Time Series Analysis

Professor Abolfazl Safikhani
School of Social Work
Columbia University
Notes by Yiqiao Yin in LaTeX

May 7, 2017

Abstract

This is the notes for STATS GR 5221 Time Series Analysis at Columbia University. Course topics include but not limit to least squares smoothing and prediction, linear systems, Fourier analysis, and spectral estimation, impulse response and transfer function, fourier series, the fast Fourier transform, autocorrelation function, and spectral density, univariate Box-Jenkins modeling and forecasting.

This document is dedicated to Professor Abolfazl Safikhani.

Contents

1 Introduction

Go back to Table of Contents. Please click TOC

This chapter we introduce some basic ideas of time series analysis and stoachastic processes. Of particular importance are the concepts of stationarity and the autocovariance and sample autocovariance functions.

1.1 Examples of Time Series

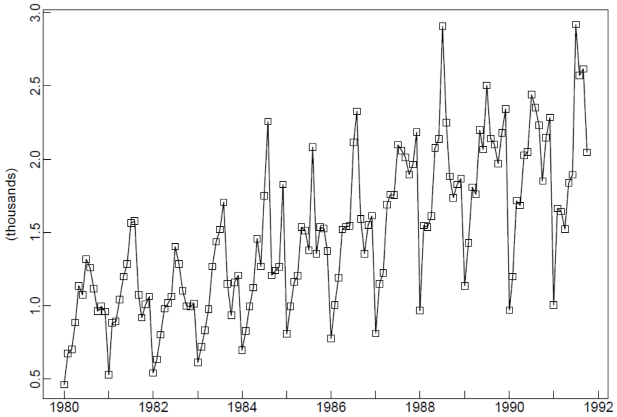
Go back to Table of Contents. Please click TOC

A **time series** is a set of observations x_t, each one being recorded at a specific time t. A *discrete-time time series* (the type to which this book is primarily devoted) is one in which the set T_0 of times at which observations are made is a discrete set, as is the case, for example, when observations are made at fixed time intervals. *Continuous-time time series* are obtained when observations are recorded continuously over some time interval, e.g., when $T_0 = [0, 1]$.

Example 1.1. Figure 1 shows the monthly sales (in kiloliters) of red wine by Australian winemakers from January 1980 through October 1991. In this case the set T_0 consists of the 142 times {(Jan. 1980), (Feb. 1980), ..., (Oct. 1991)}. In the present example this amounts to measuring time in months with (Jan. 1980) as month 1. Then T_0 is the set {1, 2, ..., 142}. it appears from the graph that the sales have an upward trend and a seasonal pattern with a peak in July and a trough in January.

□

Figure 1: The Australian red wine sales, Jan. '80-Oct. '91.



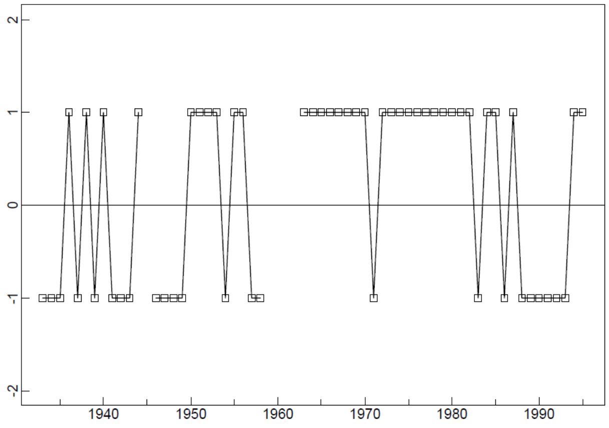
Example 1.2. Figure 2 shows the results of the all-star games by plotting x_t, where

$$x_t = \begin{cases} 1 & if\ the\ National\ League\ won\ in\ year\ t, \\ -1 & if\ the\ American\ League\ won\ in\ year\ t. \end{cases}$$

This is a series with only two possible values, ±1. It also has some missing values, since no game was played in 1945, and two games were scheduled for each of the years 1959-1962.

□

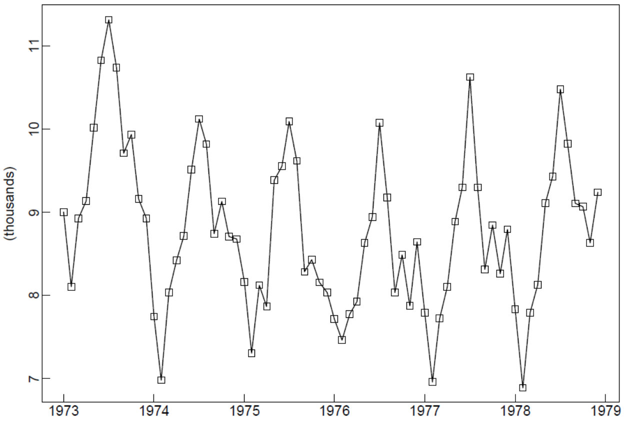
Figure 2: Results of the all-star baseball games, 1933-1995.



Example 1.3. The monthly accidental death figures show a strong seasonal pattern, with the maximum for each year occuring in July and minimum for each year occurring in February. The presence of a trend in the figure below is less apparent than in the wine sales.

□

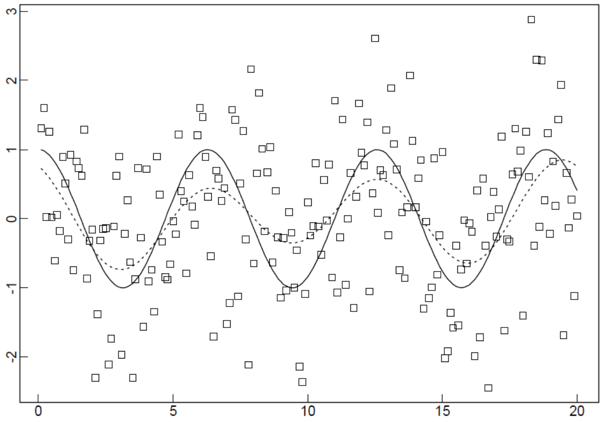
Figure 3: The monthly accidental deaths data, 1973-1978.



Example 1.4. Figure 4 shows simulated values of the series $X_t = \cos(\frac{t}{10}) + N_t,\ t = 1, 2, ..., 300,$ where $\{N_t\}$ is a sequence of independent normal random variables, with mean 0 and variance 0.25. Such a series is often referred to as *signal plus noise*, the signal being the smooth function, $S_t = \cos(\frac{t}{10})$ in this case. Given only the data X_t, how can we determine the unknown signal component? There are many approaches to this general problem under varying assumptions about the signal and the noise. One simply approach is to *smooth* the data by expressing X_t as a sum of sine waves of various frequencies (see Section 4.2) and eliminating the high-frequency components. If we do this to the values of $\{X_t\}$ shown in Figure 4 and retain only the lowest 3.5% of the frequency components, we obtain the estimate of the signal also shown in Figure 1.4. The waveform of the signal is quite close to that of the true signal in this case, although its amplitude is somewhat smaller.

□

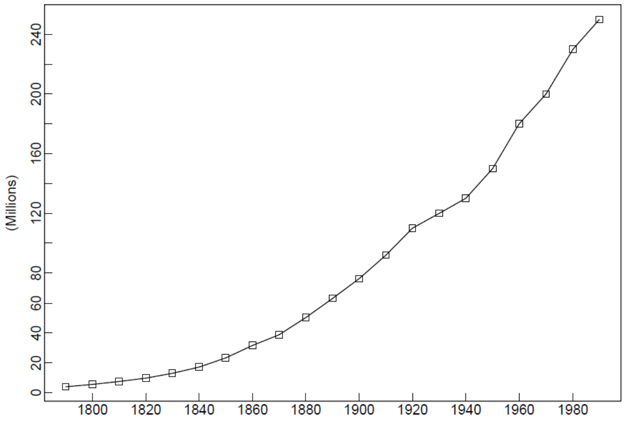
Figure 4: The series $\{X_t\}$ of Example 1.4.



Example 1.5. The population of the U.S.A., measured at ten-year intervals, is shown in Figure 5. The graph suggests the possibility of fitting a quadratic or exponential trend to the data. We shall explore in Section 1.3.

□

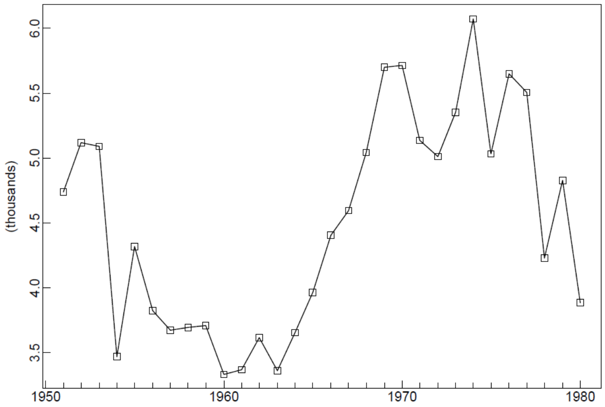
Figure 5: Population of the U.S.A. at ten-year intervals, 1790-1990.



Example 1.6. The annual numbers of strikes in the U.S.A. for the years 1951-1980 are shown in Figure 6. They appear to fluctuate erratically about a slowly changing level.

□

Figure 6: Strikes in the U.S.A., 1951-1980.



1.2 Objectives of Time Series Analysis

Go back to Table of Contents. Please click TOC

The examples considered in Section 1.1 are an extremely small sample from the multitude of time series encountered in the fields of engineering, science, socialogy, and economics. The purpose is to study techniques for drawing inferences from such series. Before we do this, however, it is necessary to set up a hypothetical probability model to represent the data.

After an appropriate family of models has been chosen, it is then possible to estimate parameters, check for goodness of fit to the data, and possibly to use the fitted model to enhance our understanding of the mechanism generating the series. Once a satisfactory model has been developed, it may be used in a variety of ways depending on the particular field of application.

1.3 Some Simple Time Series Models

Go back to Table of Contents. Please click TOC

Definition 1.7. A **time series model** for the observed data $\{x_t\}$ is specification of the joint distributions (or possibly only the means and covariances) of a sequence of random variables $\{X_t\}$ of which $\{x_t\}$ is postulated to be a realization.

Remark 1.8. We shall frequently use the term *time series* to mean both the data and the process of which it is a realization.

□

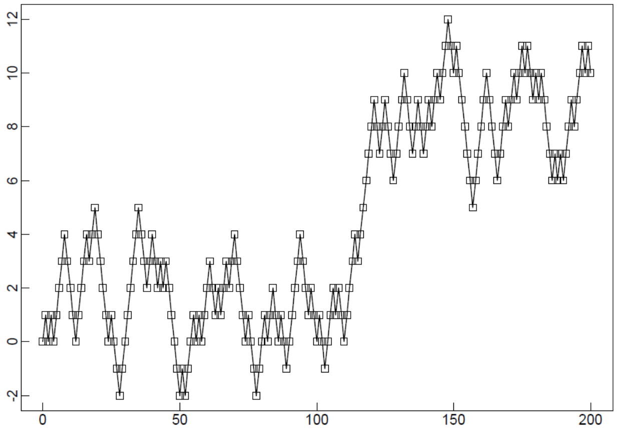
A complete probabilistic time series model for the sequence of random variables $\{X_1, X_2, ...\}$ would specify all of the **joint distributions** of the random vectors $(X_1, ..., X_n)'$, $n = 1, 2, ...$, or equivalently all of the probabilities

$$P[X_1 \leq x_1, ..., X_n \leq x_n],\ -\infty < x_1, ..., x_n < \infty,\ n = 1, 2, ...$$

We specify only the **first-** and **second-order moments** of the joint distributions, i.e. the expected values $\mathbb{E}(X_t)$ and the expected products $\mathbb{E}(X_{t+h}X_t)$, $t = 1, 2, ..., h = 0, 1, 2, ...$, focusing on properties of the sequence $\{X_t\}$ that depend only on these. Such properties of $\{X_t\}$ are referred to as **second-order properties**.

Figure 7 shows one of many possible realizations of $\{S_t, t = 1, ..., 200\}$, where $\{S_t\}$ is a sequence of random variables. In most practical problems involving time series we see only one realization.

Figure 7: One realization of a simple random walk $\{S_t,\, t = 0, 1, 2, ..., 200\}$.



1.3.1 Some Zero-Mean Models

Go back to Table of Contents. Please click TOC

Example 1.9. The simplest model for a time series is one in which there is no trend or seasonal component and in which the observations are simply independent and identically distributed (iid) random variables with zero mean. We refer to sucha sequence of random variables $X_1, X_2, ...$ as iid noise. We can write, $\forall n \in \mathbb{Z}$ and $x_1, ..., x_n \in \mathbb{R}$,

$$P[X_1 \leq x_1, ..., X_n \leq x_n] = P[X_1 \leq x_1] \dots P[X_n \leq x_n] = F(x_1) \dots F(x_n),$$

where $F(\cdot)$ is the cumulative distribution function of each of the identicallty distributed random variables $X_1, X_2, ...$ In this model, there is no dependence between observations. In particular, for all $h \geq 1$ and all $x, x_1, ..., x_n$,

$$P[X_{n+h} \leq x | X_1 = x_1, ..., X_n = x_n] = P[X_{n+h} \leq x],$$

showing that knowledge of $X_1, ..., X_n$ is of no value for predicting the behavior of X_{n+h}. Given the values of $X_1, ..., X_n$, the function f that minimizes the mean squared error $\mathbb{E}[(X_{n+h} - f(X_1, ..., X_n))^2]$ is in fact identically zero. Although this means that iid noise is a rather uninteresting process for forecasters, it plays an important role as a building block.

□

Example 1.10. Consider the sequence of iid random variables $\{X_t, t = 1, 2, ..., \}$ with

$$P[X_t = 1] = p,\ P[X_t = -1] = 1 - p,$$

where $p = \frac{1}{2}$. The time series obtained by tossing a penny repeatedly and scoring +1 for each head and -1 for each tail is usually modeled as a realization of this process. A priori we might well consider the same process as a model for baseball games in previous example.

□

Example 1.11. The random walk $\{S_t, t = 0, 1, 2, ...\}$ (starting at zero) is obtained by cumulatively summing (or "integrating") iid random variables. Thus a random walk with zero mean is obtained by defining $S_0 = 0$ and

$$S_t = X_1 + X_2 + \cdots + X_t, \; for \; t = 1, 2, ...,$$

where $\{X_t\}$ is iid noise. If $\{X_t\}$ is a binary process (just like the one aboe), then $\{S_t, t = 0, 1, 2, ...,\}$ is called a **simple symmetric random walk**. This walk can be viewed as the location of a pedestrian who starts at position zero at time zero and at each integer time tosses a fair coin, stepping one unit to the right each time a head appears and one unit to the left for each tail. A realization of length 200 of a simple symmetric random is shown in Figure 3. Notice that the outcomes of the coin tosses can be recovered from $\{S_t, t = 0, 1, ...\}$ by differencing. Thus the result of the tth toss can be found from $S_t - S_{t-1} = X_t$.

□

1.3.2 Models with Trend and Seasonality

Go back to Table of Contents. Please click TOC

In examples of Section 1.1 there is a clear trend in the data. An increasing trend is apparent in both the Australian red wine sales (Figure 1) and the population of the U.S.A. (Figure 5). In both cases a zero-mean model for the data is clearly inappropriate. The graph of the population data, which contains no apparent periodic component, suggests trying a model of the form $X_t = m_t + Y_t$, where m_t is a slowly changing function known as the **trend component** and Y_t has zero mean. A useful technique for estimating m_t is the method of least squares.

In the least squares procedure we attempt to fit a parametric family of functions, e.g.,

$$m_t = a_0 + a_1 t + a_2 t^2,$$

to the data $\{x_1, ..., x_n\}$ by choosing the parameters, in this illustration a_0, a_1, and a_2, to minimize $\sum_{t=1}^{n} (x_t - m_t)^2$. This method of curve fitting is called **least squares regression**.

Many time series are influenced by seasonally varying factors such as the weather, the effect of which can be modeled by a periodic component with fixed known period. For example, the accidental deaths series (figure 3) shows a repeating annual pattern with peaks in July and troughs in February, strongly suggesting a seasonal factor with period 12. In order to represent such a seasonal effect, allowing for noise but assuming no trend, we can use the simple model, $X_t = s_t + Y_t$, where s_t is a periodic function of t with period $d(s_{t-d} = s_t)$. A convenient choice for s_t is a sum of harmonics (or sine waves) given by

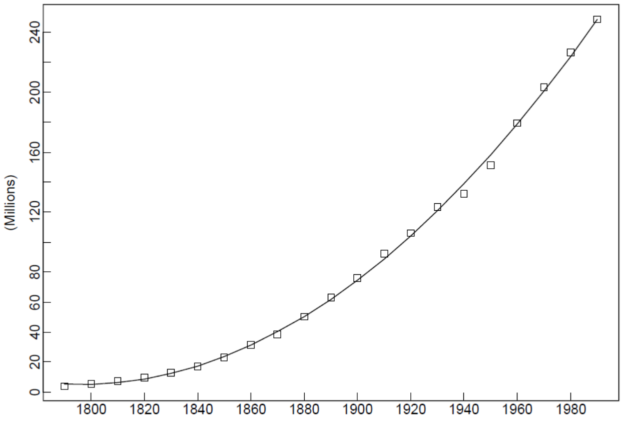
$$s_t = a_0 + \sum_{j=1}^{k} (a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t)),$$

where a_0, a_1, ..., a_k and $b_1, ..., b_k$ are unknown parameters and $\lambda_1, ..., \lambda_k$ are fixed frequencies, each being some integer multiple of $2\pi/d$. For a sine wave with period d, set $f_1 = n/d$, where n is the number of observations from beginning of the series to make it so.) The other $k-1$ Fourier indices should be positive integer multiples of the first, corresponding to harmonics of the fundamental sine wave with period d. Thus to fit a single since wave with period 365 to 365 daily observations we would choose $k = 1$ and $f_1 = 1$. To fit a linear combination of sine waves with periods $365/j$, $j = 1, ..., 4$, we would choose $k = 4$ and $f_j = j$, $j = 1, ..., 4$. Once k and $f_1, ..., f_k$ have been specified, we run least squares regression to obtain the required regression coefficients.

Example 1.12. A graph of the level in feet of Lake Huron (reduced by 570) in the years 1875-1972 is displayed in Figure 9. Since the lake level appears to decline at a roughly linear rate. A form of model can be

$$X_t = a_0 + a_1 t + Y_t,\ t = 1, ..., 98$$

Figure 8: One realization of a simple random walk $\{S_t,\ t = 0, 1, 2, ..., 200\}$.



The least squares estimates of the parameter values are

$$\hat{a}_0 = 10.202 \ and \ \hat{a}_1 = -0.0242.$$

(The resulting least squares line, $\hat{a}_0 + \hat{a}_1 t$, is also displayed in Figure 9.) The estimates of the noise, Y_t, are the residuals obtained by subtracting the least squares line from x_t and are plotted in Figure 10. There are two interesting features of the graph of the residuals. The first is the absence of any discernible trend. The second is the smoothness of the graph. Smoothness of the graph of a time series is generally indicative of the existence of some form of dependence among the observations.

Figure 9: One realization of a simple random walk $\{S_t,\ t = 0, 1, 2, ..., 200\}$.

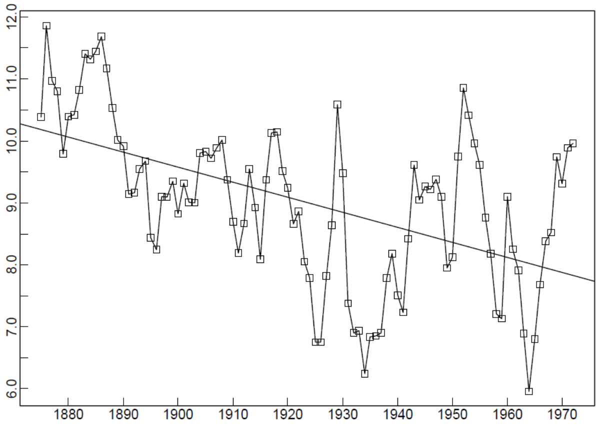
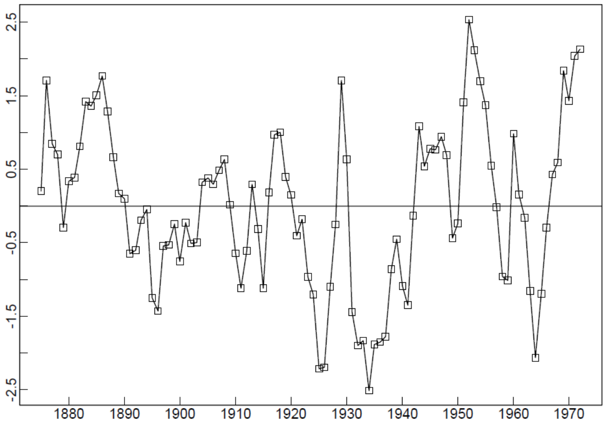


Figure 10: One realization of a simple random walk $\{S_t,\ t = 0, 1, 2, ..., 200\}$.



Such dependence can be used to advantage in forecasting future values of the series. If we were to assume the validity of the fitted model with iid residuals $\{Y_t\}$, then the minimum mean squared error predictor of the next residual (Y_{99}) would be zero. However, Figure 10 strongly suggests that Y_{99} would be positive.

□

1.3.3 A General Approach to Time Series Modeling

Go back to Table of Contents. Please click TOC

We have seen, from above, general approaches to time series analysis that will form the basis for much of what is done in this document. Here we outline the approach to provide the reader with an overview of the way in which the various ideas of this chapter fit together.

- Plot the series and examine the main features of the graph, checking in particular whether there is
  (a) a trend,
  (b) a seasonal component,
  (c) any apparent sharp changes in behavior,
  (d) any outlying observations.
- Remove the trend and seasonal components to get *stationary* residuals (as defined in section 1.4). To do this, it may sometimes be necessary to apply a preliminary transformation to the data. For example, if the magnitude of the fluctuations appears to grow roughly linearly with the level of the series, then the transformed series $\{\ln X_1, ..., \ln X_n\}$ will have fluctuations of more constant magnitude.
- Choose a model to fit the residuals, making use of various sample statistics including the sample autocorrelation function to be defined in section 1.4.
- Forecasting will be achieved by forecasting the residuals and then inverting the transformations described above to arrive at forecasts of the original series $\{X_t\}$.
- An extremely useful alternative approach touched on only briefly in this book is to express the series in terms of its Fourier components, which are sinusoidal waves of different frequencies.

1.4 Stationary Models and the Autocorrelation Function

Go back to Table of Contents. Please click TOC

Definition 1.13. Let $\{X_t\}$ be a time series with $\mathbb{E}(X_t^2) < \infty$. The **mean function** of $\{X_t\}$ is

$$\mu_X(t) = \mathbb{E}(X_t).$$

The **covariance function** of $\{X_t\}$ is

$$\gamma_X(r, s) = Cov(X_r, X_s) = \mathbb{E}[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

for all integers r and s.

Definition 1.14. $\{X_t\}$ is **(weakly) stationary** if
(i) $\mu_X(t)$ is independent of t, and
(ii) $\gamma_X(t + h, t)$ is independent of t for each h.

Remark 1.15. Strict stationarity of a time series $\{X_t, t = 0, \pm 1, ...\}$ is defined by the condition that $X_1, ..., X_n)$ and $(X_{1+h}, ..., X_{n+h})$ have the same joint distributions for all integers h and $n > 0$. It is easy to check that if $\{X_t\}$ is strictly stationary and $\mathbb{E}(X_t^2) < \infty$ for all t, then $\{X_t\}$ is also weakly stationary. Whenever we use the term *stationary* we shall mean weakly stationary as in Definition 1.14, unless we specifically indicate otherwise.

□

Remark 1.16. In view of condition (ii), whenever we use the term covariance function with reference to a *stationary* time series $\{X_t\}$ we shall mean the function γ_X of *one* variable, defined by

$$\gamma_X(h) := \gamma_X(h,0) = \gamma_X(t+h,t).$$

The function $\gamma_X(\cdot)$ will be referred to as the autocovariance function and $\gamma_X(h)$ as its value at *lag h*.

□

Definition 1.17. Let $\{X_t\}$ be a stationary time series. The **autocovariance function** (ACVF) of $\{X_t\}$ at lag h is

$$\gamma_X(h) = Cov(X_{t+h}, X_t).$$

The **autocorrelation function** (ACF) of $\{X_t\}$ at lag h is

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = Cor(X_{t+h}, X_t).$$

In the folloiwng examples we shall frequently use the easily verified **linearity property of covariances**, that if $\mathbb{E}(X^2) < \infty$, $\mathbb{E}(Y^2) < \infty$, $\mathbb{E}(Z^2) < \infty$ and a, b, and c are any real constants, then

$$Cov(aX + bY + c, Z) = aCov(X,Z) + bCov(Y,Z).$$

Example 1.18. If $\{X_t\}$ is iid noise and $\mathbb{E}(X_t^2) = \sigma^2 < \infty$, then the first requirement of Definition 1.14 is obviously satisfied, since $\mathbb{E}(X_t) = 0$ for all t. By the assumed independence,

$$\gamma_X(t+h,t) = \begin{cases} \sigma^2, & if\ h = 0, \\ 0, & if\ h \neq 0, \end{cases}$$

which does not depend on t. Hence iid noise with finite second moment is stationary. We shall use the notation $\{X_t\} \sim IID(0,\sigma^2)$ to indicate that the random variables X_t are independent and identically distributed random variables, each with mean 0 and variance σ^2.

□

Example 1.19. If $\{X_t\}$ is a sequence of uncorrelated random variables, each with zero mean and variance σ^2, then clearly $\{X_t\}$ is stationary with the same covariance function as the iid noise in Example 1.17. Such a sequence is referred to as **white noise** (with mean 0 and variance σ^2). This is indicated by the notation $\{X_t\} \sim WN(0,\sigma^2)$. Clearly, every IID(0,$\sigma^2$) sequence is WN(0,$\sigma^2$) but not conversely.

□

Example 1.20. If $\{S_t\}$ is the random walk defined in Example 1.11 with $\{X_t\}$ as in Example 1.18, then $\mathbb{E}(X_t) = 0$, $\mathbb{E}(S_t^2) = t\sigma^2 < \infty$ for all t, and, for $h \geq 0$,

$$\begin{aligned} \gamma_S(t_h,t) &= Cov(S_{t+h}, S_t) \\ &= Cov(S_t + X_{t+1} + \cdots + X_{t+h}, S_t) \\ &= Cov(S_t, S_t) \\ &= t\sigma^2. \end{aligned}$$

Since $\gamma_S(t+h,t)$ depends on t, the series $\{S_t\}$ is *not* stationary.

