fact observations of an ARMA(p, q) process, and if $\hat{p}$ and $\hat{q}$ are the estimated orders found by minimizing the BIC, then $\hat{p} \to p$ and $\hat{q} \to q$ with probability 1 as $n \to \infty$ (Hannan 1980). This property is not shared by the AICC or AIC. On the other hand, order selection by minimization of the AICC, AIC, or FPE is asymptotically efficient for autoregressive processes, while order selection by BIC minimization is not (Shibata 1980; Hurvich and Tsai 1989). Efficiency is a desirable property defined in terms of the one-step mean square prediction error achieved by the fitted model. For more details see Brockwell and Davis (1991), Section 10.3.

In the modeling of real data there is rarely such a thing as the "true order." For the process $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ there may be many polynomials $\theta(z), \phi(z)$ such that the coefficients of z^j in $\theta(z)/\phi(z)$ closely approximate ψ_j for moderately small values of j. Correspondingly, there may be many ARMA processes with properties similar to $\{X_t\}$. This problem of identifiability becomes much more serious for multivariate processes. The AICC criterion does, however, provide us with a rational criterion for choosing among competing models. It has been suggested (Duong 1984) that models with AIC values within c of the minimum value should be considered competitive (with $c = 2$ as a typical value). Selection from among the competitive models can then be based on such factors as whiteness of the residuals (Section 5.3) and model simplicity.

We frequently need, particularly in analyzing seasonal data, to fit ARMA(p, q) models in which all except $m(\leq p + q)$ of the coefficients are constrained to be zero. In such cases the definition (5.5.4) is replaced by

$$\text{AICC}(\boldsymbol{\beta}) := -2 \ln L_X(\boldsymbol{\beta}, S_X(\boldsymbol{\beta})/n) + 2(m+1)n/(n-m-2). \tag{5.5.6}$$

Example 5.5.2 Models for the Lake Data

In Example 5.2.4 we found that the minimum-AICC ARMA(p, q) model for the mean-corrected lake data is the ARMA(1,1) model (5.2.14). For this model ITSM gives the values AICC = 212.77 and BIC = 216.86. A systematic check on ARMA(p, q) models for other values of p and q shows that the model (5.2.14) also minimizes the BIC statistic. The minimum-AICC AR(p) model is found to be the AR(2) model satisfying

$$X_t - 1.0441X_{t-1} + 0.2503X_{t-2} = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 0.4789),$$

with AICC = 213.54 and BIC = 217.63. Both the AR(2) and ARMA(1,1) models pass the diagnostic checks of Section 5.3, and in view of the small difference between the AICC values there is no strong reason to prefer one model or the other.

□

Problems

5.1 The sunspot numbers $\{X_t, t = 1, \ldots, 100\}$, filed as SUNSPOTS.TSM, have sample autocovariances $\hat{\gamma}(0) = 1382.2$, $\hat{\gamma}(1) = 1114.4$, $\hat{\gamma}(2) = 591.73$, and $\hat{\gamma}(3) = 96.216$. Use these values to find the Yule–Walker estimates of ϕ_1, ϕ_2, and σ^2 in the model

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + Z_t, \quad \{Z_t\} \sim \text{WN}\big(0, \sigma^2\big),$$

for the mean-corrected series $Y_t = X_t - 46.93, t = 1, \ldots, 100$. Assuming that the data really are a realization of an AR(2) process, find 95 % confidence intervals for ϕ_1 and ϕ_2.

5.2 From the information given in the previous problem, use the Durbin–Levinson algorithm to compute the sample partial autocorrelations $\hat{\phi}_{11}, \hat{\phi}_{22}$, and $\hat{\phi}_{33}$ of the sunspot series. Is the value of $\hat{\phi}_{33}$ compatible with the hypothesis that the data are generated by an AR(2) process? (Use significance level 0.05.)

5.3 Consider the AR(2) process $\{X_t\}$ satisfying

$$X_t - \phi X_{t-1} - \phi^2 X_{t-2} = Z_t, \quad \{Z_t\} \sim \text{WN}\big(0, \sigma^2\big).$$

a. For what values of ϕ is this a causal process?

b. The following sample moments were computed after observing $X_1, \ldots, X_{200}$:

$$\hat{\gamma}(0) = 6.06, \quad \hat{\rho}(1) = 0.687.$$

Find estimates of ϕ and σ^2 by solving the Yule–Walker equations. (If you find more than one solution, choose the one that is causal.)