□

Example 1.21 (*Moving Average or MA(1)*). . Consider the series defined by the equation

$$X_t = Z_t + \theta Z_{t-1},\ t = 0, \pm 1, ...,$$

where $\{Z_t\} \sim WN(0, \sigma^2)$ and θ is a real-valued constant. From the equation above, we see that $\mathbb{E}(X_t) = 0$, $\mathbb{E}(X_t^2) = \sigma^2(1 + \theta^2) < \infty$, and

$$\gamma_X(t + h, t) = \begin{cases} \sigma^2(1 + \theta^2), & if\ h = 0, \\ \sigma^2\theta, & if\ h = \pm 1, \\ 0, & if\ |h| > 1. \end{cases}$$

Thus the requirements of Definition 1.14 are satisfied, and $\{X_t\}$ is stationary. The autocorrelation function of $\{X_t\}$ is

$$\rho_X(h) = \begin{cases} 1, & if\ h = 0, \\ \theta/(1 + \theta^2), & if\ h = \pm 1, \\ 0, & if\ |h| > 1. \end{cases}$$

□

Remark 1.22. For the above example, notice that $\mathbb{E}(X_t) = 0$, thus we have

$$\begin{aligned} Var(X_+t) &= Var(Z_t + \theta Z_{t-1} \\ &= \sigma^2 + \theta^2 Z_{t-1}^2 \\ &= \sigma^2 + \theta^2\sigma^2 \\ &= (1 + \theta^2)\sigma^2 \end{aligned}$$

Moreover, we have

$$\begin{aligned} Cov(X_{t+1}, X_t) &= Cov(Z_{t+1} + \theta Z_t, Z_t + \theta Z_{t-1}) \\ &= 0 + 0 + \theta\sigma^2 + 0 \\ &= \theta\sigma^2 \end{aligned}$$

Thus, we can calculate ACF, i.e.

$$\rho(h) \begin{cases} 1; & h = 0 \\ \frac{\theta\sigma^2}{(1+\theta^2)\sigma^2}; & h = \pm 1 \\ 0; & |h| > 1 \end{cases}$$

$\Rightarrow$

$$\rho_X(h) = \begin{cases} 1, & if\ h = 0, \\ \theta/(1 + \theta^2), & if\ h = \pm 1, \\ 0, & if\ |h| > 1. \end{cases}$$

□

Example 1.23 (*Autoregression or AR(1).*). Assume now that $\{X_t\}$ is a stationary series satisfying the equations

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

where $\{Z_t\} \sim WN(0, \sigma^2)$, $|\phi| < 1$, and Z_t is uncorrelated with X_s for each $s < t$. (We show in Section 2.2 that there is exactly one such solution.) By taking expectations on each side of equation above and using the fact that $\mathbb{E}(Z_t) = 0$, we see that $\mathbb{E}(X_t) = 0$.

To find the autocorrelation function of $\{X_t\}$ we multiply each side by X_{t-h} $(h > 0)$ and then take expectations to get

$$\begin{aligned}\gamma_X(h) &= Cov(X_t, X_{t-h}) \\ &= Cov(\phi X_{t-1}, X_{t-h}) + Cov(Z_t, X_{t-h}) \\ &= \phi\gamma_X(h-1) + 0 = \cdots = \phi^h\gamma_X(0).\end{aligned}$$

Observing that $\gamma(h) = \gamma(-h)$ and using Definition 1.16, we find that

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \phi^{|h|}, h = 0, \pm 1, ...$$

It follows from the linearity of the covariance function in each of its arguments and the fact that Z_t is uncorrelated with X_{t-1} that

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

and hence that $\gamma_X(0) = \sigma^2/(1-\phi^2)$.

□

Remark 1.24. For the above example, we consider X_t to be a combination of X_{t-1} and Z_t. Then we apply the same method for X_{t-1}. In doing so, we have the following,

$$\begin{aligned}X_t &= \phi X_{t-1} + Z_t \\ &= \phi(\phi X_{t-2} + Z_{t-1}) + Z_t \\ &= \phi^2 X_{t-2} + \phi Z_{t-1} + Z_t \\ &= \phi^2(\phi X_{t-2} + Z_{t-2}) + \phi Z_{t-1} + Z_t \\ &= \phi^3 X_{t-3} + \phi^2 Z_{t-2} + \phi Z_{t-1} + Z_t \\ &= \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j Z_{t-j}\end{aligned}$$

Thus, we can conclude

$$\lim_{n\to\infty} X_t = \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j Z_{t-j} < \infty$$

if $|\phi| < 1$. This is the reason why the assumption $|\phi| < 1$ is essential for theis argument to hold.

Taking a step further, we notice that $\gamma(h) = Cov(X_{t+h}, X_t) = \mathbb{E}\big((X_{t+h} - \mu)(X_t - \mu)\big)$ gives us estimates for ACVF, i.e., sample ACVF,

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_{t+h} - \bar{X})(X_t - \bar{X}).$$

Moreover, we have

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

which gives us estimates for ACF, i.e., sample ACF.

□

1.4.1 The Sample Autocorrelation Function

Go back to Table of Contents. Please click TOC

In practice, we start with *observed data* $\{x_1, x_2, ..., x_n\}$. To assess the degree of dependence in the data and to select a model for the data that reflects this, one of the important tools we use is the **sample autocorrelation function** (sample ACF) of the data. If we beleive that the data are realized values of a stationary time series $\{X_t\}$, then the sample ACF will provide us with an estimate of the ACF of $\{X_t\}$. This estimate may suggest which of the many possible stationary time series models is a suitable candidate.

Definition 1.25. Let $x_1, ..., x_n$ be observations of a time series. The **sample mean** of $x_1, ..., x_n$ is

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t.$$

The **sample autocovariance function** is

$$\hat{\gamma}(h) := n^{-1} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}), \ -n < h < n.$$

The **sample autocorrelation function** is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \ -n < h < n.$$

Example 1.26. Figure 11 shows 200 simulated values of normall distributed iid (0,1), denoted by IID N(0,1), noise. Figure 12 shows the corresponding sample autocorrelation function at lags 0, 1, ..., 40. Since $\rho(h) = 0$ for $h > 0$, one would also expect the corresponding sample autocorrelations to be near 0. It can be shown, in fact, that for iid noise with finite variance, the sample auto correlations $\hat{\rho}(h)$, $h > 0$, are approximately IID N(0,1/n) for n large. Hence, approximately 95% of the sample autocorrelations should fall between the bounds $\pm 1.96/\sqrt{n}$ (since 1.96 is the 0.975 quantile of the standard normal distribution).

□

Figure 11: 200 simulated values of iid N(0,1) noise.

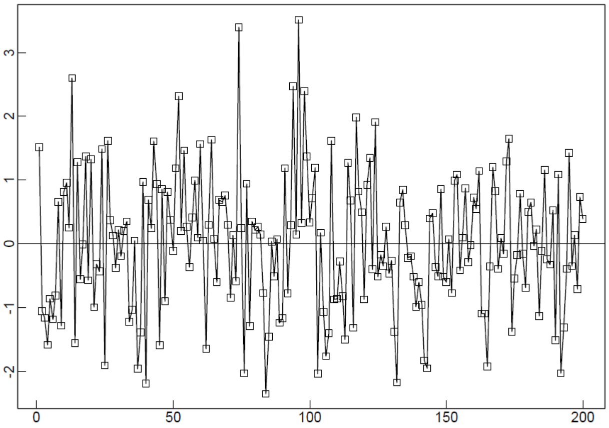
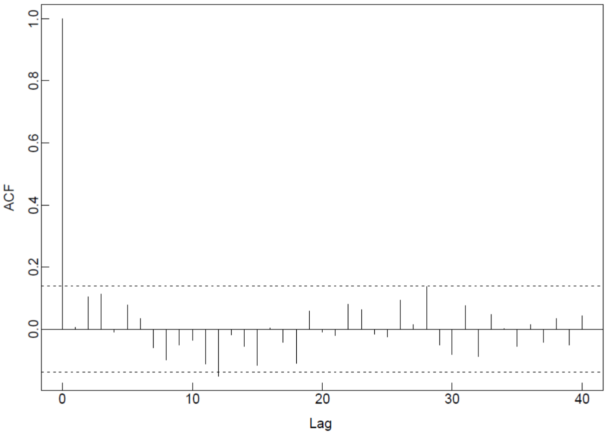


Figure 12: The sample autocorrelation function for the data of the figure above showing the bounds $\pm 1.96/\sqrt{n}$.



Remark 1.27. Note that $\hat{\gamma}(h)$ is approximately the sample covariance function of $(x_1, x_{1+h}), ..., (X_{n-h}, X_n)$. The covariance matrix $\hat{\Gamma}_n = [\hat{\gamma}(i-j)]$, for $i, j = 1, ..., n$ is non-negative definite (positive definite). If data are observations from IID noise, then we have $\hat{\rho}(h) \approx N(0, 1/n)$ and are independent for all $h \geq 1$. For IID noise, $|\hat{\rho}(h)| < 1.96n^{-.5}$ with probability 0.95.

Let us consider the following matrix data set, as an example, the sample covariance matrix takes the form

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \dots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \dots & \gamma(n-2) \\ \gamma(2) & \gamma(1) & \gamma(0) & \dots & \gamma(n-3) \\ \vdots & \vdots & \ddots & & \vdots \\ \gamma(n-2) & \vdots & & \ddots & \gamma(1) \\ \gamma(n-1) & \dots & \dots & \dots & \gamma(0) \end{bmatrix}$$

while i, jth -matrix $= Cov(X_i, X_j) = \gamma(i-j)$. In this case, $X_1, ..., X_n \sim WN(0, \sigma^2) = \Gamma = \sigma^2 I_n$. Then $A_{n \times n}$ is a positive definite if

$$0 < b'Ab = (b_1, ..., b_n) \cdot \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & & & \\ \vdots & & \ddots & \\ a_{n1} & \dots & & a_{nn} \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

In this case, we have

$$\begin{aligned} 0 < b'Ab &= Var(b' \begin{pmatrix} X_1 \\ X_2 \\ \vdots X_n \end{pmatrix}) \\ &= b'(Cov(\mathbf{X})b \\ &= b'\Gamma b \end{aligned}$$

□

1.5 Estimation and Elimination of Trend and Seasonal Components

Go back to Table of Contents. Please click TOC

The first step is to plot the data. If there are any apparent discontinuities in the series, it may be advisable to analyze the series by first breaking it into homogeneous segments. If there are outlying observations, they should be studied carefully. inspection of a graph may also suggest the possibility of representing the data as a realization of the process (the **classical decomposition** model)

$$X_t = m_t + s_t + Y_t,$$

where m_t is a slowly changing function known as a **trend component**, s_t is a function with known period d referrred to as a **seasonal component**, and Y_t is a **random noise component** that is stationary in the sense of Definition 1.3. If the seasonal and noise fluctuations appear to increase with the level of the process, then a preliminary transformation of the data is often used.

Our aim is to estimate and extract the deterministic components m_t and s_t in the hope that the residual or noise component Y_t will turn out to be a stationary time series. We can use the theory of such processes to find a satisfactory probabilistic model for the process Y_t, to analyze its properties, and to use it in conjunction with m_t and s_t for purposes of prediction and simulation of $\{X_t\}$.

Another approach, developed by Box and Jenkins (1976) [6], is to apply differencing operators repeatedly to the series $\{X_t\}$ until the differenced observations resemble a realization of some stationary time series $\{W_t\}$. We can then use the theory of stationary processes for the modeling, analysis, and prediction of $\{W_t\}$ and hence of the original process.

1.5.1 A General Approach to Time Series Modeling

Go back to Table of Contents. Please click TOC

Definition 1.28. Nonseasonal Model with Trend:

$$X_t = m_t + Y_t,\ t = 1, ..., n,$$

where $\mathbb{E}(Y_t) = 0$.

(If $\mathbb{E}(Y_t) \neq 0$, then we can replace m_t and y_t in above equation with $m_t + \mathbb{E}(Y_t)$ and $Y_t - \mathbb{E}(Y_t)$, respectively.)

Method 1: Trend Estimation

Moving average and spectral smoothing are essentially nonparametric methods for trend (or signal) estimation and not for model building. Special smoothing fitlers can also be designed to remove periodic components as described under Method S1 below. The choice of smoothing fitler requires a certain amoount of subjective judgment, and it is recommended that a variety of filters be tried in order to get a good idea of the underlying trend. Exponential smoothing, since it is based on a moving average of *past* values only, is often used for forecasting, the smoothed value at the present time being used as the forecast of the next value.

To construct a *model* for the data (with no seasonality) there are two general approaches. One is to fit a polynomial trend (by least squares), then to subtract the fitted trend from the data and to find an appropriate stationary time series model for the residuals. The other is to eliminate the trend by differencing as described in Method 2 and then to find an appropriate stationary model for the differenced series. The latter method has the advantage that it usually requires the estimation of fewer parameters and does not rest on the assumption of a trend that remains fixed throughout the observation period.

(a) *Smoothing with a finite moving average filter.* Let q be a nonnegative integer and consider the two-sided moving average

$$W_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t-j}$$

of the process $\{X_t\}$ defined by Nonseasonal Model. Then for $q + 1 \leq t \leq n - q$,

$$W_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t-j} + (2q+1)^{-1} \sum_{j=-q}^{q} Y_{t-j} \approx m_t,$$

assuming that m_t is approximately linear over the interval $[t-q, t+q]$ and that the average of the error terms over this interval is close to zero.

The moving average thus provides us with the estimates

$$\hat{m}_t = (2q+1)^{-1} \sum_{j=-q}^{q}, \; q+1 \leq t \leq n-q.$$

Since X_t is not observed for $t \leq 0$ or $t > n$, we cannot use the above equation for $t \leq q$ or $t > n-q$.

It is useful to think of $\{\hat{m}_t\}$ in the above equation as a process obtained from $\{X_t\}$ by application of a linear operator or linear fitler $\hat{m}_t = \sum\limits_{j=-\infty}^{\infty} a_j X_{t-j}$ with weights $a_j = (2q+1)^{-1}$, $-q \leq j \leq q$. This particular fitler is a **low-pass** filter in the sense that it takes the data $\{X_t\}$ and removes from it the rapidly fluctuating (or high frequency) component $\{\hat{Y}_t\}$ to leave the slowly varying estimated trend term $\{\hat{m}_t\}$.

The particular filter is only one of many that could be used for smoothing. For large q, provided that $(2q+1)^{-1} \sum\limits_{j=-q}^{q} Y_{t-j} \approx 0$, it not only will attenuate noise but at the same time will allow linear trend functions $m_t = c_0 + c_1 t$ to pass without distortion. However, we must beware of choosing q to be too large, since if m_t is not linear, the filtered process, although smooth, will not be a good estimate of m_t. Be clever choise of the weights $\{a_j\}$ it is possible to design a fitler that will not only be effective in attenuating noise in the data, but that will also allow a larger class of trend functions to pass through without distortion. The Spencer 15-point moving average is a filter that passes polynomials of degree 3 without distortion. Its weights are

$$a_j = 0, \; |j| > 7,$$

with

$$a_j = a_{-j}, \; |j| \leq 7,$$

and

$$[a_0, a_1, ..., a_7] = \frac{1}{320}[74, 67, 46, 21, 3, -5, -6, -3].$$

Applied to the process with $m_t = c_0 + c_1 t + c_2 t^2 + c_3 t^3$, it gives

$$\sum_{j=-7}^{7} a_j X_{t-j} = \sum_{j=-7}^{7} a_j m_{t-j} + \sum_{j=-7}^{7} a_j Y_{t-j} \approx \sum_{j=-7}^{7} a_j m_{t-j} = m_t,$$

where the last step depends on the assumed form of m_t.

(b) *Exponential smoothing.* For any fixed $\alpha \in [0,1]$, the one-sided moving averages $\hat{m}_t$, $t = 1, ..., n$, defined by the recursions

$$\hat{m}_t = \alpha X_t + (1-\alpha)\hat{m}_{t-1}, \; t = 2, ..., n$$

and

$$\hat{m}_1 = X_1$$

can be computed by specifying the value of α.

(c) *Smoothing by elimination of high-frequency components.* We are allowed to smooth an arbitrary series by elimination of the high-frequency components of its Fourier series expansion.

(d) *Polynomial fitting.* We showed how a trend of the form $m_t = a_0 + a_1 t + a_2 t^2$ can be fitted to the data $\{x_1, ..., x_n\}$ by choosing the parameters a_0, a_1, and a_2 to minmize the sum of squares, $\sum_{t=1}^{n}(x_t - m_t)^2$. The method of least squares estimation can also be used to estimate higher-order polynomial trends in the same way.

Method 2: Trend Elimination by Differencing

Instead of attempting to remove the noise by smoothing as in Method 1, we now attempt to eliminate the trend term by differencing. We define the lag-1 difference operator $\triangledown$ by

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

where B is the backward shift operator,

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

Powers of the operators B and $\triangledown$ are defined in the obvious way, i.e., $B^j(X_t) = X_{t-j}$ and $\triangledown^j(X_t) = \triangledown(\triangledown^{j-1}(X_t))$, $j \geq 1$, with $\triangledown^0(X_t) = X_t$. Polynomials in B and $\triangledown$ are manipulated in precisely the same way as polynomial functions of real variables. For example,

$$\begin{aligned} \triangledown^2 X_t &= \triangledown(\triangledown(X_t)) = (1 - B)(1 - B)X_t = (1 - 2B + B^2)X_t \\ &= X_t - 2X_{t-1} + X_{t-2} \end{aligned}$$

If the operator $\triangledown$ is applied to a linear trend function $m_t = c_0 + c_1 t$, then we obtain the constant function $\triangledown m_t = m_t - m_{t-1} = c_0 + c_1 t - (c_0 + c_1(t-1)) = c_1$. In the same way any polynomial trend of degree k can be reduced to a constant by application of the operator $\triangledown^k$. For example, if $X_t = m_t + Y_t$, where $m_t = \sum_{j=0}^{k} c_j t^j$ and Y_t is stationary with mean zero, application of $\triangledown^k$ gives

$$\triangledown^k X_t = k! c_k + \triangledown^k Y_t,$$

a stationary process with mean $k! c_k$. These considerations suggest the possibility, given any sequence $\{x_t\}$ of data, of applying the operator $\triangledown$ repeatedly until we find a sequence $\{\triangledown^k x_t\}$ that can plausibly be modeled as a realization of a stationary process. It is often found in practice that the order k of differencing required is quite small, frequently one or two. (This relies on the fact many functions can be well approximated, on an interval of finite length, by a polynomial of reasonably low degree.)

1.5.2 Estimation and Elimination of Both Trend and Seasonality

Go back to Table of Contents. Please click TOC

Definition 1.29. Classical Decomposition Model

$$X_t = m_t + s_t + Y_t,\ t = 1, ..., n,$$

where $\mathbb{E}(Y_t) = 0$, $s_{t+d} = s_t$, and $\sum\limits_{j=1}^{d} s_j = 0$.

Method S1: Estimation of Trend and seasonal Components

Suppose we have $\{x_1, ..., x_n\}$. The trend is first estimated by applying a moving average fitler specially chosen to eliminate the seasonal component and to dampen the noise. If the period d is even, say $d = 2q$, then we use

$$\hat{m}_t = (0.5x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + 0.5X_{t+q})/d,\ q < t \leq n - q.$$

If the period is odd, say $d = 2q+1$, then we use the simple moving average.

The second step is to estimate the seasonal component. For each $k = 1, ..., d$, we compute the average w_k of the deviations $\{(x_{k+jd} - \hat{m}_{k+jd}), q < k + jd \leq n - q\}$. Since these average deviations do not necessarily sum to zero, we estimate the seasonal component s_k as

$$\hat{s}_k = w_k - \frac{1}{d}\sum_{i=1}^{d} w_i,\ k = 1, ..., d,$$

and $\hat{s}_k = \hat{s}_{k-d}$, $k > d$.

The *deseasonalized* data is then defined to be the original series with the estimated seasonal component removed, i.e.,

$$d_t = x_t - \hat{s}_t,\ t = 1, ..., n.$$

Finally we reestimate the trend from the deseasonalized data $\{d_t\}$ using one of the methods already described. We can fit a least squares polynomial trend $\hat{m}$ to the deseasonalized series. in terms of this reestimated trend and the estimated seasonal component, the estimated noise series is then given by

$$Y_t^2 = x_t - \hat{m}_t - \hat{s}_t,\ t = 1, ..., n.$$

Method S2: Elimination of Trend and Seasonal Components by Differencing par The technique of differencing that we applied earlier to nonseasonal data can be adapted to deal with seasonality of period d by intorducing the lag-d differencing operator $\triangledown_d$ defined by

$$\triangledown_d X_t = X_t - X_{t-d} = (1 - B^d)X_t.$$

(This operator should not be confused with the operator $\triangledown^d = (1 - B)^d$ defined earlier.)

Applying the operator $\triangledown_d$ to the model

$$X_t = m_t + s_t + Y_t,$$

where $\{s_t\}$ has period d, we obtain

$$\triangledown_d X_t = m_t - m_{t-d} + Y_t - Y_{t-d},$$

which gives a decomposition of the difference $\triangledown_d X_t$ into a trend component $(m_t - m_{t-d})$ and a noise term $(Y_t - Y_{t-d})$. The trend, $m_t - m_{t-d}$, can then be eliminated using the methods already described, inparticular by applying a power of the operator $\triangledown$.

1.6 Testing the Estimated Noise Sequence

Go back to Table of Contents. Please click TOC

The next step is to model the estimated noise sequence (i.e., the **residuals** obtained either by differencing the data or by estimating and subtracting the trend and seasonal components). In this section we examine some simple tests for checking the hypothesis that the residuals from Section 1.5 are observed values of independent and identically distributed random variables.

(a) *The sample autocorrelation function.* For large n, the sample autocorrelations of an iid sequence $Y_1, ..., Y_n$ with finite variance are approximately iid with distribution N(0,1/n). Hence, if $y_1, ..., y_n$ is a realization of such an iid sequence, about 95% of the sample autocorrelations should fall between the bounds $\pm 1.96/\sqrt{n}$. If we compute the sample autocorrelations up to lag 40 and find that more than two or three values fall outside the bounds, or that one value falls far outside the bounds, we therefore reject the iid hypothesis. The bounds $\pm 1.96/\sqrt{n}$ are automatically plotted when the sample autocorrelation function is computed.

(b) *The portmanteau test.* Instead of checking to see whether sample autocorrelation $\hat{\rho}(j)$ falls inside the bounds defined in (a) above, it is also possible to consider the single statistic

$$Q = n \sum_{j=1}^{h} \hat{\rho}^2(j).$$

If $Y_1, ..., Y_n$ is a finite-variance iid sequence, then by the same result used in (a), Q is approximately distributed as the sum of squares of the independent N(0,1) random variables, $\sqrt{n}\hat{\rho}(j)$, $j = 1, ..., h$, i.e., as chi-squared with h degrees of freedom. A large value of Q suggests that the sample autocorrelations of the data are too large for the data to be a sample from an iid sequence. We therefore reject the iid hypothesis at level α if $Q > \chi^2_{1-\alpha}(h)$, where $\chi^2_{1-\alpha}(h)$ is the $1-\alpha$ quantile of the chi-squared distribution with h degrees of freedom. Some programs conducts a refinement of this test, formulated by Ljung and Box (1978), in which Q is replaced by

$$Q_{LB} = n(n+2) \sum_{j=1}^{h} \hat{\rho}(j)/(n-j),$$

whose distribution is better approximated by the chi-squared distribution with h degrees of freedom.

Another portmanteau test, formulated by Mcleod and Li (83) [24], can be used as a further test for the iid hypothesis, since if the data are iid, then the squared data are also iid. It is based on the same statistic used for Ljung-Box test, except that the sample autocorrelations of the data are replaced by the sample autocorrelations of the *squared* data, $\hat{\rho}_{WW}(k)/(n-k)$.

$$Q_{ML} = n(n+2) \sum_{k=1}^{h} \hat{\rho}_{WW}(k)/(n-k).$$

The hypothesis of iid data is then rejected at level α if the observed value of Q_{ML} is larger than the $1-\alpha$ quantile of the $\chi^2(h)$ distribution.

(c) *The turning point test.* If $y_1, ..., y_n$ is a sequence of observations, we say that there is a turning point at time i, $1 < i < n$, if $y_{i-1} < y_i$ and $y_i > y_{i+1}$ of if $y_{i-1} > y_i$ and $y_i < y_{i+1}$. If T is the number of turning points of an iid sequence of length n, then, since the probability of a turning point at time i is $\frac{2}{3}$, the expected value of T is

$$\mu_T = \mathbb{E}(T) = \frac{2}{3}(n-2).$$

It can also be shown for an iid sequence that the variance of T is

$$\sigma_T^2 = Var(T) = (16n - 29)/90.$$

A large value of $T - \mu_T$ indicates that the series is fluctuating more rapidly than expected for an iid sequence. On the other hand, a value of $T - \mu_T$ much smaller than zero indicates a positive correlation between neighboring observations. For an iid sequence with n large, it can be shown that

$$T \approx N(\mu_T, \sigma_T^2).$$

This means we can carry out a test of the iid hypothesis, rejecting it at level α if $|T - \mu_t|/\sigma_T > \Phi_{1-\alpha/2}$, where $\Phi_{1-\alpha/2}$ quantile of the standard normal distribution. (A commonly used value of α is 0.05, for which the corresponding value of $\Phi_{1-\alpha/2}$ is 1.96.)

(d) *The difference-sign test.* For this test we count the number S of values of i such that $y_i > y_{i-1}$, $i = 2, ..., n$, or equivalently the number of times the differenced series $y_i - y_{i-1}$ is positive. For an iid sequence it is clear that

$$\mu_S = \mathbb{E}(S) = \frac{1}{2}(n-1).$$

It can also be shown, under the same assumption, that

$$\sigma_S^2 = Var(S) = (n+1)/12,$$

and that for large n,

$$S \approx N(\mu_S, \sigma_S^2).$$

A large positive (or negative) value of $S - \mu_S$ indicates the presence of an increasing (or decreasing) trend in the data. we therefore reject the assumption of no trend in the data if $|S - \mu_S|/\sigma_S > \Phi_{1-\alpha/2}$.

(e) *The rank test.* The rank test is particularly useful for detecting a linear trend in the data. Define P to be the number of pairs (i, j) such that $y_j > y_i$ and $j > i$, $i = 1, ..., n-1$. There is a total of $\binom{n}{2} = \frac{1}{2}n(n-1)$ pairs (i, j) such that $j > i$. For an iid sequence $\{Y_t, ..., Y_n\}$, each event $\{Y_j > Y_i\}$ has probability $\frac{1}{2}$, and the mean of P is therefore

$$\mu_P = \frac{1}{4}n(n-1).$$

It can also be shown for an iid sequence that the variance of P is

$$\sigma_P^2 = n(n-1)(2n+5)/72$$

and that for large n,

$$P \approx N(\mu_P, \sigma_P^2)$$

see Kendall and Stuart, 1976). A large positive (negative) value of $P - \mu_P$ indicates the presence of an icnreasing (decreasing) trend in the data. The assumption that $\{y_j\}$ is a sample from an iid sequence is therefore rejected at level $\alpha = 0.05$ if $|P - \mu_P|/\sigma_P > \Phi_{1-\alpha/2} = 1.96$.

(f) *Fitting an autoregressive model.* A further test that can be carried out is to fit an autoregressive model to the data using the Yule-Walker algorithm (discussed in Section 5.1.1) and choosing the order which minimizes the AICC statistic (see Section 5.5). A selected order equal to zero suggests that the data is white noise.

(g) *Checking for normality.* If the noise process if Gaussian, i.e., if all of its joint distributions are normal, then stronger conclusions can be drawn when a model is fitted to the data.

Let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ be the order statistics of a random sample $Y_1, ..., Y_n$ from the distribution N(μ, σ^2). If $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are the order statistics from a N(0,1) sample of size n, then

$$\mathbb{E}(Y_{(j)}) = \mu + \sigma m_j,$$

where $m_j = \mathbb{E}(X_{(j)})$, $j = 1, ..., n$.

The graph of the points $(m_1, Y_{(1)}), ..., (m_n, Y_{(n)})$ is called a Gaussian **qq plot**. If the normal assumption is correct, the Gaussian qq plot should be approximately linear. Consequently, the squared correlation of the points $(m_i, Y_{(i)})$, $i = 1, ..., n$, should be near 1. The assumption of normality is therefore rejected if the squared correlation R^2 is sufficiently small. If we approximate m_i by $\Phi^{-1}((i-0.5)/n)$, then R^2 reduces to

$$R^2 = \frac{(\sum_{i=1}^{n}(Y_{(i)} - \bar{Y})\Phi^{-1}(\frac{i-0.5}{n}))^2}{\sum_{i=1}^{n}(Y_{(i)} - \bar{Y})^2 \sum_{i=1}^{n}(\Phi^{-1}(\frac{i-0.5}{n}))^2},$$

where $\bar{Y} = n^{-1}(Y_1 + \cdots + Y_n)$. Percentage points for the distribution of R^2, assuming normality of the sample values, are given by Shapiro and Francia (1972) [33] for sample sizes $n < 100$. For $n = 200$, $P(R^2 < 0.987) = 0.05$ and $P(R^2 < 0.989) = 0.10$.

2 Stationary Processes

Go back to Table of Contents. Please click TOC A key role in time series analysis is played by processes whose properties, or some of them, do not vary with time. If we wish to make predictions, then clearly we must assume that *something* does not vary with time.

2.1 Basic Properties

Go back to Table of Contents. Please click TOC

We introduced the concept of stationarity and defined the autocovariance function (ACVF) of a stationary time series $\{X_t\}$ as

$$\gamma(h) = Cov(X_{t+h}, X_t),\ \ h = 0, \pm 1, \pm 2, ...$$

The autocorrelation function (ACF) of $\{X_t\}$ was defined similarly as the function $\rho(\cdot)$ whose value at lag h is

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

The ACVF and ACF provide a useful measure of the degree of dependence among the values of a time series at different times and for this resason play an important role when we consider the prediction of future values of the series in terms of the past and present values.

Suppose that $\{X_t\}$ is a stationary Gaussian time series and that we have observed X_n. We would like to find the function of X_n that gives us the best predictor of X_{n+h}, the value of the series after another h time units have elapsed. We must first define "best". A natural and computationally convenient definition is to specify our required predictor to be the function of X_n with minimum mean squared error. The conditional distribution of X_{n+h} given that $X_n = x_n$ is

$$N(\mu + \rho(h)(x_n - \mu), \sigma^2(1 - \rho(h)^2)),$$

where μ and σ^2 are the mean and variance of $\{X_t\}$. The value of the constant c that minimizes $\mathbb{E}(X_{n+h} - c)^2$ is $c = \mathbb{E}(X_{n+h})$ and that the function m of X_n that minimizes $\mathbb{E}(X_{n+h} - m(X_n))^2$ is the conditional mean

$$m(X_n) = \mathbb{E}(X_{n+h}|X_n) = \mu + \rho(h)(X_n - \mu).$$

The corresponding mean squared error is

$$\mathbb{E}(X_{n+h} - m(X_n))^2 = \sigma^2(1 - \rho(h)^2).$$

This calculation shows that at least for stationary Gaussian time series, prediction of X_{n+h} in terms of X_n is more accurate as $|\rho(h)|$ becomes closer to 1, and in the limit as $\rho \to \pm 1$ the best predictor approaches $\mu \pm (X_n - \mu)$ and the corresponding mean squared error approaches 0.

In the preceding calculation the assumption of joint normality of X_{n+h} and X_n played a crucial role. For time series with nonnormal joint distributions the corresponding calculations are in general much more complicated. However, if instead of looking for the best function of X_n for predicting X_{n+h}, we look for the best **linear predictor**, i.e., the best predictor of the form $\ell(X_n) = aX_n + b$, then our problem becomes that of finding a and b to minimize $\mathbb{E}(X_{n+h} - aX_n - b)^2$. An elementary calculation shows that the best predictor of this form is

$$l(X_n) = \mu + \rho(h)(X_n - \mu)$$

with corresponding mean squared error

$$\mathbb{E}(X_{n+h} - \ell(X_n))^2 = \sigma^2(1 - \rho(h)^2).$$

Proposition 2.1. ***Basic Properties of*** $\gamma(\cdot)$***:***

$$\gamma(0) \geq 0,$$

$$|\gamma(h)| \leq \gamma(0) \; for \; all \; h,$$

and $\gamma(\cdot)$ *is even, i.e.,*

$$\gamma(h) = \gamma(-h) \; for \; all \; h.$$

Definition 2.2. A real-valued function κ defined on the integers is **nonnegative definite** if

$$\sum_{i,j=1}^{n} a_k \kappa(i-j) a_j \geq 0$$

for all positive integers n and vectors $\mathbf{a} = (a_1, ..., a_n)'$ with real-valued components a_i.

Theorem 2.3. *A real-valued function defined on the integers is the autocovariance function of a stationary time series if and only if it is even and nonnegative definite.*

Proof: Let $\mathbf{a}$ be any $n \times 1$ vector with real components $a_1, ..., a_n$ and let $\mathbf{X}_n = (X_n, ..., X_1)'$. Then

$$Var(\mathbf{a}'\mathbf{X}_n) = \mathbf{a}'\Gamma_n \mathbf{a} = \sum_{i,j=1}^{n} a_i \gamma(i-j) a_j \geq 0,$$

where Γ_n is the covariance matrix of the ranom vector $\mathbf{X}_n$. The last inequality, however, is precisely the statement that $\gamma(\cdot)$ is nonnegative definite. The converse result, that there exists a stationary time series with autocovariance function κ if κ is even, real-valued, and nonnegative definite, is more difficult to establish. A slightly stronger statement can be made, namely, that under the specified conditions there exists a stationary *Gaussian* time series $\{X_t\}$ with mean 0 and autocovariance function $\kappa(\cdot)$.

Q.E.D.

Remark 2.4. An autocorrelation function $\rho(\cdot)$ has all the properties of an autocovariance function and satisfies the additional condition $\rho(0) = 1$. In particular, we can say that $\rho(\cdot)$ is the autocorrelation function of a stationary process if and only if $\rho(\cdot)$ is an ACVF with $\rho(0) = 1$.

□

Remark 2.5. To verify that a given function is nonnegative definite it is often simpler to find a stationary process that has the given function as its ACVF than to verify the conditions dirrectly. For example, the function $\kappa(h) = \cos(\omega h)$ is nonnegative definite, since it is the ACVF of the stationary process

$$X_t = A\cos(\omega t) + B\sin(\omega t),$$

where A and B are uncorrelated random variables, both with mean 0 and variance 1.

□

Definition 2.6. $\{X_t\}$ is a **strictly stationary time series** if

$$(X_1, ..., X_n)' \overset{d}{=} (X_{1+h}, ..., X_{n+h})'$$

for all integers h and $n \geq 1$. (Here $\overset{d}{=}$ is used to indicate that the two random vectors have the same joint distribution function.)

Proposition 2.7. ***Properties of a Strictly Stationary Time Sereis*** $\{X_t\}$***:***

(a) The random variables X_t are identically distributed.

(b) $(X_t, X_{t+h})' \overset{d}{=} (X_1, X_{1+h})'$ for all integers t and h.

(c) $\{X_t\}$ is weakly stationary if $\mathbb{E}(X_t^2) < \infty$ for all t.

(d) Weak stationarity does not imply strict stationarity.

(e) An iid sequence is strictly stationary.

Proof: Properties (a) and (b) follow at once from Definition 2.6. If $\mathbb{E}(X_t^2) < \infty$, then by (a) and (b) $\mathbb{E}(X_t)$ is independent of t and $Cov(X_t, X_{t+h}) = Cov(X_1, X_{1+h})$, which is also independent of t, proving (c). For (d), it is in problem 1.8 of textbook [9]. If $\{X_t\}$ is an iid sequence of random variables with common distribution function F, then the joint distribution function of $(X_{1+h}, ..., X_{n+h})'$ evaluated at $(x_1, ..., x_n)'$ is $F(x_1)...F(x_n)$, which is independent of h.

Q.E.D.

One of the simplest ways to constrct a time series $\{X_t\}$ that is strictly stationary (and hence stationary if $\mathbb{E}(X_t^2) < \infty$) is to "filter" an iid sequence of random variables. Let $\{Z_t\}$ be an iid sequence, which by (e) is strictly stationary, and define

$$X_t = g(Z_t, Z_{t-1}, ..., Z_{t-q})$$

for some real-valued function $g(\cdot, ..., \cdot)$. Then $\{X_t\}$ is strictly stationary, since $(Z_{t+h}, ..., Z_{t+h-q})' \overset{d}{=} (Z_t, ..., Z_{t-q})'$ for all integers h. It follows also from the defining equation above that $\{X_t\}$ is q**-dependent**, i.e., that X_s and X_t are independent whenver $|t - s| > q$. (An iid sequence if 0-dependent.) In the same way, abopting a second-order viewpoint, we say that a stationary time series is q**-correlated** if $\gamma(h) = 0$ whever $|h| > q$. A white noise sequence is then 0-correlated, while the MA(1) process is 1-correlated. The moving-average process of order q defined below is q-correlated, and perhaps surprisingly, the converse is also true.

Proposition 2.8. ***The MA(q) Process:***
*$\{X_t\}$ is a **moving-average process of order** q if*

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

where $\{Z_t\} \sim WN(0, \sigma^2)$ and $\theta_1, ..., \theta_q$ are constants.

Proposition 2.9. *If $\{X_t\}$ is a stationary q-correlated time series with mean 0, then it can be represented as the MA(q) process in Proposition 2.8.*

2.2 Linear Processes

Go back to Table of Contents. Please click TOC

Definition 2.10. The time series $\{X_t\}$ is a **linear process** if it has the representation

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

for all t, where $\{Z_t\} \sim WN(0, \sigma^2)$ and $\{\psi_j\}$ is a sequence of constants with $\sum\limits_{j=-\infty}^{\infty} |\psi_j| < \infty$.

In terms of the backward shfit operator B, the above equation can be written more compactly as

$$X_t = \psi(B) Z_t,$$

where $\psi(B) = \sum\limits_{j=-\infty}^{\infty} \psi_j B^j$. A linear process is called a **moving average** or **MA(∞)** if $\psi_j = 0$ for all $j < 0$, i.e., if

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

Remark 2.11. The condition $\sum\limits_{j=-\infty}^{\infty} |\Psi_j| < \infty$ ensures that the infinite sum in the definition converges (with probability one), since $\mathbb{E}(|Z_t|) \leq \sigma$ and

$$\mathbb{E}(|X_t|) \leq \sum_{j=-\infty}^{\infty} (|\psi_j| \mathbb{E}(|Z_{t-j}|)) \leq \left(\sum_{j=-\infty}^{\infty} |\psi_j| \right) \sigma < \infty.$$

It also ensures that $\sum\limits_{j=-\infty}^{\infty} \psi_j^2 < \infty$ and hence that the series in definition converges in mean square, i.e., that X_t is the mean square limit of the partial sums $\sum\limits_{j=-n}^{n} \psi_j Z_{t-j}$. The condition $\sum\limits_{j=-n}^{n} |\psi_j| < \infty$ also ensures convergence in both senses of the more general series in definition considered in Proposition below.

□

Proposition 2.12. *Let $\{Y_t\}$ be a stationary time series with mean 0 and covariance function γ_Y. If $\sum\limits_{j=-\infty}^{\infty} |\psi_j| < \infty$, then the time series*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} = \psi(B) Y_t$$

is stationary with mean 0 and autocovariance function

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h + k - j).$$

In the special case where $\{X_t\}$ is a linear process,

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

Proof: With σ replaced by $\sqrt{\gamma_Y(0)}$, it shows that the series in the first equation in the Proposition 2.12 is convergent. Since $\mathbb{E}(Y_t) = 0$, we have

$$\mathbb{E}(X_t) = \mathbb{E}\bigg(\sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} \bigg) = \sum_{j=-\infty}^{\infty} \psi_j \mathbb{E}(Y_{t-j}) = 0$$

and

$$\begin{aligned} \mathbb{E}(X_{t+h}X_t) &= \mathbb{E}\bigg[\bigg(\sum_{j=-\infty}^{\infty} \psi_j Y_{t+h-j} \bigg) \bigg(\sum_{k=-\infty}^{\infty} \psi_k Y_{t-k} \bigg) \bigg] \\ &= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \mathbb{E}(Y_{t+h-j} Y_{t-k}) \\ &= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h-j+k), \end{aligned}$$

which shows that $\{X_t\}$ is stationary with covariance function $\gamma_X(h)$. (The interchange of summation and expectation operations in the above calculations can be justified by the absolute summability of ψ_j.) Finally, if $\{Y_t\}$ is the white noise sequence $\{Z_t\}$ in Definition 2.10, then $\gamma_Y(h-j+k) = \sigma^2$ if $k = j-h$ and 0 otherwise, from which the final equation in Proposition 2.12 follows.

Q.E.D.

Remark 2.13. The absolute convergence of X_t in Proposition 2.12 implies that filters of the form $\alpha(B) = \sum_{j=-\infty}^{\infty} \alpha_j B^j$ and $\beta(B) = \sum_{j=-\infty}^{\infty} \beta_j B^j$ with absolutely summable coefficients can be applied successfively to a stationary series $\{Y_t\}$ to generate a new stationary series

$$W_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j},$$

where

$$\psi_j = \sum_{k=-\infty}^{\infty} \alpha_k \beta_{j-k} = \sum_{k=-\infty}^{\infty} \beta_k \alpha_{j-k}.$$

These relations can be expressed in the equivalent form

$$W_t = \psi(B) Y_t,$$

where

$$\psi(B) = \alpha(B)\beta(B) = \beta(B)\alpha(B),$$

and the products are defined by ψ_j or equivalently by multiplying the series $\sum_{j=-\infty}^{\infty} \alpha_j B^j$ and $\sum_{j=-\infty}^{\infty} \beta_j B^j$ term by term and collecting powers of B. It is clear from the equations above the order of application of the filters $\alpha(B)$ and $\beta(B)$ is immaterial.

□

2.3 Introduction to ARMA Processes

Go back to Table of Contents. Please click TOC

Definition 2.14. The time series $\{X_t\}$ is an **ARMA(1,1) process** if it is stationary and satisfies (for every t)

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

where $\{Z_t\} \sim WN(0, \sigma^2)$ and $\phi + \theta \neq 0$.

Using the backward shift operator B, the above equation can be written more concisely as

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

where $\phi(B)$ and $\theta(B)$ are the linear filters

$$\phi(B) = 1 - \phi B \; and \; \theta(B) = 1 + \theta B,$$

respectively.

We investigate the range of values of ϕ and θ for which a stationary solution of definition exists. If $|\phi| < 1$, let $\chi(z)$ denote the power series expansion of $1/\phi 9z)$, i.e., $\sum_{j=0}^{\infty} \phi^j z^j$, which has absolutely summable coefficients. Then from $\psi(B) = \alpha(B)\beta(B) = \beta(B)\alpha(B)$ we conclude that $\chi(B)\phi(B) = 1$. Applying $\chi(B)$ to each side of $\phi(B)X_t = \theta(B)Z_t$ therefore gives

$$X_t = \chi(B)\theta(B)Z_t = \psi(B)Z_t,$$

where

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j = (1 + \phi B + \phi^2 B^2 + \dots)(1 + \theta B).$$

By multiplying out the right-hand side or using $\psi_j = \sum_{k=-\infty}^{\infty} \alpha_k \beta_{j-k} = \sum_{k=-\infty}^{\infty} \beta_k \alpha_{j-k}$, we find that

$$\psi_0 = 1 \; and \; \psi_j = (\phi + \theta)\phi^{j-1} \; for \; j \geq 1.$$

We conclude that the MA(∞) proceess

$$X_t = Z_t + (\phi + \theta) \sum_{j=1}^{\infty} \phi^{j-1} Z_{t-j}$$

is the unique stationary solution of the definition.

Now suppose that $\|phi| > 1$. We first represent $1/\phi 9z)$ as a series of powers of z with absolutely summable coefficients by expanding in powers of z^{-1}, giving

$$\frac{1}{\phi(z)} = -\sum_{j=1}^{\infty} \phi^{-j} z^{-j}.$$

Then we can apply the same argument as in the case where $|\phi| < 1$ to obtain the unique stationary solution of the definition. We let $\chi(B) = -\sum_{j=1}^{\infty}\phi^{-j}B^{-j}$ and apply $\chi(B)$ to each side of $\phi(B)X_t = \theta(B)Z_t$ to obtain

$$X_t = \chi(B)\theta(B)Z_t = -\theta\phi^{-1}Z_t - (\theta+\phi)\sum_{j=1}^{\infty}\phi^{-j-1}Z_{t+j}.$$

If $\phi = \pm 1$, there is no stationary solution of definition. Consequently, there is no such thing as an ARMA(1,1) process with $\phi = \pm 1$ according to definition.

We can now summarize our finds about the existence and nature of the stationary solutions of the ARMA(1,1) recursions as follows.

- A stationary solution of the ARMA(1,1) equations exists if and only if $\phi \neq \pm 1$.
- If $|\phi| < 1$, then the unique stationary solution is given by $X_t = Z_t + (\phi+\theta)\sum_{j=1}^{\infty}\phi^{j-1}Z_{t-j}$. In this case we say that $\{X_t\}$ is **causal** or a causal function of $\{Z_t\}$, since X_t can be expressed in terms of the current and past values Z_s, $s \leq t$.
- If $|\phi| > 1$, then the unique stationary solution is given by X_t. The solution is **noncausal**, since X_t is then a function of Z_s, $s \geq t$.

Just as causality means that X_t is expressible in terms of Z_s, $s \geq t$, the dual concept of invertibility means that Z_t is expressible in terms of X_s, $s \leq t$. We show now that the ARMA(1,1) process defined by definition is invertible if $|\theta| < 1$. To demonstrate this, let $\xi(z)$ denote the power series expansion of $1/\theta(z)$, i.e., $\sum_{j=0}^{\infty}(-\theta)^j z^j$, which has absolutely summable coeffients. From $\psi(B)$ it therefore follows that $\xi(B)\theta(B) = 1$, and applying $\xi(B)$ to each side of $\phi(B)X_t$ gives

$$Z_t = \xi(B)\phi(B)X_t = \pi(B)X_t,$$

where

$$\pi(B) = \sum_{j=0}^{\infty}\pi_j B^j = (1 - \theta B + (-\theta)^2B^2 + \dots)(1 - \phi B).$$

By multiplying out the right-hand side or using ψ_j, we find that

$$Z_t = X_t - (\phi+\theta)\sum_{j=1}^{\infty}(-\theta)^{j-1}X_{t-j}.$$

Thus the ARMA(1,1) process is **invertible**, since Z_t can be expressed in terms of the present and past values of the process X_s, $s \leq t$. An argument like the one used to show noncausality when $|\phi| > 1$ shows that eh ARMA(1,1) process is **noninvertible** when $|\theta| > 1$, since then

$$Z_t = -\phi\theta^{-1}X_t + (\theta+\phi)\sum_{j=1}^{\infty}(-\theta)^{-j-1}X_{t+j}.$$

We summarize these results as follows:

- If $|\theta| < 1$, then the ARMA(1,1) process is **invertible**, and Z_t is expressed in terms X_s, $s \leq t$, by Z_t.
- If $|\theta| > 1$, then the ARMA(1,1) process is **noninvertible**, and Z_t is expressed in terms of X_s, $s \geq t$.

2.4 Properties of the Sample Mean and Autocorrelation Function

Go back to Table of Contents. Please click TOC

A stationary process $\{X_t\}$ is characterized, at least from a second-order point of view, by its mean μ and its autocovariance function $\gamma(\cdot)$. The estimation of μ, $\gamma(\cdot)$, and the autocorrelation function $\rho(\cdot) = \gamma(\cdot)/\gamma(0)$ from observations $X_1, ..., X_n$ therefore plays a crucial role in problems of inference and in particular in the problem of constructing an appropriate model for the data. In this examine some of the properties of the sample estimates $\bar{x}$ and $\hat{\rho}(\cdot)$ of μ and $\rho(\cdot)$, respectively.

2.4.1 Estimation of μ

Go back to Table of Contents. Please click TOC

The moment estimator of the mean μ of a stationary process is the sample mean

$$\bar{x} = n^{-1}(X_1 + X_2 + \cdots + X_n).$$

It is an unbiased estimator of μ, since

$$\mathbb{E}\big(\bar{x}_n) = n^{-1}(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)\big) = \mu.$$

The mean squared error of $\bar{X}_n$ is

$$\begin{aligned}
\mathbb{E}(\bar{X}_n - \mu)^2 &= Var(\bar{X}_n) \\
&= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) \\
&= n^{-2}\sum_{i-j=-n}^{n}(n - |i-j|)\gamma(i-j) \\
&= n^{-1}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma(h).
\end{aligned}$$

Now if $\gamma(h) \to 0$ as $h \to \infty$, the right-hand side of the result above converges to zero, so that $\bar{X}_n$ converges in mean square to μ. If $\sum_{h=-\infty}^{\infty}|\gamma(h)| < \infty$, then the result gives $\lim_{n\to\infty} nVar(\bar{X}_n) = \sum_{|h|<\infty}\gamma(h)$.

Proposition 2.15. *If $\{X_t\}$ is a stationary time series with mean μ and autocovariance function $\gamma(\cdot)$, then as $n \to \infty$,*

$$Var(\bar{X}_n) = \mathbb{E}(\bar{X}_n - \mu) \to 0 \ if \ \gamma(n) \to 0,$$

$$n\mathbb{E}(\bar{X}_n - \mu)^2 \to \sum_{|h|<\infty}\gamma(h) \ if \ \sum_{h=-\infty}^{\infty}|\gamma(h)| < \infty.$$

To make inferences about μ using the sample mean $\bar{X}_n$, it is necessary to know the distribution or an approximation to the distribution of $\bar{X}_n$. If the time series is Gaussian, then

$$n^{1/2}(\bar{X}_n - \mu) \sim N\left(0, \sum_{|h|<n}\left(1 - \frac{|h|}{n}\right)\gamma(h)\right).$$

It is easy to construct exact confidence bounds for μ using this result if $\gamma(\cdot)$ is known, and approximate confidence bounds if it is necessary to estimate $\gamma(\cdot)$ from the observations.

2.4.2 Estimation of $\gamma(\cdot)$ and $\rho(\cdot)$

Go back to Table of Contents. Please click TOC

Recall that the sample autocovariance and autocorrelation functions are defined by

$$\hat{\gamma}(h) = n^{-1}\sum_{t=1}^{n-|h|}(X_{t+|h|} - \bar{X}_n)(X_t - \bar{X}_n)$$

and

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

Both the estimators $\hat{\gamma}(h)$ and $\hat{\rho}(h)$ are biased even if the factor n^{-1} is replaced by $(n-h)^{-1}$. Nevertheless, under general assumptions they are nearly unbiased for large sample sizes. The sample ACVF has the desirable property that for each $k \geq 1$ the k-dimensional sample covariance matrix

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

is nonnegative definite. To see this, first note that if $\hat{\Gamma}_m$ is nonnegative definite, then $\hat{\Gamma}_k$ is nonnegative definite for all $k < m$. So assume $k \geq n$ and write

$$\hat{\Gamma}_k = n^{-1}TT',$$

where T is the $k \times 2k$ matrix

$$T = \begin{bmatrix} 0 & \dots & 0 & 0 & Y_1 & Y_2 & \dots & Y_k \\ 0 & \dots & 0 & Y_1 & Y_2 & \dots & Y_k & 0 \\ \vdots & & & & & & & \vdots \\ 0 & Y_1 & Y_2 & \dots & Y_k & 0 & \dots & 0 \end{bmatrix}$$

$Y_i = X_i - \bar{X}_n$, $i = 1, ..., n$, and $Y_i = 0$ for $i = n+1, ..., k$. Then for any real $k \times 1$ vector $\mathbf{a}$ we have

$$\mathbf{a}'\hat{\gamma}_k\mathbf{a} = n^{-1}(\mathbf{a}'T)(T'\mathbf{a}) \geq 0,$$

and consequently the sample autocovariance matrix $\hat{\gamma}_k$ and sample autocorrelation matrix

$$\hat{R}_k = \hat{\Gamma}_k/\gamma(0)$$

are nonnegative definite. Sometimes the factor n^{-1} is replaced by $(n-h)^{-1}$ in the definition of $\hat{\gamma}(h)$, but the resulting covariance and correlation matrices $\hat{\Gamma}_n$ and $\hat{R}_n$ may not then by nonnegative definite.

Without further information beyond the observed data $X_1, ..., X_n$, it is impossible to give reasonable estimates of $\gamma(h)$ and $\rho(h)$ for $h \geq n$. Even for h slightly smaller than n, the estimates $\hat{\gamma}(h)$ and $\hat{\rho}(h)$ are unreliable, since there are so few pairs (X_{t+h}, X_t) available (only one if $h = n-1$).