5.4 Two hundred observations of a time series, $X_1, \ldots, X_{200}$, gave the following sample statistics:

sample mean:	$\bar{x}_{200} = 3.82;$
sample variance:	$\hat{\gamma}(0) = 1.15;$
sample ACF:	$\hat{\rho}(1) = 0.427;$
	$\hat{\rho}(2) = 0.475;$
	$\hat{\rho}(3) = 0.169.$

a. Based on these sample statistics, is it reasonable to suppose that $\{X_t - \mu\}$ is white noise?

b. Assuming that $\{X_t - \mu\}$ can be modeled as the AR(2) process

$$X_t - \mu - \phi_1(X_{t-1} - \mu) - \phi_2(X_{t-2} - \mu) = Z_t,$$

where $\{Z_t\} \sim \text{IID}\big(0, \sigma^2\big)$, find estimates of μ, ϕ_1, ϕ_2, and σ^2.

c. Would you conclude that $\mu = 0$?

d. Construct 95 % confidence intervals for ϕ_1 and ϕ_2.

e. Assuming that the data were generated from an AR(2) model, derive estimates of the PACF for all lags $h \geq 1$.

5.5 Use the program ITSM to simulate and file 20 realizations of length 200 of the Gaussian MA(1) process

$$X_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}(0, 1),$$

with $\theta = 0.6$.

a. For each series find the moment estimate of θ as defined in Example 5.1.2.

b. For each series use the innovations algorithm in the ITSM option `Model>Estimation>Preliminary` to find an estimate of θ. (Use the default value of the parameter m.) As soon as you have found this preliminary estimate for a particular series, select `Model>Estimation>Max likelihood` to find the maximum likelihood estimate of θ for the series.

c. Compute the sample means and sample variances of your three sets of estimates.

d. Use the asymptotic formulae given at the end of Section 5.1.1 (with $n = 200$) to compute the variances of the moment, innovation, and maximum likelihood estimators ofθ Compare with the corresponding sample variances found in (c).

e. What do the results of (c) suggest concerning the relative merits of the three estimators?

5.6 Establish the recursions (5.1.19) and (5.1.20) for the forward and backward prediction errors $u_i(t)$ and $v_i(t)$ in Burg's algorithm.

5.7 Derive the recursions for the Burg estimates $\phi_{ii}^{(B)}$ and $\sigma_i^{(B)2}$.

5.8 From the innovation form of the likelihood (5.2.9) derive the equations (5.2.10), (5.2.11), and (5.2.12) for the maximum likelihood estimators of the parameters of an ARMA process.

5.9 Use equation (5.2.9) to show that for $n > p$, the likelihood of the observations $\{X_1, \ldots, X_n\}$ of the causal AR(p) process defined by

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right),$$

is

$$L\left(\boldsymbol{\phi}, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \left(\det G_p\right)^{-1/2} \times \exp\left\{-\frac{1}{2\sigma^2}\left[\mathbf{X}_p' G_p^{-1} \mathbf{X}_p + \sum_{t=p+1}^{n} \left(X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}\right)^2\right]\right\},$$

where $\mathbf{X}_p = (X_1, \ldots, X_p)'$ and $G_p = \sigma^{-2}\Gamma_p = \sigma^{-2}E(\mathbf{X}_p\mathbf{X}_p')$.

5.10 Use the result of Problem 5.9 to derive a pair of linear equations for the least squares estimates of ϕ_1 and ϕ_2 for a causal AR(2) process (with mean zero). Compare your equations with those for the Yule–Walker estimates. (Assume that the mean is known to be zero in writing down the latter equations, so that the sample autocovariances are $\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h} X_{t+h}X_t$ for $h \geq 0$.)

5.11 Given two observations x_1 and x_2 from the causal AR(1) process satisfying

$$X_t = \phi X_{t-1} + Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right),$$

and assuming that $|x_1| \neq |x_2|$, find the maximum likelihood estimates of ϕ and σ^2.

5.12 Derive a cubic equation for the maximum likelihood estimate of the coefficient ϕ of a causal AR(1) process based on the observations $X_1, \ldots, X_n$.

5.13 Use the result of Problem A.7 and the approximate large-sample normal distribution of the maximum likelihood estimator $\hat{\boldsymbol{\phi}}_p$ to establish the approximation (5.5.1).