The sample ACF plays an important role in the selection of suitable models for the data. We have seen examples how the sample ACF can be used to test for iid noise. For systematic inference concerning $\rho(h)$, we need the sampling distribution of the estimator $\hat{\rho}(h)$. Although the distribution of $\hat{\rho}(h)$ is intractable for samples from even the simplest time series models, it can usually be well approximated by a normal distribution for large sample sizes. For linear models and in particular for ARMA models $\hat{\rho}_k = (\hat{\rho}(1), ..., \hat{\rho}(k))'$ is approximately distributed for large n as $N(\rho_k, n^{-1}W)$, i.e.,

$$\hat{\rho} \approx N(\rho, n^{-1}W),$$

where $\rho = (\rho(1), ..., \rho(k))'$, and W is the covariance matrix whose (i, j) element is given by **Bartlett's formula**

$$\begin{aligned} w_{ij} \quad = \quad & \sum_{k=-\infty}^{\infty} \{\rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho^2(k) \\ & -2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+i)\} \end{aligned}$$

Simple algebra shows that

$$\begin{aligned} w_{ij} \quad = \quad & \sum_{k=1}^{\infty} \{\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\} \\ & \times\{\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\}, \end{aligned}$$

which is a more convenient form of w_{ij} for computational purposes.

Example 2.16. If $\{X_t\} \sim IID(0, \sigma^2)$, then $\rho(h) = 0$ for $|h| > 0$, so from w_{ij} (Bartlett's formula) we obtain

$$w_{ij} = \begin{cases} 1 & if\ i = j, \\ 0 & if\ otherwise. \end{cases}$$

For large n, therefore, $\hat{\rho}(1), ..., \hat{\rho}(h)$ are approximately independent and identically distributed normal random variables with mean 0 and variance n^{-1}. This result is the basis for the test that data are generated from iid noise using the sample ACF.

□

Example 2.17. If $\{X_t\}$ is the MA(1) process of Example 1.12 (page 13), i.e., if

$$X_t = Z_t + \theta Z_{t-1},\ t = 0, \pm 1, ...,$$

where $\{Z_t\} \sim WN(0, \sigma^2)$, then from Bartlett's formula,

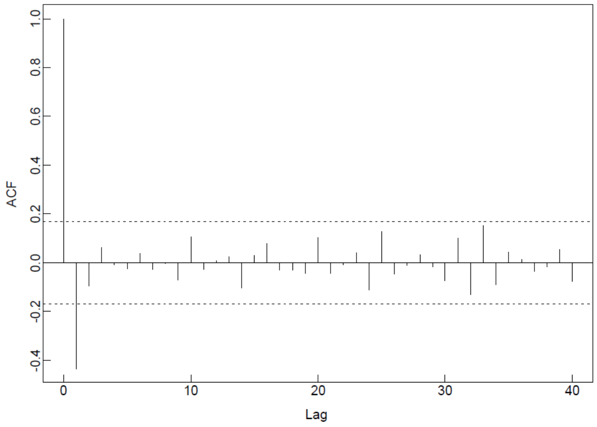
$$w_{ii} = \begin{cases} 1 - 3\rho^2(1) + 3\rho^4(1), & if\ i = 1, \\ 1 + 2\rho^2(1), & if\ i > 1, \end{cases}$$

is the approximate variance of $n^{-1/2}(\hat{\rho}(i) - \rho(i))$ for large n. In Figure 13 we have plotted the sample autocorrelation function $\hat{\rho}(k)$, $k = 0, ..., 40$, for 200 observations from the MA(1) model

$$X_t = Z_t - 0.8Z_{t-1},$$

where $\{Z_t\}$ is a sequence of iid N(0,1) random variables. Here $\rho(1) = -0.8/1.64 = -0.4878$ $\rho(h) = 0$ for $h > 1$. The lag-one sample ACF is found to be $\hat{\rho}(1) = -0.4333 = -6.128n^{-1/2}$, which would cause us (in the absence of our prior knowledge of $\{X_t\}$) to reject the hypothesis that the data are a sample from an iid noise sequence. The fact that $|\hat{\rho}(h)| \leq 1.96n^{-1/2}$ for $h = 2, ..., 40$ strongly suggests that the data are from a model in which observations are uncorrelated past lag 1. In figure 13, we have plotted the bounds $\pm 1.96n^{-1/2}(1 + 2\rho^2(1))^{1/2}$, indicating the compatibility of the data with othe model, $X_t = Z_t - 0.8Z_{t-1}$. Since, however, $\rho(1)$ is not normally known in advance, the autocorrelations $\hat{\rho}(2), ..., \hat{\rho}(40)$ would in practice have been compared with the more stringent bounds $\pm 1.96n^{-1/2}$ or with the bounds $\pm 1.96n^{-1/2}(1 + 2\hat{\rho}(1))^{1/2}$ in order to check the hypothesiss that the data are generated by moving-average process of order 1. Finally, it is worth noting that the lag-one correlation -0.4878 is well inside the 95% confidence bounds for $\rho(1)$ given by $\hat{\rho}(1) \pm 1.96n^{-1/2}(1 - 3\hat{\rho}^2(1) + 4\hat{\rho}^4(1))^{1/2} = -0.4333 \pm 0.1053$. This further supports the compatibility of the data with the model $X_t = Z_t - 0.8Z_{t-1}$.

Figure 13: The sample autocorrelation function of $n = 200$ observations of the MA(1) process, showing the bounds $\pm 1.96n^{-1/2}(1 + 2\hat{\rho}^2(1))^{1/2}$.



□

Example 2.18. For the AR(1) process,

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

where $\{Z_t\}$ is iid noise and $|\phi| < 1$, we have, from Bartlett's formula with $\rho(h) = \phi^{|h|}$,

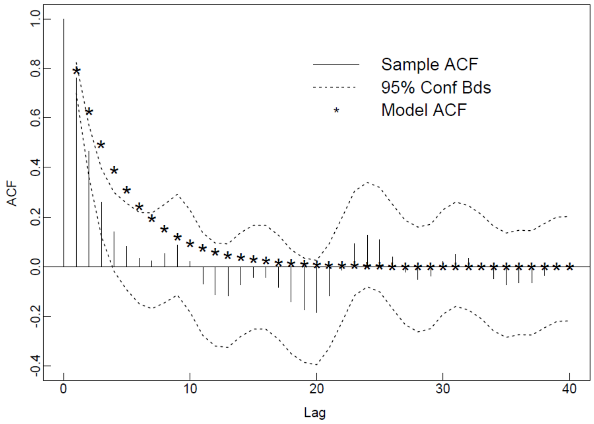
$$\begin{aligned} w_{ii} &= \sum_{k=1}^{i} \phi^{2i}(\phi^{-k} - \phi^{k})^2 + \sum_{k=i+1}^{\infty} \phi^{2k}(\phi^{-i} - \phi^{i})^2 \\ &= (1 - \phi^{2i})(1 + \phi^2)(1 - \phi^2)^{-1} - 2i\phi^{2i}, \end{aligned}$$

$i = 1, 2, \ldots$. In Figure 14 we have plotted the sample ACF of the Lake Huron residuals $y_1, \ldots, y_{98}$ together with 95% confidence bounds for $\rho(i)$, $i = 1, \ldots, 40$, assuming that data are generated from the AR(1) model

$$Y_t = 0.791 Y_{t-1} + Z_t$$

The confidence bounds are computed from $\hat{\rho} \pm 1.96 n^{-1/2} w_{ii}^{1/2}$, where w_{ii} is given with $\phi = -.791$. The model ACF, $\rho(i) = (0.791)^i$, is also plotted in figure below. Notice that the model ACF lies just outside the confidence bounds at lags 2-6. This suggests some incompatibility of the data with the model above. A much better fit to the residuals is provided by the second-order autoregressing, $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + Z_t$ (see text page 23 [9].

Figure 14: The sample autocorrelation function of the Lake Huron residuals showing the bounds $\hat{\rho}(i) \pm 1.96 n^{-1/2} w_{ii}^{1/2}$ and the model ACF $\rho(i) = (0.791)^i$.



□

2.5 Forecasting Stationary Time Series

Go back to Table of Contents. Please click TOC

Now consider the problem of predicting the values X_{n+h}, $h > 0$, of a stationary time series with known mean μ and autocovariance function γ in terms of the values $\{X_n, \ldots, X_1\}$, up to time n. Our goal is to find the *linear combination* of $1, X_n, X_{n-1}, \ldots, X_1$, that forecasts X_{n+h}

with minimum mean squared error. The best linear predictor in terms of $1, X_n, ..., X_1$ will be denoted by P_nX_{n+h} and clearly has the form

$$P_nX_{n+h} = a_0 + a_1X_n + \cdots + a_nX_1.$$

It remains only to determine the coefficients $a_0, a_1, ..., a_n$, by finding the values that minimize

$$S(a_0, ..., a_n) = \mathbb{E}(X_{n+h} - a_0 - a_1X_n - \cdots - a_nX_1)^2.$$

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

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

Evaluation of the derivatives in equation above gives the equivalent equations

$$\mathbb{E}\left[X_{n+h} - a_0 - \sum_{i=1}^{n} a_iX_{n+1-i}\right] = 0,$$

$$\mathbb{E}\left[(X_{n+h} - a_0 - \sum_{i=1}^{n} a_iX_{n+1-i})X_{n+1-j}\right] = 0, \; j = 1, ..., n.$$

These equations can be written more neatly in vector notation as

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

and

$$\Gamma_n \mathbf{a}_n = \gamma_n(h),$$

where

$$\mathbf{a}_n = (a_1, ..., a_n)', \; \Gamma_n = [\gamma(i-j)]_{i,j=1}^{n},$$

and

$$\gamma_n(h) = (\gamma(h), \gamma(h+1), ..., \gamma(h+n-1))'.$$

Hence,

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

where $\mathbf{a}_n$ satisfies $\Gamma_n\mathbf{a}_n$. From P_nX_{n+h} the expected value of the prediction error $X_{n+h} - P_nX_{n+h}$ is zero, and the mean square prediction error is therefore

$$\begin{aligned}
\mathbb{E}(X_{n+h} - P_nX_{n+h})^2 &= \gamma(0) - 2\sum_{i=1}^{n} a_i\gamma(h+i-1) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i\gamma(i-j)a_j \\
&= \gamma(0) - \mathbf{a}_n'\gamma_n(h), \\
&= \gamma(0) - \mathbf{a}_n'\Gamma_n\mathbf{a}_n
\end{aligned}$$

Remark 2.19. To show that equations $\mathbb{E}\left[X_{n+h} - a_0 - \sum_{i=1}^{n} a_iX_{n+1-i}\right] = 0$, and $\mathbb{E}\left[(X_{n+h} - a_0 - \sum_{i=1}^{n} a_iX_{n+1-i})X_{n+1-j}\right] = 0, \; j = 1, ..., n.$ determine

P_nX_{n+h} uniquely, let $\{a_j^{(1)}, j = 0, ..., n\}$ and $\{a_j^{(2)}, j = 0, ..., n\}$ be two solutions and let Z be the difference between the corresponding predictors, i.e.,

$$Z = a_0^{(1)} - a_0^{(2)} + \sum_{j=1}^{n}(a_j^{(1)} - a_j^{(2)})X_{n+1-j}.$$

Then

$$Z^2 = Z\biggl(a_0^{(1)} - a_0^{(2)} + \sum_{j=1}^{n}(a_j^{(1)} - a_j^{(2)})X_{n+1-j}\biggr).$$

But from the two expected values above we have $\mathbb{E}(Z) = 0$ and $\mathbb{E}(ZX_{n+1-j}) = 0$ for $j = 1, ..., n$. Consequently, $\mathbb{E}(Z^2) = 0$ and hence $Z = 0$.

□

Proposition 2.20. ***Properties of*** P_nX_{n+h}***:***

(1) $P_nX_{n+h} = \mu + \sum\limits_{i=1}^{n} a_i(X_{n+1-i} - \mu)$, *where* $\boldsymbol{a}_n = (a_1, ..., a_n)'$ *satisfies* $\Gamma_n \boldsymbol{a}_n$.

(2) $\mathbb{E}(X_{n+h} - P_nX_{n+h})^2 = \gamma(0) - \boldsymbol{a}_n'\gamma_n(h)$, *where* $\gamma_n(h) = (\gamma(h), ..., \gamma(h+n-1))'$.

(3) $\mathbb{E}(X_{n+h} - P_nX_{n+h}) = 0$.

(4) $\mathbb{E}[X_{n+h} - P_nX_{n+h})X_j] = 0, j = 1, ..., n$.

Remark 2.21. Notice that properties 3 and 4 are exactly equivalent to $\mathbb{E}\biggl[X_{n+h} - a_0 - \sum\limits_{i=1}^{n} a_iX_{n+1-i}\biggr] = 0$ and $\mathbb{E}\biggl[(X_{n+h} - a_0 - \sum\limits_{i=1}^{n} a_iX_{n+1-i})X_{n+1-j}\biggr] = 0$, $j = 1, ..., n$. They can be written more succinctly in the form $\mathbb{E}[(Error)\times(Predictor\ Variable)] = 0$, which uniquely determine P_nX_{n+h}.

□

Example 2.22. Consider now the stationary time series defined by

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

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

$$P_nX_{n+1} = \mathbf{a}_n'\mathbf{X}_n,$$

where $\mathbf{X}_n = (X_n, ..., X_1)'$ and

$$\begin{bmatrix} 1 & \phi & \phi^2 & \dots & \phi^{n-1} \\ \phi & 1 & \phi & \dots & \phi^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \dots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \phi \\ \phi^2 \\ \vdots \\ \phi^n \end{bmatrix}$$

A solution of the above equation is

$$\mathbf{a}_n = (\phi, 0, ..., 0)',$$

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

$$P_nX_{n+1} = \mathbf{a}_n'\mathbf{X}_n = \phi X_n,$$

with mean squared error

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

□

Prediction of Second-Order Random Variables

Suppose that Y and $W_n, ..., W_1$ are *any* random variables with finite second moments and that the means $\mu = \mathbb{E}(Y)$, $\mu_i = \mathbb{E}(W_i)$ and covariances $Cov(Y, Y)$, $Cov(Y, W_i)$, and $Cov(W_i, W_j)$ are all known. It is convenient to introduce the random vector $\mathbf{W} = (W_n, ..., W_1)'$, the corresponding vector of means $\mu_W = (\mu_n, ..., \mu_1)'$, the vector of covariances

$$\gamma = Cov(Y, \mathbf{W}) = (Cov(Y, W_n), Cov(Y, W_{n-1}), ..., Cov(Y, W_1))',$$

and the covariance matrix

$$\Gamma = Cov(\mathbf{W}, \mathbf{W}) = [Cov(W_{n+1-i}, W_{n+1-j})]_{i,j=1}^n.$$

Then by the same arguments used in the calculation of $P_n X_{n+h}$, the best linear predictor of Y in terms of $\{1, W_n, .., W_n\}$ is found to be

$$P(Y|\mathbf{W}) = \mu_Y + \mathbf{a}(\mathbf{W} - \mu_W),$$

where $\mathbf{a} = (a_1, ..., a_n)'$ is any solution of

$$\Gamma\mathbf{a} = \gamma.$$

The mean squared error of the predictor is

$$\mathbb{E}[(Y - P(Y|\mathbf{W}))^2] = Var(Y) - \mathbf{a}'\gamma.$$

Example 2.23. Consider again the stationary series by

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

where $|\phi| < 1$ and $\{Z_t\} \sim WN(0, \sigma^2)$. Suppose that we observe the series at times 1 and 3 and wish to use these observations to find the linear combination of 1, X_1, and X_3 that estimates X_2 wth minimum mean squared error. The solution to this problem can be obtained directly from $P(Y|\mathbf{W}) = \mu_Y + \mathbf{a}(\mathbf{W} - \mu_W)$, and $\Gamma\mathbf{a} = \gamma$. by setting $Y = X_2$ and $\mathbf{W} = (X_1, X_3)'$. This gives the equations

$$\begin{bmatrix} 1 & \phi^2 \\ \phi^2 & 1 \end{bmatrix} \mathbf{a} = \begin{bmatrix} \phi \\ \phi \end{bmatrix},$$

with solution

$$\mathbf{a} = \frac{1}{1+\phi^2} \begin{bmatrix} \phi \\ \phi \end{bmatrix}.$$

The best estimator of X_2 is thus

$$P(X_2|\mathbf{W}) = \frac{\phi}{1+\phi^2}(X_1 + X_3),$$

with mean squared error

$$\mathbb{E}[(X_2 - P(X_2|\mathbf{W}))^2] = \frac{\sigma^2}{1-\phi^2} - \mathbf{a}' \begin{bmatrix} \frac{\phi\sigma^2}{1-\phi^2} \\ \frac{\phi\sigma^2}{1-\phi^2} \end{bmatrix} = \frac{\sigma^2}{1+\phi^2}.$$

□

Proposition 2.24. ***Properties of the Prediction Operator*** $P(\cdot|\,\boldsymbol{W})$***:***
Suppose that $\mathbb{E}(U^2) < \infty$, $\mathbb{E}(V^2) < \infty$, $\Gamma = cov(\boldsymbol{W}, \boldsymbol{W})$, *and* β, α_1, ..., α_n *are constants.*

(1) $P(U|\,\boldsymbol{W}) = \mathbb{E}(U) + \boldsymbol{a}'(\boldsymbol{W} - \mathbb{E}(\boldsymbol{W}))$, *where* $\Gamma\boldsymbol{a} = cov(U, \boldsymbol{W})$.

(2) $\mathbb{E}[(U - P(U|\,\boldsymbol{W}))\,\boldsymbol{W}] = 0$ *and* $\mathbb{E}[U - P(U|\,\boldsymbol{W})] = 0$.

(3) $\mathbb{E}[(U - P(U|\,\boldsymbol{W}))^2] = var(U) - \boldsymbol{a}'cov(U, \boldsymbol{W})$.

(4) $P(\alpha_1 U + \alpha_2 V + \beta|\,\boldsymbol{W}) = \alpha_1 P(U|\,\boldsymbol{W}) + \alpha_2 P(V|\,\boldsymbol{W}) + \beta$.

(5) $P(\sum\limits_{i=1}^{n}\alpha_i W_i + \beta|\,\boldsymbol{W}) = \sum\limits_{i=1}^{n}\alpha_i W_i + \beta$.

(6) $P(U|\,\boldsymbol{W}) = \mathbb{E}(U)$ *if* $cov(U, \boldsymbol{W}) = 0$.

(7) $P(U|\,\boldsymbol{W}) = P(P(U|\,\boldsymbol{W}, \boldsymbol{V})|\,\boldsymbol{W})$ *if* $\boldsymbol{V}$ *is a random vector such that the components of* $\mathbb{E}(\boldsymbol{V}\boldsymbol{V}')$ *are all finite.*

2.5.1 The Durbin-Levinson Algorithm

Go back to Table of Contents. Please click TOC

Algorithm 2.25. ***The Durbin-Levinson Algorithm:***
The coefficients $\phi_{n1}, ..., \phi_{nn}$ *can be computed recursively from the equations*

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

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

and

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

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

Proof: The definition of ϕ_{11} ensures that the equation

$$R_n\phi_n = \rho_n$$

(where $\rho_n = (\rho(1), ..., \rho(n))')$ is satisfied for $n = 1$. The first step in the proof is to show that ϕ_n, defined recursively by ϕ_{nn} and $\begin{bmatrix}\phi_{n1}\\ \vdots\\ \phi_{n,n-1}\end{bmatrix}$, satisfies $R_n\phi_n = \rho_n$ for $n = k$. Then, partitioning R_{k+1} and defining

$$\rho_k^{(r)} := (\rho(k), \rho(k-1), ..., \rho(1))'$$

and

$$\phi_k^{(r)} := (\phi_{kk}, \phi_{k,k-1}, ..., \phi_{k1})',$$

we see that the recursions imply

$$\begin{aligned} R_{k+1}\phi_{k+1} &= \begin{bmatrix} R_k & \rho_k^{(r)} \\ \rho_k^{(r)\prime} & 1 \end{bmatrix} \begin{bmatrix} \phi_k - \phi_{k+1,k+1}\phi_k^{(r)} \\ \phi_{k+1,k+1} \end{bmatrix} \\ &= \begin{bmatrix} \rho_k - \phi_{k+1,k+1}\rho_k^{(r)} + \phi_{k+1,k+1}\rho_k^{(r)} \\ \rho_k^{(r)\prime}\phi_k - \phi_{k+1,k+1}\rho_k^{(r)\prime}\phi_k^{(r)} + \phi_{k+1,k+1} \end{bmatrix} \\ &= \rho_{k+1}, \end{aligned}$$

as required. Here we have used the fact that if $R_k\phi_k = \rho_k$, then $R_k\phi_k^{(r)} = \rho_k^{(r)}$. This is easily checked by writing out the component equations in reverse order. Since $R_n\phi_n$ is satisfied for $n = 1$, it follows by induction that the coefficient vectors ϕ_n defined recursively by

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

and

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

to satisfy $R_n\phi_n = \rho_n$ for all n.

It remains only to establish that the mean squared errors

$$v := \mathbb{E}(Xn + 1 - \phi_n'\mathbf{X}_n)^2$$

satisfy $v_0 = \gamma(0)$ and $v_n = v_{n-1}[1 - \phi_{nn}^2]$. The fact that $v_0 = \gamma(0)$ is an immediate consequence of the definition $P_0X_1 := \mathbb{E}(X_1) = 0$. Since we have shown that $\phi_n'\mathbf{X}_n$ is the best linear predictor of X_{n+1}, we can writ,e

$$v_n = \gamma(0) - \phi_n'\gamma_n = \gamma(0) - \phi_{n-1}'\gamma_{n-1} + \phi_{nn}\phi_{n-1}^{(r)\prime}\gamma_{n-1} - \phi_{nn}\gamma(n).$$

Applying

$$\begin{aligned} \mathbb{E}(X_{n+h} - P_nX_{n+h})^2 &= \gamma(0) - 2\sum_{i=1}^{n} a_i\gamma(h+i-1) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i\gamma(i-j)a_j \\ &= \gamma(0) - \mathbf{a}_n'\gamma_n(h), \\ &= \gamma(0) - \mathbf{a}_n'\Gamma_n\mathbf{a}_n \end{aligned}$$

again gives us

$$v_n = v_{n-1} + \phi_{nn}\left(\phi_{n-1}^{(r)\prime}\gamma_{n-1} - \gamma(n)\right),$$

and hence, by

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

there is

$$v_n = v_{n-1} - \phi_{nn}^2(\gamma(0) - \phi_{n-1}'\gamma_{n-1}) = v_{n-1}(1 - \phi_{nn}^2).$$

Q.E.D.

Proof (Here is an alternative approach): Consider the Hilbert space $\mathcal{H} = \{X : \mathbb{E}(X^2) < \infty\}$ with inner product $< X, Y >= \mathbb{E}(XY)$ with norm $||X||^2 =< X, X >$. By the definition of $\hat{X}_{n+1}$, we can view $\hat{X}_{n+1}$ is in the linear space of $\mathcal{H}$ spanned by $\{X_n, ..., X_1\}$, denoted by $\overline{sp}\{X_n, ..., X_1\} \doteq \{Y : Y = a_1X_n + \cdots + a_nX_1 \; where \; a_1, ..., a_n \in \mathbb{R}\}$. Since $X_1 - \overline{\mathbb{P}}(X_1|X_n, ..., X_2)$ is orthogonal to $X_n, ..., X_2$; i.e.,

$$< X_1 - \overline{\mathbb{P}}(X_1|X_n, ..., X_2),\ X_k. = 0,\ k = 2, ..., n.$$

We have

$$\begin{aligned}\overline{sp}\{X_n, ..., X_2, X_1\} &= \overline{sp}\{X_n, ..., X_2, X_1 - \overline{\mathbb{P}}(X_1|X_n, ..., X_2)\} \\ &= \overline{sp}\{X_n, ..., X_2\} + \overline{sp}\{X_1 - \overline{\mathbb{P}}(X_1|X_n, ..., X_2)\}.\end{aligned}$$

Thus

$$\hat{X}_{n+1} = \overline{\mathbb{P}}(X_{n+1}|X_n, ..., X_2) + a\{X_1 - \overline{\mathbb{P}}(X_1|X_n, ..., X_2)\},$$

where

$$a = \frac{< X_{n+1}, X_1 - \overline{\mathbb{P}}(X_1|X_n, ..., X_2) >}{||X_1 - \overline{\mathbb{P}}(X_1|X_n, ..., X_2)||^2}.$$

By stationary, we have

$$\overline{\mathbb{P}}(X_1|X_n, ..., X_2) = \sum_{j=1}^{n-1} \phi_{n-1,j} X_{j+1}$$

$$\overline{\mathbb{P}}(X_{n+1}|X_n, ..., X_2) = \sum_{j=1}^{n-1} \phi_{n-1,j} X_{n+1-j}$$

Then from $\hat{X}_{n+1}$, $\overset{\mathbb{P}}{(}X_1|X_n, ..., X_2)$, and $\overline{\mathbb{P}}(X_{n+1}|X_n, ..., X_2)$ we have

$$\hat{X}_{n+1} = aX_1 + \sum_{j=1}^{n-1} (\phi_{n-1,j} - a\phi_{n-1,n-j})X_{n+1-j},$$

where from equation a and $\overline{\mathbb{P}}(X_1|X_n, ..., X_2)$, there is

$$\begin{aligned} a &= \big(< X_{n+1}, X_1 > - \sum_{j=1}^{n-1} \phi_{n-1,j} < X_{n+1}, X_{j+1} > \big) v_{n-1}^{-1} \\ &= \big\{\gamma_X(n) - \sum_{j=1}^{n-1} \phi_{n-1,j}\gamma_X(n-j)\big\} v_{n-1}^{-1}. \end{aligned}$$

When $\gamma_X(h) \to 0$ as $h \to \infty$ guarantees that the representation

$$\hat{X}_{n+1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}$$

is unique. Therefore we deduce

$$\phi_{nn} = a$$

and

$$\phi_{nj} = \phi_{n-1,j} - a\phi_{n-1,n-j}, \;\; j = 1, ..., n-1.$$

Lastly,

$$\begin{array}{rcl} v_n & = & ||X_{n+1} - \hat{X}_{n+1}||^2 \\ & = & ||X_{n+1} - \overline{\mathbb{P}}(X_{n+1}|X_n, ..., X_2) - a\{X_1 - \overline{\mathbb{P}}(X_1|X_n, ..., X_2)\}||^2 \\ & = & ||X_{n+1} - \overline{\mathbb{P}}(X_{n+1}|X_n, ..., X_2)||^2 + a^2||X_1 - \overline{\mathbb{P}}(X_1|X_n, ..., X_2)\}||^2 \\ & & -2a < X_{n+1} - \overline{\mathbb{P}}(X_{n+1}|X_n, ..., X_2), X_1 - \overline{\mathbb{P}}(X_1|X_n, ..., X_2) > \\ & = & v_{n-1} + a^2 < X_{n+1}, X_1 - \overline{\mathbb{P}}(X_1|X_n, ..., X_2) > \end{array}$$

Thus, we conclude

$$v_n = v_{n-1} + a^2 v_{n-1} - 2a^2 v_{n-1} = v_{n-1}(1 - a^2).$$

Q.E.D.

Example 2.26. Consider the following:

$$\begin{array}{rl} \phi_0 = 0, & \phi_{00} = 0; \\ \phi_1 = \phi_{11}, & \phi_{11} = \frac{\gamma(1)}{\gamma(0)}; \\ \phi = \begin{pmatrix} \phi_{n-1} - \phi_{nn}\tilde{\phi}_{n-1} \\ \phi_{nn} \end{pmatrix}, & \phi_{nn} = \frac{\gamma(n) - \phi'_{n-1}\tilde{\gamma}_{n-1}}{\gamma(0) - \phi'_{n-1}\gamma_{n-1}}. \end{array}$$

THis algorithm computes ϕ_1, ϕ_2, ϕ_3, ..., where

$$X_2^1 = X_1\phi_1, \;\; X_3^2 = (X_2, X_1)\phi_2, \;\; X_4^3 = (X_3, X_2, X_1)\phi_3, ...$$

□

Remark 2.27. Durbin-Levinson, as famous as it is, is the evolution of mean square error. One can consider the following

$$\begin{array}{rcl} P_{n+1}^n & = & \gamma(0) - \phi'_n\gamma_n \\ & = & \gamma(0) - \begin{pmatrix} \phi_{n-1} - \phi_{nn}\tilde{\phi}_{n-1} \\ \phi_{nn} \end{pmatrix}' \begin{pmatrix} \gamma_{n-1} \\ \gamma(n) \end{pmatrix} \\ & = & P_n^{n-1} - \phi_{nn}\big(\gamma(n) - \tilde{\phi}'_{n-1}\gamma_{n-1}\big) \\ & = & P_n^{n-1} - \phi_{nn}^2\big(\gamma(0) - \phi'_{n-1}\gamma_{n-1}\big) \\ & = & P_n^{n-1}(1 - \phi_{nn}^2). \end{array}$$

hence, the variance reduces by a factor $1 - \phi_{nn}^2$.

2.5.2 The Innovations Algorithm

Go back to Table of Contents. Please click TOC

Suppose then that $\{X_t\}$ is a zero-mean series with $\mathbb{E}(|X_t|^2) < \infty$ for each t and

$$\mathbb{E}(X_i, X_j) = \kappa(i, j).$$

It will be convenient to introduce

$$\hat{X}_n = \begin{cases} 0, & if \;\; n = 1, \\ P_{n-1}X_n, & if \;\; n = 2, 3, ..., \end{cases}$$

and

$$v_n = \mathbb{E}(X_{n+1} - P_nX_{n+1})^2.$$

We shall also introduce the **innovations**, or one-step prediction errors,

$$U_n = X_n - \hat{X}_n.$$

In terms of the vectors $\mathbf{U}_n = (U_1, ..., U_n)'$ and $\mathbf{X}_n = (X_1, ..., X_n)'$ the last equations can be

$$\mathbf{U}_n = A_n \mathbf{X}_n,$$

where

$$A_n = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ a_{11} & 1 & 0 & \dots & 0 \\ a_{22} & a_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \dots & 1 \end{bmatrix}$$

(If $\{X_t\}$ is stationary, then $a_{ij} = -a_j$ with a_j as in

$$\Gamma_n \mathbf{a}_n = \gamma_n(h),$$

with $h = 1$.) This implies that A_n is nonsingular, with inverse C_n of the form

$$C_n = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \theta_{11} & 1 & 0 & \dots & 0 \\ \theta_{22} & \theta_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \dots & 1 \end{bmatrix}$$

The vector of one-step predictors $\hat{X}_n := (X_1, P_1X_2, ..., P_{n-1}X_n)'$ can therefore be expressed as

$$\hat{\mathbf{X}}_n = \mathbf{X}_n - \mathbf{U}_n = C_n \mathbf{U}_n - \mathbf{U}_n = \Theta_n(\mathbf{X}_n - \hat{\mathbf{X}}_n),$$

where

$$\Theta_n = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \theta_{11} & 1 & 0 & \dots & 0 \\ \theta_{22} & \theta_{21} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \dots & 1 \end{bmatrix}$$

and $\mathbf{X}_n$ itself satisfies

$$\mathbf{X}_n = C_n(\mathbf{X}_n - \hat{\mathbf{X}}_n).$$

Then,

$$\hat{\mathbf{X}}_n = \mathbf{X}_n - \mathbf{U}_n = C_n \mathbf{U}_n - \mathbf{U}_n = \Theta_n(\mathbf{X}_n - \hat{\mathbf{X}}_n),$$

can be rewritten as

$$\hat{X}_{n+1} = \begin{cases} 0, & if\ n = 0, \\ \sum\limits_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), & if\ n = 1, 2, ..., \end{cases}$$

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

Algorithm 2.28. ***The Innovations Algorithm:***
The coefficients $\theta_{n1}, ..., \theta_{nn}$ can be computed recursively from the equations

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

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

and

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

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

Remark 2.29. While the Durbin-Levinson recursion gives the coefficients of $X_n, ..., X_1$ in the representation $\hat{X}_{n+1} = \sum\limits_{j=1}^{n}\phi_{nj}X_{n+1-j}$, the innovations algorithm gives the coefficients of $(X_n - \hat{X}_)n), ..., (X_1 - \hat{X}_1)$, in the expansion $\hat{X}_{n+1} = \sum\limits_{j=1}^{n}\theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j})$. The latter expansion has a number of advantages deriving from the fact that the innovations are uncorrelated. It can also be greatly simplified in the case of ARMA(p,q) series, as we shall see in Section 3.3. An immediate consequence of

$$\hat{X}_{n+1} = \begin{cases} 0, & if\ n = 0, \\ \sum\limits_{j=1}^{n}\theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), & if\ n = 1,2,..., \end{cases}$$

is the innovations representation of X_{n+1} itself. Thus (defining $\theta_{n0} := 1$),

$$X_{n+1} = X_{n+1} - \hat{X}_{n+1} + \hat{X}_{n+1} = \sum_{j=0}^{n}\theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}),\ n = 0,1,2,...$$

Example 2.30. If $\{X_t\}$ is the time series defined by

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

then $\kappa(i,j) = 0$ for $|i-j| > 1$, $\kappa(i,j) = \sigma^2(1+\theta^2)$, and $\kappa(i,i+1) = \theta\sigma^2$. Application of the innovations algorithm leads at once to the recursions

$$\theta_{nj} = 0,\ 2 \le j \le n,$$

$$\theta_{n1} = v_{n-1}^{-1}\theta\sigma^2,$$
$$v_0 = (1+\theta)^2\sigma^2,$$

and

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

□

2.5.3 Prediction of a Stationary Process in Terms of Infinitely Many Past Values

Go back to Table of Contents. Please click TOC

It is often useful, when many past observations $X_m, ..., X_0, X_1, ..., X_n$ $(m < 0)$ are available, to evaluate the best linear predictor of X_{n+h} in terms of $1, X_m, ..., X_0, ..., X_n$. This predictor, which we shall denote by $P_{m,n}X_{n+h}$, can easily be evaulated by the methods described above. If $|m|$ is large, this predictor can be approximated by the sometimes more easily calculated mean square limit

$$\tilde{P}_n X_{n+h} = \lim_{m \to -\infty} P_{m,n} X_{n+h}.$$

We shall refer to $\tilde{P}_n$ as the **prediction operator based on the infinite past,** $\{X_t, -\infty < t \leq n\}$. Analogously we shall refer to P_n as the **prediction operator based on the finite past**, $\{X_t, ..., X_n\}$.

Determination of $\tilde{P}_n X_{n+h}$

Like $P_n X_{n+h}$, the best linear predictor $\tilde{P}_n X_{n+h}$ when $\{X_t\}$ is a zero-mean stationary process with autocovariance function $\gamma(\cdot)$ is characterized by the equations

$$\mathbb{E}[(X_{n+h} - \tilde{P}_n X_{n+h}) X_{n+1-i}] = 0, \ i = 1, 2, ...$$

If we can find a solution to these equations, it will necessarily by the uniquely defined predictor $\tilde{P}_n X_{n+h}$. An approach to this problem that is often effective is to assume that $\tilde{P}_n X_{n+h}$ can be expressed in the form

$$\tilde{P}_n X_{n+h} = \sum_{j=1}^{\infty} \alpha_j X_{n+1-j},$$

in which case the preceding equations reduce to

$$\mathbb{E}\left[\left(X_{n+h} - \sum_{j=1}^{\infty} \alpha_j X_{n+1-j}\right) X_{n+1-i}\right] = 0, \ i = 1, 2, ...,$$

or equivalently,

$$\sum_{j=1}^{\infty} \gamma(i-j)\alpha_j = \gamma(h+i-1), \ i = 1, 2, ...$$

This is an infinite set of linear equations for the unknown coefficients α_i that determine $\tilde{P}_n X_{n+h}$, provided that the resulting series converges.

Proposition 2.31. ***Properties of*** $\tilde{P}_n$***:***
Suppose that $\mathbb{E}(U^2) < \infty$, $\mathbb{E}(V^2) < \infty$, *a, b, and c are constants, and* Γ=*Cov(* $\boldsymbol{W}$, $\boldsymbol{W}$*).*

(1) $\mathbb{E}[(U - \tilde{P}_n(U))X_j] = 0, \ j \leq n.$

(2) $\tilde{P}_n(aU + bV + c) = a\tilde{P}_n(U) + b\tilde{P}_n(V) + c.$

(3) $\tilde{P}_n(U) = U$ *if* U *is a limit of linear combinations of* X_j, $j \leq n$.

(4) $\tilde{P}_n(U) = \mathbb{E}(U)$ *if Cov(U,*X_j*) = 0 for all* $j \leq n$.

Example 2.32. Consider the causal invertible ARMA(1,1) process $\{X_t\}$ defined by

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

We know

$$X_{n+1} = Z_{n+1}(\phi + \theta)\sum_{j=1}^{\infty}\phi^{j-1}Z_{n+1-j}$$

and

$$Z_{n+1} = X_{n+1} - (\phi + \theta)\sum_{j=1}^{\infty}(-\theta)^{j-1}X_{n+1-j}.$$

Applying the operator $\tilde{P}_n$ to the second equation and using the properties of $\tilde{P}_n$ gives

$$\tilde{P}_n X_{n+1} = (\phi + \theta)\sum_{j=1}^{\infty}(-\theta)^{j-1}X_{n+1-j}.$$

Applying the operator $\tilde{P}_n$ to the first equation and using the properties of $\tilde{P}_n$ gives

$$\tilde{P}_n X_{n+1} = (\phi + \theta)\sum_{j=1}^{\infty}\phi^{j-1}Z_{n+1-j}.$$

Hence,

$$X_{n+1} - \tilde{P}_n X_{n+1} = Z_{n+1},$$

and so the mean squared error of the predictor $\tilde{P}_n X_{n+1}$ is $\mathbb{E}(Z_{n+1}^2) = \sigma^2$.

□

2.6 The Wold Decomposition

Go back to Table of Contents. Please click TOC

Consider the stationary process

$$X_t = A\cos(\omega t) + B\sin(\omega t),$$

where $\omega \in (0, \pi)$ is constant and A, B are uncorrelated random variables with mean 0 and variance σ^2. Notice that

$$X_n = (2\cos\omega)X_{n-1} - X_{n-2} = \tilde{P}_{n-1}X_n,\ n = 0, \pm 1, ...,$$

so that $X_n - \tilde{P}_{n-1}X_n = 0$ for all n. Processes with latter property are said to be **deterministic**.

The Wold Decomposition:

If $\{X_t\}$ is a nondeterministic stationary time series, then

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

where

(1) $\psi_0 = 1$ and $\sum_{j=0}^{\infty}\psi_j^2 < \infty$,

(2) $\{Z_t\} \sim WN(0, \sigma^2)$,

(3) $\text{Cov}(Z_s, V_t) = 0$ for all s and t,

(4) $Z_t = \tilde{P}_t Z_t$ for all t,

(5) $V_t = \tilde{P}_s V_t$ for all s and t, and

(6) $\{V_t\}$ is deterministic.

We also present the following alternative approach to describe *Wold Decomposition.*

Theorem 2.33. *Suppose that $\{x_t\}$ is a covariance stationary process with $\mathbb{E}(x_t) = 0$ and covariance function, $\gamma(j) = \mathbb{E}(x_t x_{t-j})$, $\forall j$. Then*

$$x_t = \sum_{j=0}^{\infty} d_j \epsilon_{t-j} + \eta_t$$

where

$$\begin{aligned} d_0 &= 1,\ \sum_{j=0}^{\infty} d_j^2 < 0,\ \mathbb{E}(\epsilon_t^2) = \sigma_\epsilon^2,\ \mathbb{E}(\epsilon_t \epsilon_s) = 0\ for\ t \neq s, \\ \mathbb{E}(\epsilon_t) &= 0,\ \mathbb{E}(\eta_t \epsilon_s = 0\ \forall t, s, \\ P[\eta_{t+s} | x_{t-1}, x_{t-2}, ...] &= \eta_{t+s},\ s \geq 0. \end{aligned}$$

The first part of the representation of x_t looks like the MA(∞) with square summable moving average terms that we have worked with, while the second part, η_t, is something new. That part is called the deterministic part of x_t because η_t is perfectly predictable based on past observations on x_t.

The style of proof is constructive. We will show that given only covariance stationarity, we can build the Wold representation with the indicated properties. We will not provide a fully rigorous proof and a key result will simply be assumed. The proof is an application of linear projections, and the orthogonality and recursive properties of projections. The proof follows that in Sargent (1979) [30].

Proof:

We first find the d_j's and ϵ_t and establish the required properties. Then, we find the projection error, η_t.

We begin with a preliminary result. Let x_t be a covariance stationary process. Let

$$\hat{x}_t^{(n)} = P[x_t | x_{t-1}, ..., x_{t-n}],$$

and write

$$x_t = \hat{x}_t^{(n)} + \epsilon_t^{(n)}.$$

From the orthogonality property of projections we know that

$$\begin{aligned} & \epsilon_t^{(n)} \perp (x_{t-1}, ..., x_{t-n}) \\ \mathbb{E}(\epsilon_t^{(n)} &= \sigma^{2(n)}. \end{aligned}$$

We assume, without proof, the following result:

$$\begin{aligned} \hat{x}_t^{(n)} &\rightarrow \hat{x} = P[x_t | x_{t-1}, x_{t-2}, ...] \\ x_t &= \hat{x}_t + \epsilon_t,\ \mathbb{E}(\epsilon_t^2) = \sigma^2 \\ & \epsilon_t \perp (x_{t-1}, x_{t-2}, ...). \end{aligned}$$

The disturbance, ϵ_t, is known as the "innovation" in x_t or its "one-step-ahead forecast error". It is easy to see that ϵ_t is a serially uncorrelated process. In particular,

$$\epsilon_t = x_t - P[x_t|x_{t-1}, x_{t-2}, ...],$$

so that it is a linear combination of current and past x_t's. It follows that since ϵ_t is orthogonal to past x_t's, it is also orthogonal to past ϵ_t's.

Q.E.D.

The Wold representation is the **unique** linear representation where the innovations are linear forecast errors.

Remark 2.34. We have the following definition as premise.

Definition 2.35. A zero-mean nondeterministic covariance-stationary process, $\{x_t; t \in \mathbb{Z}\}$, is called purely nondeterministic (or regular) if $d_t = 0$.

Remark 2.36. The Wold Theorem plays a central role in time series analysis. It implies that the dynamic of any purely nondeterministic covariance-stationary process can be arbitrarily well approximated by an ARMA process.

By Wold's Decomposition Theorem, we have that any purely nondeterministic covariance-stationary process can be written as a linear combination of lagged values of a white noise process (MA(∞)) representation, that is $X_t = \sum\limits_{j=0}^{\infty} \psi_j u_{t-j}$. Now, we note that, under general conditions, the infinite lag polynomial of the Wold decomposition can be approximated by the ratio of two finite-lag polynomials:

$$\Psi(L) \approx \frac{\theta(L)}{\phi(L)}.$$

Therefore x_t can be accurately approximated by a ARMA process

$$x_t^* = \frac{\theta(L)}{\phi(L)} u_t.$$

Any purely nondeterministic covariance-stationary process has an ARMA representation. This means that the stationary ARMA(p, q) models are a class of linear stochastic processes that are general enough.

Example 2.37. Are the covariance-stationary ARMA processes purely nondeterministic processes? Consider a covariance-stationary ARMA(p, q) process defined by

$$\phi(L)x_t = \theta(L)u_t, \ u_t \sim WN(0, \sigma^2)$$

where

$$\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$$

and

$$\theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$$

Suppose that this representation is causal and invertible. The causality assumption implies that there exists constants $\psi_0, \psi_1, ...$ such that

$$\sum_{j=0}^{\infty} |\psi_j| < \infty, \ with \ \psi_0 = 1$$

and

$$x_t = \sum_{j=0}^{\infty} \psi_j u_{t-j}, \ \forall t.$$

$$\sum_{j=0}^{\infty} |\psi_j| < \infty \Rightarrow \sum_{j=1}^{\infty} \psi_j^2 < \infty$$

$u_t \sim WN(0, \sigma_u^2)$ and the invertibility condition implies that u_t is the limit of linear combinations of $x_s,\ s \leq t$. We can conclude that the covariance-stationary ARMA(p, q) process, x_t, is a purely nondeterministic process.

Now, suppose the representation

$$\phi(L)x_t = \theta(L)u_t, \ u_t \sim WN(0, \sigma_u^2)$$

where

$$\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$$

and

$$\theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$$

is not causal and not invertible. It is possible to show that if $\theta(z) \neq 0$ when $|z| = 1$, then it is always possible to find polynomials $\phi^*(L)$ and $\theta^*(L)$ and a white noise $v_t \sim WN(0, \sigma_v^2)$ such that the representation

$$\phi^*(L)x_t = \theta^*(L)v_t, \ v_t \sim WN(0, \sigma_v^2)$$

is causal and invertible.

Thus a covariance-stationary ARMA(p, q) process defined by

$$\phi(L)x_t = \theta(L)u_t, \ u_t \sim WN(0, \sigma^2)$$

where

$$\phi(L) = 1 - \phi_q L - \cdots - \phi_p L^p$$

and

$$\theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$$

with $\theta(z) \neq 0$ if $|z| = 1$ is a purely nondeterministic process. A covariance-stationary ARMA process, with $\theta(z) \neq 0$ if $|z| = 1$, is a purely nondeterministic process.

3 ARMA Models

Go back to Table of Contents. Please click TOC

In this chapter we introduce an important parametric family of stationary time series, the autoregressive moving-average, or ARMA, processes. For a large class of autocovariance functions $\gamma(\cdot)$ it is possible to find an ARMA process $\{X_t\}$ with ACVF $\gamma_X(\cdot)$ such that $\gamma(\cdot)$ is well approximated by $\gamma_X(\cdot)$. In particular, for any positive integer K, there exists an ARMA process $\{X_t\}$ such that $\gamma_X(h) = \gamma(h)$ for $h = 0, 1, ..., K$.

3.1 ARMA(p,q) Processes

Go back to Table of Contents. Please click TOC

We introduced an ARMA(1,1) process and discussed some of its key properties. We extend these notions to the general ARMA(p, q) process.

Definition 3.1. $\{X_t\}$ is an **ARMA(p,q) process** if $\{X_t\}$ is stationary and if for every t,

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

where $\{Z_t\} \sim WN(0, \sigma^2)$ and the polynomials $(1 - \phi_1 z - \cdots - \phi_p Z^p)$ and $(1 + \theta_1 z + \cdots + \theta_q z^q)$ have no common factors.

The process $\{X_t\}$ is said to be an **ARMA(p,q) process with mean** μ if $\{X_t - \mu\}$ is an ARMA(p,q) process.

It is convenient to use the more concise form of $X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$, which is

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

where $\phi(\cdot)$ and $\theta(\cdot)$ are the pth and qth-degree polynomials

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$

and

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q,$$

and B is the backward shift operator ($B^j X_t = X_{t-j}$, $B^j X_t = Z_{t-j}$, $j = 0, \pm 1, ...$). The time series $\{X_t\}$ is said to be an **autoagressive process of order** p (or AR(p)) if $\theta(z) \equiv 1$, and a **moving-average process of order** q (or MA(q)) if $\phi(z) \equiv 1$.

Theorem 3.2. ***Existence and Uniqueness:***
A stationary solution $\{X_t\}$ of equations

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

exists (and is also the unique stationary solution) if and only if

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \ for \ all \ |z| = 1.$$

Definition 3.3. Causality
An ARMA(p, q) process $\{X_t\}$ is **causal**, or a **causal function** of $\{Z_t\}$, if there exist constants $\{\psi_j\}$ such that $\sum\limits_{j=0}^{\infty} \psi_j < \infty$ and

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

Causality is equivalent to the condition

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \ \forall |z| \leq 1.$$

The sequence $\{\psi_j\} = \sum\limits_{j=0}^{\infty} \psi_j Z_{t-j}$ for all t is determined by the relation $\psi(z) = \sum\limits_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$, or equivalently by the identity

$$(1 - \phi_1 z - \cdots - \phi_p z^p)(\psi_0 + \psi_1 z + \dots) = 1 + \theta_1 z + \cdots + \theta_q z^q.$$

Equating coefficients of z^j, $j = 0, 1, ...$, we find that

$$\begin{array}{rcl} 1 & = & \psi_0, \\ \theta_1 & = & \psi_1 - \psi_0 \phi_1, \\ \theta_2 & = & \psi_2 - \psi_1 \phi_1 - \psi_0 \phi_2 \\ \vdots & & \end{array}$$

or equivalently,

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = \theta_j, \ j = 0, 1, 2, ...,$$

where $\theta_0 := 1$, $\theta_j := 0$ for $j > q$, and $\psi_j := 0$ for $j < 0$.

Definition 3.4. Invertibility:

An ARMA(p,q) process $\{X_t\}$ is **invertible** if there exist constants $\{\pi_j\}$ such that $\sum\limits_{j=0}^{\infty} |\pi_j| < \infty$ and

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \ \forall t.$$

Invertibility is equivalent to the condition

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \neq 0 \ \forall |z| \leq 1.$$

Example 3.5. Consider the ARMA(1,1) process $\{X_t\}$ satisfying the equations

$$X_t - .5X_{t-1} = Z_t + .4Z_{t-1}, \ \{Z_t\} \sim WN(0, \sigma^2).$$

Since the autoregressive polynomial $\phi(z) = 1 - .5z$ has a zero at $z = 2$, which is located outside the unit circle, we conclude that there exists a unique ARMA process satisfying the equation above that is also causal. The coefficients $\{\psi_j\}$ in the MA(∞) representation of $\{X_t\}$ are found directly from

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = \theta_j, \ j = 0, 1, 2, ...$$

that

$$\begin{array}{rcl} \psi_0 & = & 1, \\ \psi_1 & = & .4 + .5, \\ \psi_2 & = & .5(.4 + .5), \\ \psi_j & = & ,5^{j-1}(.4 + .5), \ j = 1, 2, ... \end{array}$$

The MA polynomial $\theta(z) = 1 + .4z$ has a zero at $z = -1/.4 = -2.5$, which is also located outside the unit circle. This implies that $\{X_t\}$ is invertible with coefficients $\{\pi_j\}$ given by $\pi_j + \sum\limits_{k=1}^{q} \theta_k \pi_{j-k}$, $j = 0, 1, 2, ...$, that

$$\begin{aligned}
\pi_0 &= 1, \\
\pi_1 &= -(.4 + .5), \\
\pi_2 &= -(.4 + .5)(-.4), \\
\pi_j &= -(.4 + .5)(-.4)^{j-1},\ j = 1, 2, ...
\end{aligned}$$

(A direct derivation of these formulas for $\{\psi_j\}$ and $\{\pi_j\}$ was given in Section 2.3 without appealing to the recursions of

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = \theta_j,\ j = 0, 1, 2, ...,$$

and

$$\pi_j + \sum_{k=1}^{q} \theta_k \pi_{j-k} = -\phi_j,\ j = 0, 1, ...,$$

□

Example 3.6. Consider the ARMA(2,1) process defined by the equations

$$X_t - .75X_{t-1} + .5625X_{t-2} = Z_t + 1.25Zt - 1,\ \{Z_t\} \sim WN(0, \sigma^2).$$

The AR polynomial $\phi(z) = 1 - .75z + .5625z^2$ has zeros at $z = 2(1 \pm i\sqrt{3})/3$, which lie outside the unit circle. The process is therefore causal. On the other hand, the MA polynomial $\theta(z) = 1 + 1.25z$ has a zero at $z = -.8$, and hence $\{X_t\}$ is not invertible.

□

Example 3.7. Let $\{X_t\}$ be the AR(2) process

$$X_t = 0.7X_{t-1} - .1X_{t-2} + Z_t,\ \{Z_t\} \sim WN(0, \sigma^2).$$

The auto regressive polynomial for this process has the factorization $\phi(z) = 1 - .7z + .1z^2 = (1 - .5z)(1 - .2z)$, and is therefore zero at $z = 2$ and $z = 5$. Since these zeros lie outside the unit circle, we conclude that $\{X_t\}$ is a causal AR(2) process with coefficients $\{\psi_j\}$ given by

$$\begin{aligned}
\pi_0 &= 1, \\
\pi_1 &= .7, \\
\pi_2 &= .7^2 - .1, \\
\pi_j &= .7\psi_{j-1} - .1\psi_{j-2},\ j = 2, 3, ...
\end{aligned}$$

While it is a simple matter to calculate ψ_j numerically for any j, it is possible also to give an explicit solution of these difference equations using the theory of linear difference equations.

3.2 The ACF and PACF of an ARMA(p,q) Processes

Go back to Table of Contents. Please click TOC

In this section, we discuss three methods for computing the autocovariance function $\gamma(\cdot)$ of a causal ARMA process $\{X_t\}$. The autocorrelation function is readily found from the ACVF on dividing by $\gamma(0)$. The partial autocorrelation function (PACF) is also found from the function $\gamma(\cdot)$.

3.2.1 Calculation of the ACVF

Go back to Table of Contents. Please click TOC

First we determine the ACVF $\gamma(\cdot)$ of the causal ARMA(p,q) process defined by

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

where $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$. The causality assumption implies that

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

where $\sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$, $|z| \leq 1$. The calculation of the sequence $\{\psi_j\}$ was discussed earlier.

First Method

From proposition (see remark below) and the representation $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, we obtain

$$\gamma(h) = \mathbb{E}(X_{t+h}X_t) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}.$$

Remark 3.8. Recall Proposition earlier, if $\{X_t\}$ is a staionary q-correlated time series with mean 0, then it can be represented as the MA(q) process in $X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$, where $\{Z_t\} \sim WN(0, \sigma^2)$ and $\theta_1, ..., \theta_q$ are constants.

Example 3.9. *The ARMA(1,1) process*

Substituting from $X_t = Z_t + (\phi + \theta) \sum_{j=1}^{\infty} \phi^{j-1} Z_{t-j}$ into $\gamma(h) = \mathbb{E}(X_{t+h}X_t) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}$, we find that the ACVF of the process defined by

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

with $|\phi| < 1$ is given by

$$\begin{aligned}\gamma(0) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 \\ &= \sigma^2 \left[1 + (\theta+\phi)^2 \sum_{j=0}^{\infty} \phi^{2j} \right] \\ &= \sigma^2 \left[1 + \tfrac{(\theta+\phi)^2}{1-\phi^2} \right], \\ \gamma(1) &= \sigma^2 \sum_{j=0}^{\infty} \psi_{j+1}\psi_j \\ &= \sigma^2 \left[\theta + \phi + (\theta+\phi)^2 \phi \sum_{j=0}^{\infty} \phi^{2j} \right] \\ &= \sigma^2 \left[\theta + \phi + \tfrac{(\theta+\phi)^2\phi}{1-\phi^2} \right],\end{aligned}$$

and

$$\gamma(h) = \phi^{h-1}\gamma(1), \; h \geq 2.$$

□

Example 3.10. *The MA(q) process*

For the process

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \; \{Z_t\} \sim WN(0, \sigma^2),$$

equation $\gamma(h) = \mathbb{E}(X_{t+h}X_t) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}$ immediately gives the result

$$\gamma(h) = \begin{cases} \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|}, & if \ |h| \leq q, \\ 0, & if \ |h| > q, \end{cases}$$

where θ_0 is defined to be 1. The ACVF of the MA(q) process thus has the distinctive feature of vanishing at lags greater than q. Data for which the sample ACVF is small for lags greater than q therefore suggest that an appropriate model might be a moving average of order q (or less). Recall from proposition (see Remark 3.8 above) that every zero-mean stationary process with correlations vanishing at lags greater than q can be represented as a moving-average process of order q or less.

□

Second Method

If we multiply each side of the equations

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

by X_{t-k}, $k = 0, 1, 2, ...$, and take expectations on each side, we find that

$$\gamma(k) - \phi_1 \gamma(k-1) - \cdots - \phi_p \gamma(k-p) = \sigma^2 \sum_{j=0}^{\infty} \theta_{k+j} \psi_j, \; 0 \leq k < m,$$

and

$$\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = 0,\ k \geq m,$$

where $m = \max(p, q+1)$, $\psi_j := 0$ for $j < 0$, $\theta_0 := 1$, and $\theta_j := 0$ for $j \notin \{0, ..., q\}$. In calculating $\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = 0,\ k \geq m$ we have made use of the expansion $X_t = \sum\limits_{j=0}^{\infty}\psi_j Z_{t-j}$. Equations $\gamma(k) - \phi_1\gamma(l-1) - \cdots - \phi_p\gamma(l-p) = 0,\ k \geq m$ are a set of homogeneous linear difference equations with constant coefficients, for which the solution is well known to be of the form

$$\gamma(h) = \alpha_1\xi_1^{-h} + \alpha_2\xi_2^{-h} + \cdots + \alpha_p\xi_p^{-h},\ h \geq m-p,$$

where $\xi_1, ..., \xi_p$ are the roots (assumed to be distinct) of the equation $\phi(z) = 0$, and $\alpha_1, ..., \alpha_p$ are arbitrary constants. We are looking for the solution of $\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = 0,\ k \geq m$ that also satisfies $\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = \sigma^2\sum\limits_{j=0}^{\infty}\theta_{k+j}\psi_j,\ 0 \leq k < m$. We therefore substitute the solution $\gamma(h) = \alpha_1\xi_1^{-h} + \alpha_2\xi_2^{-h} + \cdots + \alpha_p\xi_p^{-h},\ h \geq m-p$ into $\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = \sigma^2\sum\limits_{j=0}^{\infty}\theta_{k+j}\psi_j,\ 0 \leq k < m$ to obtain a set of m linear equations that then uniquely determine the constants $\alpha_1, ..., \alpha_p$ and the $m-p$ autocovariances $\gamma(h)$, $0 \leq h < m-p$.

Example 3.11. *The ARMA(1,1) process*

For the causal ARMA(1,1) process defined

$$\gamma(0) - \phi\gamma(1) = \sigma^2(1 + \theta(\theta + \phi))$$

and

$$\gamma(1) - \phi\gamma(0) = \sigma^2\theta.$$

Equation $\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = 0,\ k \geq m$ takes the form

$$\gamma(k) - \phi\gamma(k-1) = 0, k \geq 2.$$

The solution is

$$\gamma(h) = \alpha\phi^h,\ h \geq 1.$$

Substituting this expression for $\gamma(h)$ into two preceding equations, $\gamma(0) - \phi\gamma(1) = \sigma^2(1+\theta(\theta+\phi))$ and $\gamma(1) - \phi\gamma(0) = \sigma^2\theta$, gives two linear equations for α and the unknown autocovariance $\gamma(0)$. These equations are esaily solved, giving the autocovariances.

□

Example 3.12. **The general AR(2) process**

For the causal AR(2) process defined by

$$(1 - \xi_1^{-1}B)(1 - \xi_2^{-1}B)X_t = Z_t,\ |\xi_1|, |\xi_2| > 1,\ \xi_1 \neq \xi_2,$$

we esaily find from $\gamma(h) = \alpha_1\xi_1^{-h} + \alpha_2\xi_2^{-h} + \cdots + \alpha_p\xi_p^{-h},\ h \geq m-p$ and $\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = \sigma^2\sum\limits_{j=0}^{\infty}\theta_{k+j}\psi_j,\ 0 \leq k < m$ (from *Second Method*) using the relatins

$$\phi_1 = \xi_1^{-1} + \xi_2^{-1}$$

and

$$\phi_2 = -\xi_1^{-1}\xi_2^{-1}$$

that

$$\gamma(h) = \frac{\sigma^2\xi_1^2\xi_2^2}{(\xi_1\xi_2 - 1)(\xi_2 - \xi_1)}[(\xi_1^2 - 1)^{-1}\xi_1^{1-h} - (\xi_2^2 - 1)^{-1}\xi_2^{1-h}].$$

Figures 3.1-3.4 illustrate some of the possible forms of $\gamma(\cdot)$ for different values of ξ_1 and ξ_2. Notice that in the case of complex conjugate roots $\xi_1 = re^{i\theta}$ and $\xi_2 = re^{-i\theta}$, $0 < \theta < \pi$, we can write $\gamma(h) = \frac{\sigma^2\xi_1^2\xi_2^2}{(\xi_1\xi_2-1)(\xi_2-\xi_1)}[(\xi_1^2 - 1)^{-1}\xi_1^{1-h} - (\xi_2^2 - 1)^{-1}\xi_2^{1-h}]$ in the more illuminating from

$$\gamma(h) = \frac{\sigma^2 r^4 \cdot r^{-h}\sin(h\theta + \psi)}{(r^2 - 1)(r^4 - 2r^2\cos 2\theta + 1)\sin\theta},$$

where

$$\tan\psi = \frac{r^2 + 1}{r^2 - 1}\tan\theta$$

and $\cos\psi$ has the same sign as $\cos\theta$. Thus in this case $\gamma(\cdot)$ has the form of a damped sinusoidal function with damping factor r^{-1} and period $2\pi/\theta$. If the roots are close to the unit circle, then r is close to 1, the damping is slow, and we obtain a nearly sinusoidal autocovariance function.

Figure 15: The model ACF of the AR(2) series of Example 3.2.4 from text [9] with $\xi_1 = 2$ and $\xi_2 = 5$..

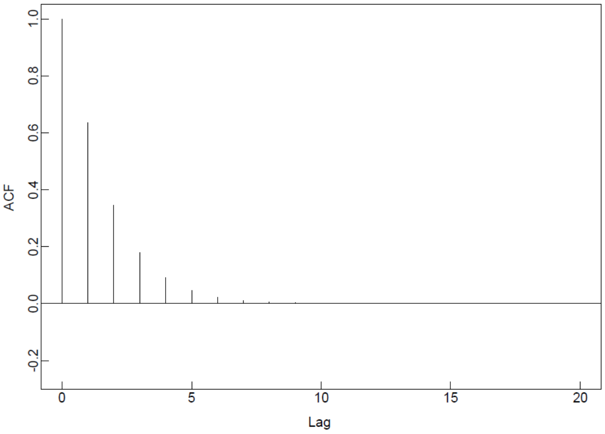


Figure 16: The model ACF of the AR(2) series of Example 3.2.4 from text [9] with $\xi_1 = 10/9$ and $\xi_2 = 2$..

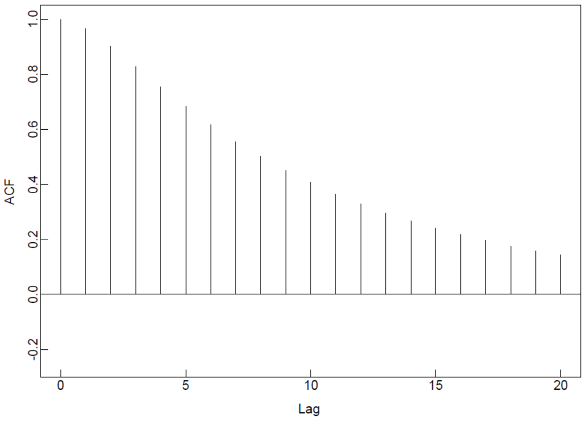


Figure 17: The model ACF of the AR(2) series of Example 3.2.4 from text [9] with $\xi_1 = -10/9$ and $\xi_2 = 2$..

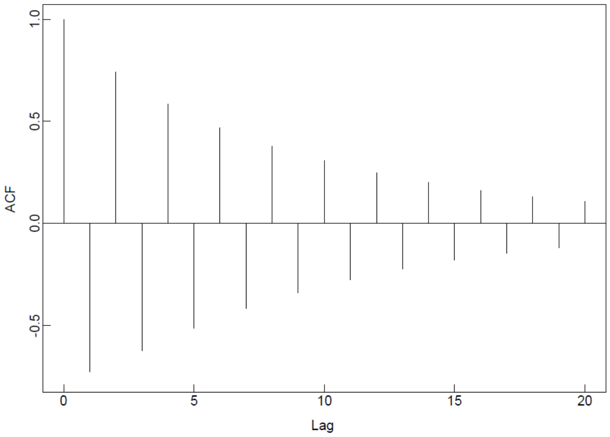
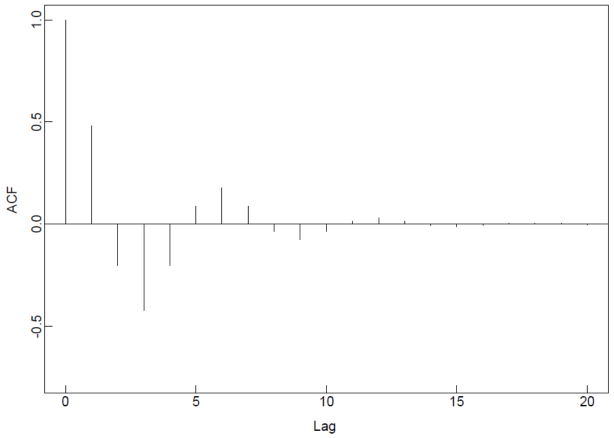


Figure 18: The model ACF of the AR(2) series of Example 3.2.4 from text [9] with $\xi_1 = 2(1+i\sqrt{3})/3$ and $\xi_2 = 2(1-i\sqrt{3})/3$.



□

3.2.2 The Autocorrelation Function

Go back to Table of Contents. Please click TOC

Recall that the ACF of an ARMA process $\{X_t\}$ is the function $\rho(\cdot)$found immediately from the ACVF $\gamma(\cdot)$ as

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

Likewise, for any set of observations $\{x_1, ..., x_n\}$, the sample ACF $\hat{\rho}(\cdot)$ is computed as

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

The Sample ACF of an MA(q) Series.

Given observations $\{x_1, ..., x_n\}$ of a time series, one approach to the fitting of a model to the data is to match the sample ACF of the data with the ACF of the model. In particular, if the sample ACF$\hat{\rho}(h)$ is significantly different from $0 \leq h \leq q$ and negligible for $h > q$, Example 3.2.2 from text [9] suggests that an MA(q) model might provide a good representation of the data. In order to apply this criterion we need to take into account the random variation expected in the sample autocorrelation function before we can classify ACF values as "negligible". To resolve this problem we can use bartlett's formula, which implies that for a large sample of size n from an MA(q) process, the sample ACF values at lags greater than q are approximately normally distributed with means 0 and variances $w_{hh}/n = (1 + 2\rho^2(1) + \cdots + 2\rho^2(q))/n$. This means that if the sample is from an MA(q) process and if $h > q$, then $\hat{\rho}(h)$ should fall between the bounds $\pm 1.96\sqrt{w_{hh}/n}$ with probability approximately 0.95. In practice we frequently use the more stringent values $\pm 1.96/\sqrt{n}$ as the bounds between which sample autocovariances are considered "negligible". A more effective and systematic approach to the problem of model selection, which also applies to ARMA(p,q) models with $p > 0$ and $q > 0$, will be discussed in Section 5.5.

3.2.3 The Partial Autocorrelation Function

Go back to Table of Contents. Please click TOC

The **partial autocorrelation function (PACF)** of an ARMA process $\{X_t\}$ is the function $\alpha(\cdot)$ defined by the equations

$$\alpha(0) = 1$$

and

$$\alpha(h) = \phi_{hh},\ h \geq 1,$$

where ϕ_{hh} is the last component of

$$\phi_h = \Gamma_h^{-1}\gamma_h,$$

$\Gamma_h = [\gamma(i-j)]_{i,j=1}^{h}$, and $\gamma_h = [\gamma(1), \gamma(2), ..., \gamma(h)]'$.

For any set of observations $x_1, ..., x_n\}$ with $x_i \neq x_j$ for some i and j, the **sample PACF** $\hat{\alpha}(h)$ is given by

$$\hat{\alpha}(0) = 1$$

and

$$\hat{\alpha}(h) = \hat{\phi}_{hh},\ h \geq 1,$$

where $\hat{\phi}_h = \hat{\Gamma}_h^{-1}\hat{\gamma}_h$.

We show in the next example that the PACF of a causal AR(p) process is zero for lags greater than p. Both sample and model partial autocorrelation functions can be computed numerically using programs. Algebraic calculation of the PACF is quite complicated except when q is zero or p and q are both small.

Example 3.13. For the causal AR(p) process defined by

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p},\ \{Z_t\} \sim WN(0, \sigma^2),$$

we know that for $h \geq p$ the best linear predictor of X_{h+1} in terms of $1, X_1, ..., X_h$ is

$$\hat{X}_{h+1} = \phi_1 X_h + \phi_2 X_{h-1} + \cdots + \phi_p X_{h+1-p}.$$

Since the coefficient ϕ_{hh} of X_1 is ϕ_p if $h = p$ and 0 and if $h > p$, we conclude that the PACF $\alpha(\cdot)$ of the process $\{X_t\}$ has the properties

$$\alpha(p) = \phi_p$$

and

$$\alpha(h) = 0\ for\ h > p.$$

For $h < p$ the values of $\alpha(h)$ can easily be computed from $\phi_h = \Gamma_h^{-1}\gamma_h$.

3.3 Forecasting ARMA Processes

Go back to Table of Contents. Please click TOC

The innovations provided us with a recursive method for forecasting second-order zero-mean processes that are not necessarily stationary. For the causal ARMA process

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

it is possible to simplify the application of the algorithm drastically. The idea is to apply it not to the process $\{X_t\}$ itself, but to the transformed process (cf. Ansley, 1979) [4]

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

where

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

For notational convenience we define $\theta_0 := 1$ and $\theta_j := 0$ for $j > q$. We shall also assume that $p \geq 1$ and $q \geq 1$. (There is no loss of generality in these assumptions, since in the analysis that follows we may take any of the coefficients ϕ_i and θ_i to be zero.)

The autocovariance function $\gamma_X(\cdot)$ of $\{X_t\}$ can easily be computed using any of the methods described in Section 3.2.1. The autocovariances $\kappa(i, j) = \mathbb{E}(W_iW_j)$, $i, j \geq 1$, are then found from

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

Applying the innovations algorithm to the process $\{X_t\}$ we obtain

$$\begin{cases} \hat{W}_{n+1} = \sum\limits_{j=1}^{n}\theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}), & 1 \leq n < m, \\ \hat{W}_{n+1} = \sum\limits_{j=1}^{q}\theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}), & n \geq m, \end{cases}$$

where the coefficients θ_{nj} and the mean squared errors $r_n = \mathbb{E}(W_{n+1} - \hat{W}_{n+1})^2$ are found recursively from the innovations algorithm with κ defined as in $\kappa(i, j)$. The notable feature of the predictors in the equation above is the vanishing of θ_{nj} when both $n \geq m$ and $j > q$. This is a consequence of the innovations algorithm and the fact that $\kappa(r, s) = 0$ if $r > m$ and $|r - s| > q$.

Observe now that the group of equations allow each X_n, $n \geq 1$, to be written as a linear combination of W_j, $1 \leq j \leq n$, and, conversely, each W_n, $n \geq 1$, to be written as a linear combination of X_j, $1 \leq j \leq n$. This means that the best linear predictor of any random variable Y in terms of $\{1, X_1, ..., X_n\}$. We shall denote this predictor by P_nY. In particular, the one-step predictors of W_{n+1} and X_{n+1} are given by

$$\hat{W}_{n+1} = P_nW_{n+1}$$

and

$$\hat{X}_{n+1} = P_n X_{n+1}.$$

Using the linearity of P_n and equations we see that

$$\begin{cases} \hat{W}_t = \sigma^{-1}\hat{X}_t, & t = 1, ..., m, \\ \hat{W}_t = \sigma^{-1}[\hat{X}_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}], & t > m, \end{cases}$$

which, together with

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

shows that

$$X_t - \hat{X}_t = \sigma[W_t - \hat{W}_t], \ \forall t \geq 1.$$

Replacing $(W_j - \hat{W}_j)$ by $\sigma^{-1}(X_j - \hat{X}_j)$ in $\kappa(i,j)$ and then substituting into

$$\begin{cases} \hat{W}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}), & 1 \leq n < m, \\ \hat{W}_{n+1} = \sum_{j=1}^{q} \theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}), & n \geq m, \end{cases}$$

we finally obtain

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

and

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

where θ_{nj} and r_n are found from the innovations algorithm with κ as in $\kappa(i,j)$. Equations $\hat{X}_{n+1}$ determine the one-step predictors $\hat{X}_2, \hat{X}_3, ...$ recursively.

Example 3.14. Applying

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

to the ARMA(p,1) process with $\phi_1 = 0$, we easily find that

$$\hat{X}_{n+1} = \phi_1 X_n + \cdots + \phi_p X_{n+1-p}, \ n \geq p.$$

□

Example 3.15. Applying

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

to the ARMA(1,q) process with $\phi_1 = 0$ gives

$$\hat{X}_{n+1} = \sum_{j=1}^{\min(n,q)} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}, \ n \geq 1,$$

where the coefficients θ_{nj} are found by applying the innovations algorithm to the covariances $\kappa(i,j)$ defined before. Since in this case the processes $\{X_t\}$ and $\{\sigma^{-1}W_t\}$ are identical, these covariances are simply

$$\kappa(i,j) = \sigma^{-2}\gamma_X(i-j) = \sum_{r=0}^{q-|i-j|} \theta_r\theta_{r+|i-j|}.$$

□

Example 3.16. If

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

and $|\phi| < 1$, then equations

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

reduce to the single equation

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

To compute θ_{n1} we first use Example 3.14 to find $\gamma_X(0) = \sigma^2(1 + 2\theta\phi + \theta^2)/(1-\phi^2)$. Substituting in $\kappa(i,j)$ then gives, for $i, j \geq 1$,

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

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

$$\begin{aligned} r_0 &= (1 + 2\theta\phi + \theta^2)/(1-\phi^2), \\ \theta_{n1} &= \theta/r_{n-1}, \\ r_n &= 1 + \theta^2 - \theta^2/r_{n_1}, \end{aligned}$$

which can be solved explicitly.

□

4 Spectral Analysis

Go back to Table of Contents. Please click TOC

The spectral representation of a stationary time series $\{X_t\}$ essentially decomposes $\{X_t\}$ into a sum of sinusoidal components with uncorrelated random coefficients. In conjunction with this decomposition there is a corresponding decomposition into sinusoids of the autocovariance function of $\{X_t\}$. The spectral decomposition is thus an analogue for stationary processes of the more familiar Fourier representation of deterministic functions. The analysis of stationary processes by means of their spectral representation is often referred to as the "frequency domain analysis" of time series or "spectral analysis". It is equivalent to "time domain" analysis based on the autocovariance function, but provides an alternative way of viewing the process, which for some applications may be more illuminating.

4.1 Spectral Densities

Go back to Table of Contents. Please click TOC

Suppose that $\{X_t\}$ is a zero-mean stationary time series with autocovariance function $\gamma(\cdot)$ satisfying $\sum\limits_{h=-\infty}^{\infty} |\gamma(h)| < \infty$. The **spectral density** of $\{X_t\}$ is the function $f(\cdot)$ defined by

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda}\gamma(h),\ -\infty < \lambda < \infty,$$

where $e^{i\lambda} = \cos(\lambda) + i\sin(\lambda)$ and $i = \sqrt{-1}$. The summability of $|\gamma(\cdot)|$ implies that series of $f(\lambda)$ defined above converges absolutely (since $|e^{ih\lambda}|^2 = \cos^2(h\lambda) + \sin^2(h\lambda) = 1$). Since cos and sin have period 2π, so also does f, and it suffices to confine attention to the values of f, on the interval $(-\pi, \pi]$.

Proposition 4.1. ***Basic Properties of*** f***:***

(a) *f is even, i.e., $f(\lambda) = f(-\lambda)$,*

(b) *$f(\lambda) \geq 0$ for all $\lambda \in (-\pi, \pi]$,*

(c) *$\gamma(k) = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda = \int_{-\pi}^{\pi} \cos(k\lambda) f(\lambda) d\lambda$.*

Proof: Since $\sin(\cdot)$ is an odd function and $\cos(\cdot)$ and $\gamma(\cdot)$ are even functions, we have

$$\begin{aligned} f(\lambda) &= \tfrac{1}{2\pi} \sum_{h=-\infty}^{\infty} (\cos(h\lambda) - i\sin(h\lambda))\gamma(h) \\ &= \tfrac{1}{2\pi} \sum_{h=-\infty}^{\infty} \cos(-h\lambda)\gamma(h) + 0 \\ &= f(-\lambda). \end{aligned}$$

For each positive integer N define

$$\begin{aligned} f_N(\lambda) &= \tfrac{1}{2\pi N}\mathbb{E}\Big(|\sum_{r=1}^{N} X_r e^{-ir\lambda}|^2\Big) \\ &= \tfrac{1}{2\pi N}\mathbb{E}\Big(\sum_{r=1}^{N} X_r e^{-ir\lambda}\sum_{s=1}^{N} X_s e^{is\lambda}\Big) \\ &= \tfrac{1}{2\pi N}\sum_{|h|<N}(N-|h|)e^{-ih\lambda}\gamma(h), \end{aligned}$$

where $\Gamma_N = [\gamma(i-j)]_{i,j=1}^{N}$. Clearly, the function f_N is nonnegative for each N, and since $f_N(\lambda) \to \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} e^{-ih\lambda}\gamma(h) = f(\lambda)$ as $N \to \infty$, f must also be nonnegative. This proves Proposition 4.1 (b). Turning to Proposition 4.1 (c),

$$\begin{aligned} \textstyle\int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda)d\lambda &= \textstyle\int_{-\pi}^{\pi} \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} e^{i(k-h)\lambda}\gamma(h)d\lambda \\ &= \tfrac{1}{2\pi}\sum_{h=-\infty}^{\infty}\gamma(h)\textstyle\int_{-\pi}^{\pi} e^{i(k-h)\lambda}d\lambda \\ &= \gamma(k), \end{aligned}$$

since the only nonzero summand in the second line is the one for which $h = k$.

Q.E.D.

Definition 4.2. A function f is the **spectral density** of a stationary time series $\{X_t\}$ with ACVF $\gamma(\cdot)$ if

(i) $f(\lambda) \geq 0$ for all $\lambda \in (0, \pi]$,

(ii) $\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)d\lambda$ for all integers h.

Remark 4.3. Spectral densities are essentially unique. That is, if f and g are two spectral densities corresponding to the autocovariance function $\gamma(\cdot)$, i.e., $\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)d\lambda = \int_{-\pi}^{\pi} e^{ih\lambda} g(\lambda)d\lambda$ for all integers h, then f and h have the same Fourier coefficients and hence are equal.

□

Proposition 4.4. *A real-valued function f defined on $(-\pi, \pi]$ is the spectral density of a stationary process if and only if*

(i) $f(\lambda) = f(-\lambda)$,

(ii) $f(\lambda) \geq 0$,

(iii) $\int_{-\pi}^{\pi} f(\lambda)d\lambda < \infty$.

Proof: If $\gamma(\cdot)$ is aboslutely summable, then (i)-(iii) follow from the basic properties of f, Proposition 4.1.

Conversely, suppose f satisfies (i)-(iii). Then it is easy to check, using (i), that the function defined by

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)d\lambda$$

is even. Moreover, if $a_r \in \mathbb{R}$, $r = 1, ..., n$, then

$$\begin{aligned}\sum_{r,s=1}^{n} a_r \gamma(r-s) a_s &= \int_{-\pi}^{\pi} \sum_{r,s=1}^{n} a_r a_s e^{i\lambda(r-s)} f(\lambda) d\lambda \\ &= \int_{-\pi}^{\pi} \left| \sum_{r=1}^{n} a_r e^{i\lambda r} \right|^2 f(\lambda) d\lambda \\ &\geq 0,\end{aligned}$$

so that $\gamma(\cdot)$ is also nonnegative definite and therefore, by Theorem 2.3 (also Theorem 2.1.1 of text [9]), is an autocovariance function.

Remark 4.5. Recall the theorem below. A real-valued function defined on the integers is the autocovariance function of a stationary time series if and only if it is even and nonnegative definite.

Corollary 4.6. *An absolutely summable function $\gamma(\cdot)$ is the autocovariance function of a stationary time series if and only if it is even and*

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \gamma(h) \geq 0, \ \forall \lambda \in (-\pi, \pi],$$

in which case $f(\cdot)$ is the spectral density of $\gamma(\cdot)$.

Proof: We have already established the necessity of $f(\lambda)$. now suppose it holds. Applying Proposition we conclude f is the spectral density of some autocovariance function. But this ACVF must be $\gamma(\cdot)$, since $\gamma(k) = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda$ for all integers k.

Q.E.D.

Example 4.7. Using Corollary 4.6, it is a simple matter to show that the function defined by

$$\kappa(h) = \begin{cases} 1, & if\ h = 0, \\ \rho, & if\ h = \pm 1, \\ 0, & otherwise, \end{cases}$$

is the ACVF of a stationar time series if and only $|\rho| \leq \frac{1}{2}$. Since $\kappa(\cdot)$ is even and nonzero only at lags 0, ±1, it follows from the corollary that κ is an ACVF if and only if the function

$$f(\lambda) = \frac{1}{2} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \gamma(h) = \frac{1}{2\pi} [1 + 2\rho \cos \lambda]$$

is nonnegative for all $\lambda \in [-\pi, \pi]$. But this occurs if and only if $|\rho| \leq \frac{1}{2}$.

□

Theorem 4.8. *(Spectral Representation of the ACVF) A function $\gamma(\cdot)$ defined on the integers is the ACVF of a stationary time series if and only if there eixsts a right-continuous, nondecreasing, bounded function F on $[-\pi, \pi]$ with $F(-\pi) = 0$ such that*

$$\gamma(h) = \int_{(-\pi,\pi]} e^{ih\lambda} dF(\lambda)$$

for all integers h. (For real-valued time series, F is symmetric in the sense that $\int_{(a,b]} dF(x) = \int_{[-b,-a)} e^{ih\lambda} dF(x)$ for all a and b such that $0 < a < b$.)

Remark 4.9. The function F is a **generalized distribution function** on $[-\pi, \pi]$ in the sense that $G(\lambda) = F(\lambda)/F(\pi)$ is a probability distribution function on $[-\pi, \pi]$. Note that since $F(\pi) = \gamma(0) = Var(X_1)$, the ACF of $\{X_t\}$ has spectral representation

$$\rho(h) = \int_{(-\pi,\pi]} e^{ih\lambda} dG(\lambda).$$

The function F in Theorem 4.8 is called the **spectral distribution function** of $\gamma(\cdot)$. If $F(\lambda)$ can be expressed as $F(\lambda) = \int_{-\pi}^{\lambda} f(y)dy$ for all $\lambda \in [-\pi, \pi]$, then f is the **spectral density function** and the time series is said to have a **continuous spectrum**. If F is a discrete distribution (i.e., if G is a discrete probability distribution), then the time series is said to have a **discrete spectrum**. The time series $X_t = A\cos(\omega t) + B\sin(\omega t)$, has a discrete spectrum.

Example 4.10. If $\{X_t\} \sim WN(0, \sigma^2)$, then $\gamma(0) = \sigma^2$ and $\gamma(h) = 0$ for all $|h| > 0$. This process has a flat spectral density

$$f(\lambda) = \frac{\sigma^2}{2\pi}, \quad -\pi \le \lambda \le \pi.$$

A process with this spectral density is called **white noise**, since each frequency in the spectrum contributes equally to the variance of the process.

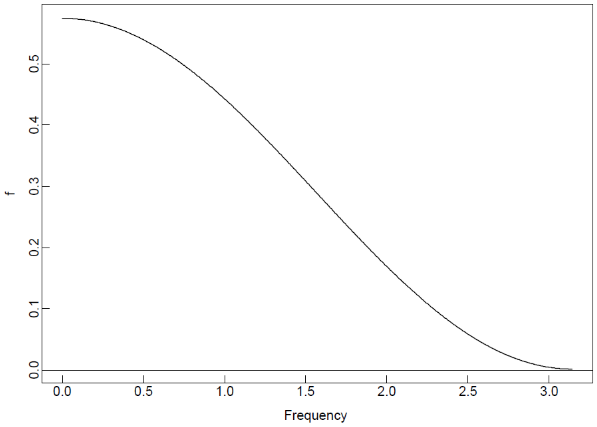
□

4.2 The Periodogram

Go back to Table of Contents. Please click TOC

If $\{X_t\}$ is a stationary time series $\{X_t\}$ with ACVF $\gamma(\cdot)$ and spectral density $f(\cdot)$, then just as the sample ACVF $\hat{\gamma}(\cdot)$ of the observations $\{x_1, ..., x_n\}$ can be regarded as a sample analogue of $\gamma(\cdot)$, so also can the periodogram $I_n(\cdot)$ of the observations be regarded as a sample analogue of $2\pi f(\cdot)$.

Figure 19: The spectral density $f(\lambda)$, $0 \le \lambda \le \pi$, of $X_t = Z_t + .9Z_{t-1}$ where $\{Z_t\} \sim WN(0, \sigma^2)$.



To introduce th periodogram, we consider the vector of complex numbers

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{C}^n,$$

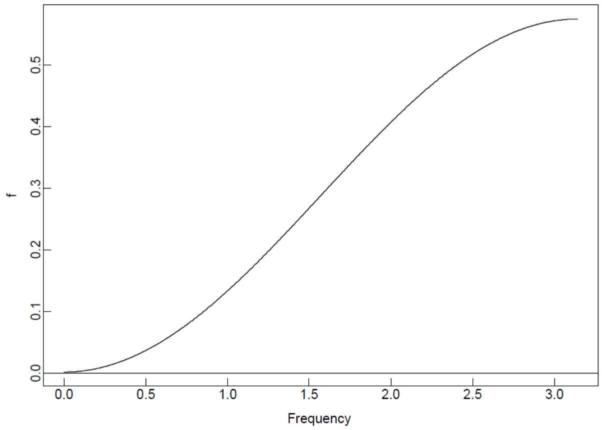
where $\mathbb{C}^n$ denotes the set of all column vectors with complex-valued components. Now let $\omega_k = 2\pi k/n$, where k is any integer between $-(n-1)/2$ and $n/2$ (includsive), i.e.,

$$\omega_k = \frac{2\pi k}{n}, \ k = -\left[\frac{n-1}{2}\right], \ldots, \left[\frac{n}{2}\right],$$

where $[y]$ denotes the largest integer less than or equal to y. We shall refer to the set F_n of these values as the **Fourier frequencies** associated with sample size n, noting that F_n is a subset of the interval $(-\pi, \pi]$. Correspondingly, we introduce the n vectors

$$\mathbf{e}_k = \frac{1}{\sqrt{n}} \begin{bmatrix} e^{i\omega_k} \\ e^{2i\omega_k} \\ \vdots \\ e^{ni\omega_k} \end{bmatrix}, \ k = -\left[\frac{n-1}{2}\right], \ldots, \left[\frac{n}{2}\right].$$

Figure 20: The spectral density $f(\lambda)$, $0 \le \lambda \le \pi$, of $X_t = Z_t - .9Z_{t-1}$ where $\{Z_t\} \sim WN(0, \sigma^2)$.



Now $\mathbf{e}_1, ..., \mathbf{2}_n$ are **orthonormal** in the sense that

$$\mathbf{e}_j^* \mathbf{e}_k = \begin{cases} 1, & if\ j = k, \\ 0, & if\ j \neq k, \end{cases}$$

where $\mathbf{e}_j^*$ denotes the row vector whose kth component is the complex conjugate of the kth component of $\mathbf{e}_j$. This implies that $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ is a basis for $\mathbb{C}^n$, so that any $\mathbf{x} \in \mathbb{C}^n$ can be expressed as the sum of n components,

$$\mathbf{x} = \sum_{k=-[(n-1)/2]}^{[n/2]} a_k \mathbf{e}_k.$$

The coefficients a_k are easily found by multiplying $\mathbf{X}$ on the left by $\mathbf{e}_k^*$ and using $\mathbf{e}_j^* \mathbf{e}_k$. Thus,

$$a_k = \mathbf{e}_k^* \mathbf{x} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t e^{-it\omega_k}.$$

The sequence $\{a_k\}$ is called the **discrete Fourier transform** of the sequence $\{x_1, ..., x_n\}$.

Remark 4.11. The tth component of $\mathbf{x}$ can be written as

$$x_t = \sum_{k=-[(n-1)/2]}^{[n/2]} a_k [\cos(\omega_k t) + i \sin(\omega_k t)],\ t = 1, ..., n,$$

showing that $\mathbf{x}$ is just a way of representing x_t as a linear combination of sine waves with frequencies $\omega_k \in F_n$.

Definition 4.12. The **periodogram** of $\{x_1, ..., x_n\}$ is the function

$$I_n(\lambda) = \frac{1}{n} \left| \sum_{t=1}^{n} x_t e^{-it\lambda} \right|^2.$$

Remark 4.13. If λ is one of the Fourier frequencies ω_k, then $I_n(\omega_k) = |a_k|^2$, and so from $\mathbf{x}$ and $\mathbf{e}_j^*\mathbf{e}_k$ we find at once that the squared length of $\mathbf{x}$ is

$$\sum_{t=1}^{n} |x_t|^2 = \mathbf{x}^*\mathbf{x} = \sum_{k=-[(n-2)/2]}^{[n/2]} |a_k|^2 = \sum_{k=-[(n-2)/2]}^{[n/2]} I_n(\omega_k).$$

The value of the periodogram at frequency ω_k is thus the contribution to this sum of squares from the "frequency ω_k" term $a_k\mathbf{e}_k$ in $\mathbf{x}$.

□

Proposition 4.14. *If $x_1, ..., x_n$ are any real numbjers and ω_k is any of the nonzero Fourier frequencies $2\pi k/n$ in $(-\pi, \pi]$, then*

$$I_n(\omega_k) = \sum_{|h|<n} \hat{\gamma}(h) e^{-ih\omega_k},$$

where $\hat{\gamma}(h)$ is the sample ACVF of $x_1, ..., x_n$.

Proof: Since $\sum\limits_{t=1}^{n} e^{-it\omega_k} = 0$ if $\omega_k \neq 0$, we can subtract the sample mean $\bar{x}$ from x_t in the defining equation of $I_n(\lambda)$ of $I_n(\omega_k)$. Hence,

$$\begin{aligned} I_n(\omega_k) &= n^{-1} \sum_{s=1}^{n}\sum_{t=1}^{n} (x_s - \bar{x})(x_t - \bar{x}) e^{-i(s-t)\omega_k} \\ &= \sum_{|h|<n} \hat{\gamma}(h) e^{-ih\omega_k}. \end{aligned}$$

□

In view of the similarity between $2\pi f(\lambda)$ and $I_n(\omega_k)$, a natural estimate of the spectral density $f(\lambda)$ is $I_n(\lambda)/(2\pi)$. For a very large class of stationary time series $\{X_t\}$ with strictly positive spectral density, it can be shown that for any fixed frequencies $\lambda_1, ..., \lambda_m$ such that $0 < \lambda_1 < \cdots < \lambda_m < \pi$, the joint distribution function $F_n(x_1, ..., x_m)$ of the periodogram values $(I_n(\lambda_1), ..., I_n(\lambda_m))$ converges, as $n \to \infty$, to $F(x_1, ..., x_m)$, where

$$F(x_1, ..., x_m) = \begin{cases} \prod\limits_{i=1}^{m} \left(1 - \exp\left\{\frac{-x_i}{2\pi f(\lambda_i)}\right\}\right), & \textit{if } x_1, ..., x_m > 0, \\ 0, & \textit{otherwise}, \end{cases}$$

Thus for large n the periodogram ordinates $(I_n(\lambda_1), ..., I_n(\lambda_m))$ are approximately distributed as independent exponential random variables with means $2\pi f(\lambda_1), ..., 2\pi f(\lambda_m)$, respectively. In particular, for each fixed $\lambda \in (0, \pi)$ and $\epsilon > 0$,

$$P[|I_n(\lambda) - 2\pi f(\lambda)| > \epsilon] \to p > 0, \ \textit{as } n \to \infty,$$

so the probability of an estimation error larger than ϵ cannot be made arbitrarily small by choosing a sufficiently large sample size n. Thus, $I_n(\lambda)$ is not a consistent estimator of $2\pi f(\lambda)$.

Since for large n the periodogram ordinates at fixed frequencies are approximately independent with variances changing only slightly over small frequency intervals, we might hope to construct a consistent estimator of

$f(\lambda)$ by averaging the periodogram estimates in a small frequency interval containing λ, provided that we can choose the interval in such a way that its width decreases to zero while at the same time the number of Fourier frequencies in the interval increases to ∞ as $n \to \infty$. This can be done, since the number of Fourier frequencies in any *fixed* frequency interval increases approximately linearly with n. Consider, for example, the estimator

$$\bar{f}(\lambda) = \frac{1}{2\pi} \sum_{|j| \leq m} (2m+1)^{-1} I_n(g(n,\lambda) + 2\pi j/n),$$

where $m = \sqrt{n}$ and $g(n,\lambda)$ is the multiple of $2\pi/n$ closest to λ. The number of periodogram ordinates being averaged is approximately $2\sqrt{n}$, and the width of the frequency interval over which the average is taken is approximately $2\sqrt{n}$, and the width of the frequency interval over which the average is taken is approximately $4\pi/\sqrt{n}$.

Definition 4.15. A **discrete spectral average estimator** of the spectral density $f(\lambda)$ has the form

$$\hat{f}(\lambda) = \frac{1}{2\pi} \sum^{|j| \leq m_n} W_n(j) I_n(g(n,\lambda) + 2\pi j/n),$$

where the **bandwidths** m_n satisfy

$$m_n \to \infty \; and \; m_n/n \to 0 \; as \; n \to \infty,$$

and the **weight functions** $W_n(\cdot)$ satisfy

$$W_n(j) = W_n(-j), \; W_n(j) \geq 0 \; for \; all \; j,$$

$$\sum_{|j| \leq m_n} W_n(j) = 1,$$

and

$$\sum_{|j| \leq m_n} W_n^2(j) \to 0 \; as \; n \to \infty.$$

Remark 4.16. The conditions imposed on the sequences $\{m_n\}$ and $\{W_n(\cdot)\}$ ensure consistency of $\hat{f}(\lambda)$ for $f(\lambda)$ for a very large class of stationary processes.

4.3 Time-Invariant Linear Filters

Go back to Table of Contents. Please click TOC

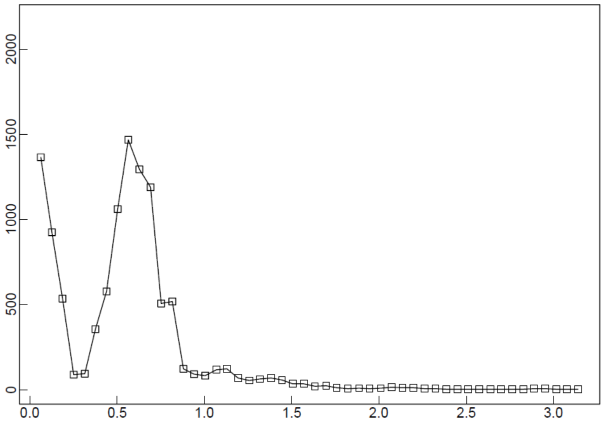
In Section 1.5 we saw the utility of time-invariant linear fitlers for smoothing the data, estimating the trend, eliminating the seasonal and/or trend components of the data. A linear process is the output of a time-invariant linear fitler (TLF) applied to a whtie noise input series. More generally, we say that the process $\{Y_t\}$ is the output of a linear filter $C = \{c_{t,k}, t, k = 0, \pm 1, ...\}$ applied to an input process $\{X_t\}$ if

$$Y_t = \sum_{k=-\infty}^{\infty} c_{t,k} X_k, \; t = 0, \pm 1, ...$$

The filter is said to be **time-invariant** if the weights $c_{t,t-k}$ are independent of t, i.e., if

$$c_{t,t-k} = \psi_k.$$

Figure 21: The spectral density estimate, $f(\lambda)$, $0 < \lambda \leq \pi$, of the sunspot numbers, 1770-1869, with weights $\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$.



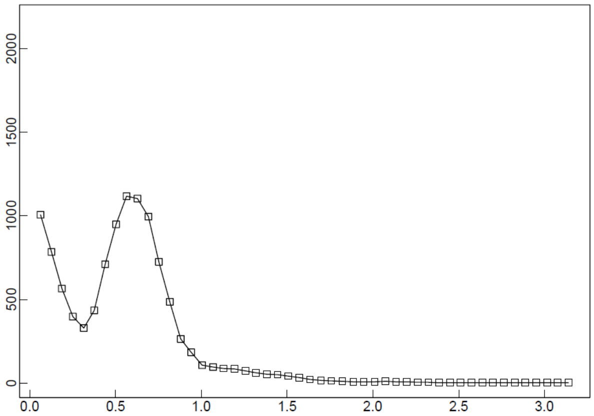
In this case,

$$Y_t = \sum_{k=-\infty}^{\infty} \psi_k X_{t-k}$$

and

$$Y_{t-s} = \sum_{k=-\infty}^{\infty} \psi_k X_{t-s-k},$$

Figure 22: The spectral density estimate, $f(\lambda)$, $0 < \lambda \leq \pi$, of the sunspot numbers, 1770-1869, with weights $\{\frac{1}{15}, \frac{2}{15}, \frac{3}{15}, \frac{3}{15}, \frac{3}{15}, \frac{2}{15}, \frac{1}{15}.\}$.



so that the time-shifted process $\{Y_{t-s}, t = 0, \pm 1, ...\}$ is obtained from $\{X_{t-s}, t = 0, \pm 1, ...\}$ by application of the same linear filter $\psi = \{\psi_j, j = 0, \pm 1, ...\}$. The TLF ψ is said to be **causal** if

$$\psi_j = 0 \ for \ j < 0,$$

since then Y_t is expressible in terms only of X_s, $s \leq t$.

Example 4.17. The filter defined by

$$Y_t = aX_{-t}, \ t = 0, \pm 1, ...,$$

is linear but not time-invariant, since $c_{t,t-k} = 0$ except when $2t = k$. Thus, $c_{t,t-k}$ depends on the value of t.

□

Example 4.18. The filter

$$Y_t = (2q+1)^{-1} \sum_{|j| \leq q} X_{t-j}$$

is a TLF with $\psi_j = (2q+1)^{-1}$, $j = -q, ..., q$, and $\psi_j = 0$ otherwise.

□

Spectral methods are particularly valuable in describing the behavior of time-invariant linear filters as well as in designing filters for particular purposes such as the suppression of high-frequency components. The following proposition shows how the spectral density of the output of a TLF is related to the spectral density of the input – a fundamental result in the study of time-invariant linear filters.

Proposition 4.19. *Let $\{X_t\}$ be a stationary time series with mean zero and spectral density $f_X(\lambda)$. Suppose that $\Psi = \{\psi_j, \ j = 0, \pm 1, ...\}$ is an absolutely summable TLF (i.e., $\sum_{j=-\infty}^{\infty} \Psi_j| < \infty$). Then the time series*

$$Y_t = \sum_{j=-\infty}^{\infty} \Psi_j X_{t-j}$$

is stationary with mean zero and spectral density

$$f_Y(\lambda) = |\Psi(e^{-i\lambda})|^2 f_X(\lambda) = \Psi(e^{-i\lambda})\Psi(e^{i\lambda}) f_X(\lambda),$$

*where $\Psi(e^{-i\lambda}) = \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda}$. (The function $\Psi(e^{-i\cdot})$ is called the **transfer function** of the filter, and the squared modulus $|\Psi(e^{-i\cdot})|^2$ is referred to as the **power transfer function** of the filter.)*

Proof: Applying Proposition 2.12 (also Proposition 2.2.1 from text [9]), we see that

$$\gamma_Y(h) = \sum_{j,k=-\infty}^{\infty} \psi_j \psi_k \gamma_X(h + k - j).$$

Since $\{X_t\}$ has spectral density $f_X(\lambda)$, we have

$$\gamma_X(h+k-j) = \int_{-\pi}^{\pi} e^{i(h-j+k)\lambda} f_X(\lambda) d\lambda,$$

which, by substituting $\gamma_X(h+k-j)$ into $\gamma_Y(h)$, gives

$$\begin{aligned}
\gamma_Y(h) &= \sum_{j,k=-\infty}^{\infty} \psi_j \psi_k \int_{-\pi}^{\pi} e^{i(h-j+k)\lambda} f_X(\lambda) d\lambda \\
&= \int_{-\pi}^{\pi} \left(\sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda} \right) \left(\sum_{k=-\infty}^{\infty} \psi_k e^{ik\lambda} \right) e^{ih\lambda} f_X(\lambda) d\lambda \\
&= \int_{-\pi}^{\pi} e^{ih\lambda} \left| \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda} \right|^2 f_X(\lambda) d\lambda.
\end{aligned}$$

The last expression immediately identifies the spectral density function of $\{Y_t\}$ as

$$f_Y(\lambda) = |\Psi(e^{-i\lambda})|^2 f_X(\lambda) = \Psi(e^{-i\lambda})\Psi(e^{i\lambda}) f_X(\lambda).$$

Q.E.D.

Remark 4.20. Recall the proposition. Let $\{Y_t\}$ be a stationary time series with mean 0 and covariance function γ_Y. If $\sum_{j=-\infty}^{\infty} |\Psi_j| < \infty$, then the time series

$$X_t = \sum_{j=-\infty}^{\infty} \Psi_j Y_{t-j} = \Psi(B) Y_t$$

is stationary with mean 0 and autocovariance function

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \Psi_j \Psi_k \gamma_Y(h+k-j).$$

In the special case where $\{X_t\}$ is a linear process,

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \Psi_j \Psi_{j+h} \sigma^2.$$

Figure 23: The spectral density $f(\lambda)$, $0 \le \lambda \le \pi$, of $X_t = Z_t - .9Z_{t-1}$ where $\{Z_t\} \sim WN(0, \sigma^2)$.

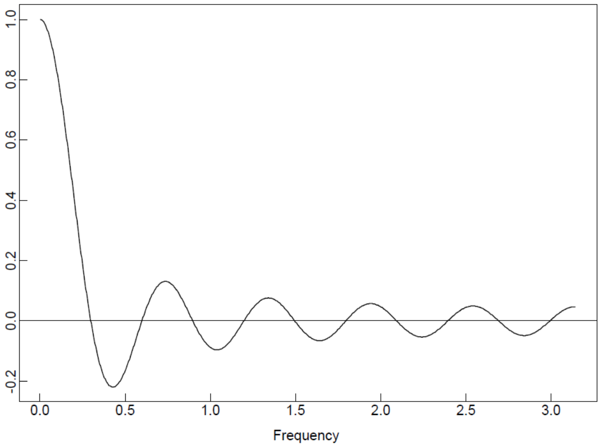
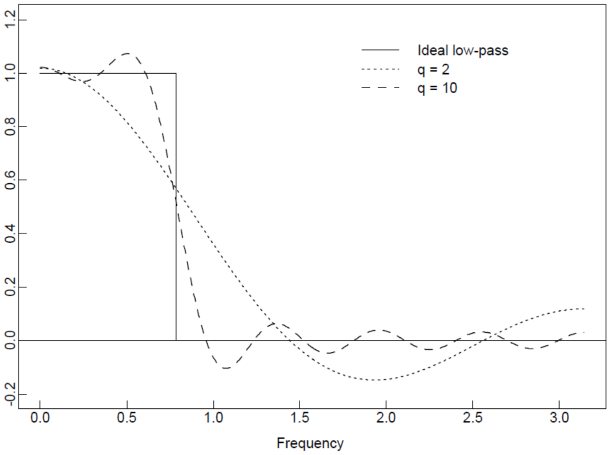


Figure 24: The spectral density $f(\lambda)$, $0 \le \lambda \le \pi$, of $X_t = Z_t - .9Z_{t-1}$ where $\{Z_t\} \sim WN(0, \sigma^2)$.



4.4 The Spectral Density of an ARMA Process

Go back to Table of Contents. Please click TOC In Section 4.1 the spectral density was computed for an MA(1) and for an AR(1) process. As an application of Proposition 4.19, we can now easily derive the spectral density of an arbitrary ARMA(p,q) process.

Theorem 4.21. ***Spectral Density of an ARMA(p,q) Process:*** *If*

$\{X_t\}$ is a causal ARMA(p,q) process satisfying $\phi(B)X_t = \theta(B)Z_t$, then

$$f_X(\lambda) = \frac{\sigma^2|\theta(e^{-i\lambda})|^2}{2\pi|\phi(e^{-i\lambda}|^2}, \ -\pi \leq \lambda \leq \pi.$$

Because the spectral density of an ARMA process is a ratio of trigonometric polynomials, it is often called a **rational spectral density**.

Proof: From, $X_t = \chi(B)\phi(B)X_t = \chi(B)\theta(B)Z_t = \psi(B)Z_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$, $\{X_t\}$ is obtained from $\{Z_t\}$ by application of the TLF with transfer function

$$\Psi(e^{-i\lambda}) = \frac{\theta(e^{-i\lambda})}{\phi(e^{i\lambda}}.$$

Since $\{Z_t\}$ has spectral density $f_Z(\lambda) = \sigma^2/(2\pi)$, the result now follows from 4.19.

Q.E.D.

Example 4.22. For an AR(2) process becomes

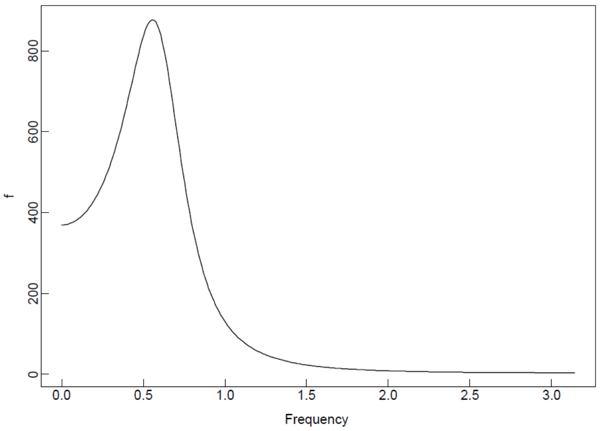
$$\begin{aligned} f_X(\lambda) &= \tfrac{\sigma^2}{2\pi(1-\phi_1 e^{-i\lambda}-\phi_2 e^{-2i\lambda})(1-\phi_1 e^{i\lambda}-\phi_2 e^{2i\lambda})} \\ &= \tfrac{\sigma^2}{2\pi(1+\phi_1^2+2\phi_2+\phi_2^2+2(\phi_1\phi_2-\phi_1)\cos\lambda-4\phi_2\cos^2\lambda)} \end{aligned}$$

Figure 25 shows the spectral fitted to the mean-corrected sunspot series. Notice the well-defined peak in the model spectral density. The frequency at which this peak occurs can be found by differentiating the denominator of the spectral density with repect to $\cos\lambda$ and setting the derivative equal to zero. This gives

$$\cos\lambda = \frac{\phi_1\phi_2 - \phi_1}{4\phi_2} = 0.849.$$

The corresponding freuqency is $\lambda = 0.556$ radians per year, or equivalently $c = \lambda/(2\pi) = 0.0885$ cycles a per year, and the corresponding period is therefore $1/0.0885 = 11.3$ years. The model thus reflects the approximate cyclic behavior of the data already pointed out in Example 4.2.2 of text [9]. The model spectral density in Figure 25 should be compared with the rescaled periodogram of the data nad the nonparametric spectral density estimates of Figures 4.9-4.11 [9].

Figure 25: The spectral density $f_X(\lambda)$, $0 \leq \lambda \leq \pi$ of the AR(2) model $X_t - 1.318X_{t-1} + 0.634X_{t-2} = Z_t$, $\{Z_t\} \sim WN(0, 289.2)$ [9] fitted to the mean-corrected sunspot series.



□

5 Modeling and Forecasting with ARMA Processes

Go back to Table of Contents. Please click TOC

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 $\phi_i, i = 1, ..., p\}$, $\{\theta_i, i = 1, ..., q\}$, 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.

This chapter is primarily devoted to the problem of estimating the parameters $\phi = (\phi_i, ..., \phi_p)h'$, $\theta = 9\theta_i, ..., \theta_q)h'$, 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 is is appropriate to fit z zero-mean ARMA model to the adjusted data $x_1, ..., x_n$. If the model fitted to the mean-corrected data is

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

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 ϕ and θ 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 $\phi_1, ..., \phi_p, \theta_1, ..., \theta_q$ 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. Maximization of the likelihood and selection of the minimum AICC model over a specified range of p and q values can be found in programs.

5.1 Preliminary Estimation

Go back to Table of Contents. Please click TOC

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

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

The Yule-Walker and Burg procedures apply to the fitting of pure autoregressive models.

5.1.1 Yule-Walker Estimation

Go back to Table of Contents. Please click TOC

For a pure autoregressive model the moving-average polynomial $\theta(z)$ is identically 1, and the causality assumption in $\phi(B)X_t = \theta(B)Z_t,\ \{Z_t\} \sim WN(0, \sigma^2)$ allows us to write X_t in the form

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

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

$$\Gamma_p \phi = \gamma_p$$

and

$$\sigma^2 = \gamma(0) - \phi' \gamma_p,$$

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

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

$$\hat{\Gamma}_p \hat{\phi} = \hat{\gamma}_p$$

and

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

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

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

Definition 5.1. Sample Yule-Walker Equations:

$$\hat{\phi} = (\hat{\phi}_1, ..., \hat{\phi}_p)' = \hat{R}_p^{-1}\hat{\rho}_p$$

and

$$\hat{\sigma}^2 = \hat{\gamma}(0)[1 - \hat{\rho}_p\hat{R}_p^{-1}\hat{\rho}_p],$$

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

With $\hat{\phi}$ as defined by $\hat{\phi} = (\hat{\phi}_1, ..., \hat{\phi}_p)' = \hat{R}_p^{-1}\hat{\rho}_p$ in the definition, it can be shown that $1 - \hat{\phi}_1 z - \cdots - \hat{\phi}_p z^p \neq 0$ for $|z| \leq 1$. Hence the fitted model

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

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

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

However, from $\Gamma_p\phi = \gamma_p$ and $\sigma^2 = \gamma(0) - \phi'\gamma_p$ we see that the solution of these equations is $\gamma_F(h) = \hat{\gamma}(h)$, $h = 0, ..., p$, so that the autocovariances of the fitted model at lags $0, 1, ..., p$ 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 auto covariances at lags $0, ..., p$ are $\gamma(0), ..., \gamma(p)$. There may not, however, be an MA(p) process with this property.

It is often the case that moment estimators, i.e., estimators that (like $\hat{\phi}$) 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 $\phi_1, ..., \phi_p$ of an AR(p) process have approximately the same distribution for large samples as the corresponding maximum likelihood estimators. For our purposes it suffices to note the following:

Definition 5.2. Large-Sample Distribution of Yule-Walker Estimators:

For a large sample from an AR(p) process,

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

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 ϕ and each of its components as 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.

- Some guidance in the choice of order is provided by a large-sample result, which states that if $\{X_t\}$ is the causal AR(p) process defined by $\phi(B)X_t = \theta(B)Z_t$, $\{Z_t\} \sim iid(o0, \sigma^2)$ 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
$$\phi_m = \hat{R}_m^{-1}\hat{\rho},\ m > p,$$
then the *last component*, $\hat{\phi}_{mm}$, of the vector $\hat{\phi}_m$ is approximately normally distributed with mean 0 and variance $1/n$, Notice that $\hat{\phi}_{mm}$ is exactly the sample partial autocorrelation at lag m.
- A more systematic approach to order selection is to find the values of p and ϕ_p that minimize the AICC statistic
$$AICC = -2\ln L(\phi_p, S(\phi_p)/n) + 2(p+1)n/(n-p-2),$$
where L is the Gaussian likelihood in
$$L(\phi, \theta, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n r_0 \dots r_{n-1}}} \exp\left\{ -\frac{1}{2\sigma^2}\sum_{j=1}^{n}\frac{(X_j - \hat{X}_j)^2}{r_{j-1}} \right\}$$
and S is defined in
$$S(\hat{\phi}, \hat{\theta}) = \sum_{j=1}^{n}(X_j - \hat{X}_j)^2/r_{j-1}.$$

Definition 5.3. The **fitted Yule-Walker AR(m) model** is
$$X_t - \hat{\phi}_{m1}X_{t-1} - \cdots - \hat{\phi}_{mm}X_{t-m} = Z_t,\ \{Z_t\} \sim WN(0, \hat{v}_m),$$
where
$$\hat{\phi}_m = (\hat{\phi}_{m1}, ..., \hat{\phi}_{mm})' = \hat{R}_m^{-1}\hat{\rho}$$
and
$$\hat{v}_m = \hat{\gamma}(0)[1 - \hat{\rho}_m'\hat{R}_m^{-1}\hat{\rho}_m].$$

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 2.5.1, where we derived a recursive scheeme for solving the equations $\Gamma_p\phi = \gamma_p$ and $\sigma^2 = \gamma(0) - \phi'\gamma_p$, with p successively taking the values 1,2, ... Here we can use exactly the same scheme (the Durbin-Levinson algorithm) to solve the Yule-Walker equations $\hat{\Gamma}_p\hat{\phi} = \hat{\gamma}_p$ and $\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\phi}'\hat{\gamma}_p$, the only difference being that the covariances in the original formulas are replaced by their sample sample counterparts.

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 $\hat{\phi}_p$ to derive approximate lrage-sample confidence regions for the true coefficient vector ϕ_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

$$\{\phi \in \mathbb{R}^p : \ (\hat{\phi}_p - \phi)'\hat{\Gamma}_p(\hat{\phi}_p - \phi) \leq n^{-1}\hat{v}_p\chi_{1-\alpha}(p)\}$$

contains ϕ_p with probability close to $(1-\alpha)$. 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 $\hat{v}_p\Gamma_p^{-1}$, then for large n the interval bounded by

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

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

Relative Efficiency of Estimators

The performance of two copmeting 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 $X_1, ..., X_n$. If $\hat{\theta}_n^{(i)}$ is approximately $\mathrm{N}(\theta, \sigma_i^2(\theta))$ 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

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

If $e(\theta, \hat{\theta}^{(1)}) \leq 1$ 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, $e(\theta, \hat{\theta}^{(1)}, \hat{\theta}^{(2)}) \leq 1$ for for some $\theta \in \Theta$).

5.1.2 Burg's Algorithm

Go back to Table of Contents. Please click TOC

The Yule-Walker coefficients $\hat{\phi}_{p1}, ..., \hat{\phi}_{pp}$ are precisely the coefficients of the best linear predictor of X_{p+1} in terms of $\{X_p, ..., X_1\}$ under the assumption that the ACF of $\{X_t\}$ coincides with the sample ACF at lags $1, ..., p$.

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

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

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

and

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

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

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

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 equations above with $i = 2$. Then we minimize

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

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

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

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

and

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

The resulting estimtates of ϕ_{pj}, $j = 1, ..., p$, are the coefficient estimates of the Burg AR(p) model for the data $\{x_1, ..., x_n\}$. The Burg estimate of the whtie 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 to solving the following recursions:

Algorithm 5.4. ***Burg's Algorithm:***

$$\begin{aligned}
d(1) &= \sum_{t=2}^{n} (u_0^2(t-1) + v_0^2(t), \\
\phi_{ii}^{(B)} &= \tfrac{2}{d(i)} \sum_{t=i+1}^{n} v_{t-1}(t) u_{t-1}(t-1), \\
d(i+1) &= (1 - \phi_{ii}^{(B)2}) d(i) - v_i^2(i+1) - u_i^2(n), \\
\sigma_i^{(B)2} &= [(1 - \phi_{ii}^{(B)2} d(i)]/[2(n-i)].
\end{aligned}$$

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, $N(\phi, n^{-1}\sigma^2\Gamma_p^{-1})$. 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.

5.1.3 The Innovations Algorithm

Go back to Table of Contents. Please click TOC

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

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

of orders $m = 1, 2, ...$ by means of the innovations algorithm (Section 2.5.2). The estimated coefficient vectors $\hat{\theta}_m := (\hat{\theta}_{m1}, ..., \hat{\theta}_{mm})'$ and white noise variances $\hat{v}_m$, $m = 1, 2, ...$, are specified in the following definition.

Definition 5.5. The **fitted innovations MA(m) model** is

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

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

Remark 5.6. If can be shown (see Brockwell and Davis, 1988) [10] that if $\{X_t\}$ is an invertible MA(q) process

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

with $\mathbb{E}(Z_t^4) < \infty$, 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

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

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

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

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

□

Remark 5.7. 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 $\hat{\phi}_p$ are consistent estimators of $(\phi_1, ..., \phi_p)'$ as the sample size $n \to \infty$. However, for an MA(q) process the estimator $\hat{\theta}_q = (\theta_{q1}, ..., \theta_{qq})'$ is not consistent for $(\theta_1, ..., \theta_q)'$. For consistency it is necessary to use the estimators $(\theta_{m1}, ..., \theta_{mq})'$ with $m(n)$ satisfying the conditions of Remark 5.6. The choise of m for any fixed sample size can be made by increasing m until the vector $(\theta_{m1}, ..., \theta_{mq})'$ stabilizes. It is found in practice that there is a large range of values of m for which the fluctuations in θ_{nj} are small compared with the estimated asymptotic standard deviation $n^{-1/2}(\sum_{i=0}^{j-1} \hat{\theta}_{mi}^2)^{1/2}$ as found from $a_{ij} = \sum_{r=1}^{\min(i,j)} \theta_{i-r}\theta_{j-r}$ 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.

- From Section 3.2.2 that for an MA(q) processthe 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 $n^{-1}[1 + 2\rho^2(1) + \cdots + 2\rho^2(q)]$. This result enables us to use the graph of $\hat{\rho}(m)$, $m = 1, 2, ...$, 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".
- If in addition to examining $\hat{\rho}(m)$, $m = 1, 2, ...$, we examine the coefficient vectors $\hat{\theta}_m$, $m = 1, 2, ...$, 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 $\hat{\theta}_{m1}, ..., \hat{\theta}_{mq}$ of the coefficients. By inspecting the estimated coefficients $\hat{\theta}_{m1}, ..., \hat{\theta}_{mm}$ for $m = 1, 2, ...$ and the ratio of each coefficient estimate $\hat{\theta}_{mj}$ to 1.96 times its approximate standard deviation $\sigma_j = n^{-1/2}[\sum_{t=0}^{j-1} \hat{\theta}_{ni}^2]^{1/2}$, 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 propgram, but it may be altered manually. As m is increased the values $\hat{\theta}_{m1}, ..., \hat{\theta}_{mm}$ 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 $\hat{\theta}_q = (\hat{\theta}_{m1}, ..., \hat{\theta}_{mq})'$ that minimize the AICC statistic

$$AICC = -2 \ln L(\theta_q, S(\theta_q)/n) + 2(q+1)n/(n-q-2),$$

where L is the Gaussian likelihood defined in

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

and S is defined in

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

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

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

where the coefficients ψ_j satisfy

$$\psi_j = \theta_j + \sum_{i=1}^{\min(j,p)} \phi_i \psi_{j=i},\ j = 0, 1, ...,$$

and we define $\theta_0 := 1$ and $\theta_j := 0$ for $j > q$. To estimate $\psi_1, ..., \psi_{p+q}$ we can use the innovation estimates $\hat{\theta}_{m1}, ..., \hat{\theta}_{m,p+q}$, whose large-sample behavior is specified in Remark 5.6. Replacing ψ_j by $\hat{\theta}_{mj}$ in ψ_j and solving the resulting equations

$$\hat{\theta}_{mj} = \theta_j + \sum_{\substack{i=1\\ \min(j,p)}} \phi_i \hat{\theta}_{m,j-i},\ j = 1, ..., p+q,$$

for ϕ and θ, we obtain initial parameter estimates $\hat{\phi}$ and $\hat{\theta}$. To solve $\hat{\theta}_{mj}$ we first find ϕ from the last q equations:

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

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

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

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

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

where $\hat{X}_t$ is the one-step predictor of X_t computed from the fitted coefficient vectors $\hat{\phi}$ and $\hat{\theta}$, and r_{t-1} is defined in page 101 of text [9]

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

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

$$AICC = -2 \ln L(\phi_p, \theta_q, S(\phi_p, \theta_q)/n) + 2(p + q + 1)n/(n - p - q - 2),$$

which is discussed in more detail in Section 5.5.

5.1.4 The Hannan-Rissanen Algorithm

Go back to Table of Contents. Please click TOC

The defining equations for a causal AR(p) model have the form of a linear regression model with coefficient vector $\phi = (\phi_1, ..., \phi_p)'$. 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 $X_{t-1}, ..., X_{t-p}$, but also on unobserved quantities $Z_{t-1}, ..., Z_{t-q}$. Nevertheless, it is still possible to apply least squares regression to the estimation of ϕ and θ by first replacing the unobserved quantities $Z_{t-1}, ..., Z_{t-q}$ in $\phi(B)X_t = \theta(B)Z_t$, $\{Z_t\} \sim WN(0, \sigma^2)$ by estimated values $\hat{Z}_{t-1}, ..., \hat{Z}_{t-q}$. The parameters ϕ and θ are then estimated by regressing X_t onto $X_{t-1}, ..., X_{t-p}, \hat{Z}_{t-1}, ..., \hat{Z}_{t-q}$. These are the main steps.

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

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

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

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

with respect to β. This gives the Hannan-Rissanen estimator

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

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

$$Z = \begin{bmatrix} X_{m+q} & X_{m+q-1} & \dots & X_{m+q+1-p} & \hat{Z}_{m+q} & \hat{Z}_{m+q-1} & \dots & \hat{Z}_{m+1} & \\ X_{m+q+1} & X_{m+q} & \dots & X_{m+q+2-p} & \hat{Z}_{m+q+1} & \hat{Z}_{m+q} & \dots & \hat{Z}_{m+2} & \\ \vdots & \vdots & \ddots & \dots & \dots & \dots & \dots & \ddots & \vdots \\ X_{n-1} & X_{n-2} & \dots & X_{n-p} & \hat{Z}_{n-1} & \hat{Z}_{n2} & \dots & \hat{Z}_{n-q} & \end{bmatrix}$$

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

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

5.2 Maximum Likelihood Estimation

Go back to Table of Contents. Please click TOC Suppose that $\{X_t\}$ is a Gaussian time series with mean zero and autocovariance function $\kappa(i,j) = \mathbb{E})X_iX_j)$. Let $\mathbf{X}_n = (X_1, ..., X_n)'$ and let $\hat{\mathbf{X}}_n = (\hat{X}_1, ..., \hat{X}_n)'$, where $\hat{X}_1 = 0$ and $\hat{X}_j = \mathbb{E}(X_j | X_1, ..., X_{j-1}) = P_{j-1}X_j$, $j \geq 2$. Let Γ_n denote the covariance matrix $\Gamma_n = \mathbb{E}(\mathbf{X}_n\mathbf{X}_n')$, and assume that Γ_n is nonsingular.

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

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

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}, $j = 1, ..., n$, both of which are easily calculated recursively from the innovations algorithm.

Let θ_{ij}, $j = 1, ..., oi$, $i = 1, 2, ...$, 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 2.5.2. From $\mathbf{X}_n = C_n(\mathbf{X}_n - \hat{\mathbf{X}}_n)$ we have the identify

$$\mathbf{X}_n = C_n(\mathbf{X}_n - \hat{X}_n).$$

We also know from remark below (from Section 2.5.2.) that the components of $\mathbf{X}_n - \hat{\mathbf{X}}_n$ are uncorrelated. Consequently, by the definition of v_j, $\mathbf{X}_n - \hat{X}_n$ has the diagonal covariance matrix

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

From $\mathbf{X}_n = C_n(\mathbf{X}_n - \hat{\mathbf{X}}_n)$ and $\sum_{\mathbf{YY}} = B\sum_{\mathbf{XX}} B'$ we conclude that

$$\Gamma_n = C_nD_nC_n'.$$

From $\mathbf{X}_n = C_n(\mathbf{X}_n - \hat{X}_n)$ and $\Gamma_n = C_nD_nC_n'$ we see that

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

and

$$\det\Gamma_n = (\det C_n)^2(\det D_n) = v_0v_1 \dots v_{n-1}.$$

The likelihood $L(\Gamma_n) = (2\pi)^{-n/2}(\det\Gamma_n)^{-1/2}\exp\bigg(-\frac{1}{2}\mathbf{X}_n'\Gamma_n^{-1}\mathbf{X}_n\bigg)$ of the vector $\mathbf{X}_n$ therefore reduces to

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

If Γ_n is expressible in terms of a finite number of unknown parameters $\beta_1, ..., \beta_r$ (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 $X_1, X_2, ..., X_n$ 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., Lehamnn, 1983) [21].

Remark 5.8. From Remark 5 on page 73 in text [9]: Whitle the Durbin-Levinson recursion gives the coefficients of $X_n, ..., X_1$ in the representation $\hat{X}_{n+1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}$, the innovations algorithms gives the coefficients of $(X_n - \hat{X}_n), ..., (X_1 - \hat{X}_1)$, in the expansion $\hat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j})$. The latter expansion has a number of advantages deriving from the fact that the innovations are uncorrelated. It can also be greatly simplified in the case of ARMA(p,q) series, as we shall see in Section 3.3. An immediate consequence is the innovations representation of $X_n + 1$ itself. Thus (defining $\theta_{n0} := 1$),

$$X_{n+1} = X_{n+1} - \hat{X}_{n+1} + \hat{X}_{n+1} = \sum_{j=0}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), n = 0, 1, 2, ...$$

□

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

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

and

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

where θ_{nj} and r_n are determined by the innovations algorithm with κ as in $\kappa(i,j)$ in Section 3.3 and $m = \max(p,q)$. Subsituting in the general expression $L(\Gamma_n) = \frac{1}{\sqrt{(2\pi)^n v_0 \ldots v_{n-1}}} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} (X_j - \hat{X}_j)^2 / v_{j-1} \right)$, we obtain the following:

Definition 5.9. The Gaussian Likelihood for an ARMA Process:

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

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

Definition 5.10. Maximum Likelihood Estimators:

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

where

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

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

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

Least Squares Estimation for Mixed Models

The least squares estimates $\tilde{\phi}$ and $\tilde{\theta}$ of ϕ and θ are obtained by minimizing the function S rather than l, subject to the constraints that the model be causal and invertible. The least squares estimate of σ^2 is

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

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:

Definition 5.11. AICC Criterion:

Choose p, q, ϕ_p, and θ_q to minimize

$$AICC = -2\ln L(\phi_p, \theta_q, S(\phi_p, \theta_q)/n) + 2(p+q+1)n/(n-p-q-2).$$

For any fixed p and q it is clear that the AICC is minimized when ϕ_p and θ_q are the vectors that minimze $-2\ln L(\phi_p, \theta_q, S(\phi_p, \theta_q)/n)$, 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 tis justification are discussed in detail in Section 5.5.

Confidence Regions for the Coefficients

For large sample size the maximum likelihood estimator $\hat{\beta}$ of $\beta := (\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)'$ is approximately normally distributed with mean β and covariance matrix $[n^{-1}V(\beta)]$ which can be approximated by $2H^{-1}(\beta)$, where H is the Hessian matrix $[\partial^2 \ell(\beta)/\partial\beta_i\partial\beta_j]_{i,j=1}^{p+q}$.

Definition 5.12. Large-Sample Distribution of Maximum Likelihood Estimators:

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

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

Example 5.13. **An AR(p) model**

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

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

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

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

□

Example 5.14. **An MA(q) model**

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

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

Then it can be shown that

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

Inspection of the results of Example 5.14 and replacement of ϕ_i by $-\theta_i$ yields

$$\begin{aligned} MA(1) : V(\theta) &= (1 - \phi_1^2), \\ MA(2) : V(\theta) &= \begin{bmatrix} 1 - \theta_2^2 & \theta_1(1 - \theta)2) \\ \theta_1(1 - \theta_2) & 1 - \theta_2^2 \end{bmatrix} \end{aligned}$$

□

Example 5.15. **An ARMA(1,1) model**

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

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

□

5.3 Diagnostic Checking

Go back to Table of Contents. Please click TOC

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 $\hat{\phi}$, $\hat{\theta}$, and $\hat{\sigma}^2$ of the parameters ϕ, θ, and σ. In the course of this procedure the predicted values $\hat{X}_t(\hat{\phi}, \hat{\theta})$ of X_t based on $X_1, ..., X_{t-1}$ are computed for the fitted model. The **residuals** are then defined

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

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

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

Moreover, $W_t(\phi, \theta)$ approximates the white noise term in the defining equation $\phi(B)X_t = \theta(B)Z_t$ in the sense that $\mathbb{E}(W_t(\phi, \theta) - Z_t)^2 \to 0$

as $t \to \infty$. Consequently, the properties of the residuals $\{\hat{W}_t\}$ should reflect those of the white noise sequence $\{Z_t\}$ generating the underlying ARMA(p,q) process. In particular, the sequence $\{\hat{W}_t\}$ should be approximately i uncorrelated if $\{Z_t\} \sim WN(0, \sigma^2)$, (ii) independent if $\{Z_t\} \sim IID(0, \sigma^2)$, and (iii) normally distributed if $Z_t \sim N(0, \sigma^2)$.

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

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

5.4 Forecasting

Go back to Table of Contents. Please click TOC

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.

5.5 Order Selection

Go back to Table of Contents. Please click TOC

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 froecasting, 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. 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.

5.5.1 The FPE Criterion

Go back to Table of Contents. Please click TOC

The FPE criterion was developed by Akaike (1969) to select the appropriate order of an AR process to fit to a time series $\{X_t, ..., X_n\}$ [1]. Instead to choose the order p to make the estmiated whtie noise variance small, 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 $\{X_1, ..., X_n\}$ is a realization of an AR(p) process with coefficients $\phi_1, ..., \phi_p$, are the maximum likelihood estimators of the coefficients based on $\{X_1, ..., X_n\}$, are the maximum likelihood estimators of the coefficients based on $\{X_1, ..., X_n\}$ 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

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

where $\phi_p' = (\phi_1, ..., \phi_p)'$, $\hat{\phi}_p' = (\hat{\phi}_1, ..., \hat{\phi}_p)'$, 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 $X_1, ..., X_n$, nad using the independence of $\{X_1, ..., X_n\}$ and $\{Y_1, ..., Y_n\}$, we obtain

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

where $\Gamma_p = \mathbb{E}[Y_i Y_j]_{i,j=1}^p$. We can approximate the last term by assuming that $n^{-1/2}(\hat{\phi}_p - \phi_p)$ has its large-sample distribution $N(\mathbf{0}, \sigma^2 \Gamma_p^{-1})$. From Problem 5.13 of text [9], we have

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

If $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2, then for large n, $n\hat{\sigma}^2/\sigma^2$ is distributed approximatedly as chi-squared with $(n-p)$ degrees of freedom. Therefore we replace σ^2 in $\mathbb{E}(Y_{n+1} - \hat{\phi}_1 Y_n - \cdots - \hat{\phi}_p Y_{n+1-p})^2 \approx \sigma^2(1 + \frac{p}{n})$ by the estimator $n\hat{\sigma}^2/(n-p)$ to get the estimated mean square prediction error of Y_{n+1},

$$FPE_p = \hat{\sigma}^2 \frac{n+p}{n-p}.$$

5.5.2 The AICC Criterion

Go back to Table of Contents. Please click TOC

A more generally applicable criterion for model selection that the FPE is the information criterion of Akaike (1973) [**?**], known as the AIC. The 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) [19].

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

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

where

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

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

1974) [25],

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

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

Given observations $X_1, ..., X - n$ of an ARMA process with unknown parameters $\theta = (\beta, \sigma^2)$, 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 $\theta = (\beta, \sigma^2)$, $f(\cdot;\theta)$ is the probability density of $(Y_t, ..., Y_n)'$, where $\{Y_t\}$ is a Gaussian ARMA(p, q) process with coefficient vector β and white noise variance σ^2. (The dependence of θ on p adn q is through the dimension of the autoregressive and moving-average coefficients inβ.)

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

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

where L_x, L_y, S_X, and S_Y are defined as in

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

and

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

Hence,

$$\begin{aligned} \mathbb{E}_\theta(\triangle(\hat{\theta}|\theta)) &= \mathbb{E}_{\beta,\sigma^2}(-2\ln L_Y(\hat{\beta}, \hat{\sigma}^2)) \\ &= \mathbb{E}_{\beta,\sigma^2}(-2\ln L_X(\hat{\beta}, \hat{\sigma}^2)) + \mathbb{E}_{\beta,\sigma^2}\left(\tfrac{S_Y(\hat{\beta})}{\hat{\sigma}^2}\right) - n. \end{aligned}$$

It can be shown using large-sample approximations that

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

from which we see that $-2\ln L_X(\hat{\beta}, \hat{\sigma}^2) + 2(p+q+1)n/(n-p-q-2)$ is an approximately unbiased estimator of the expected Kullback-Leibler index $\mathbb{E}_\theta(\triangle(\hat{\theta}|\theta))$ in $\mathbb{E}_\theta(\triangle(\hat{\theta}|\theta))$. Since the preceding calculation (and the maximum likeihood estimators $\hat{\beta}$ 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 AICC($\hat{\beta}$), where

$$AICC(\beta) := -2 \ln L_X(\beta, S_X(\beta)/n) + 2(p+q+1)n/(n-p-q-2).$$

The AIC statistic, defined as

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

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

For fitting autoregressive models, Monte Carlo studies (Jones, 1975; Shibata, 1976) [20] [31] 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) [2] to be

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

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

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 $\{X_1, ..., X_n\}$ 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) [16]. 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) [32] [19]. efficiency is a desirable property defined in terms of the one-step mean square prediction error achieved by the fitted model.

6 Nonstationary and Seasonal Time Series Models

Go back to Table of Contents. Please click TOC

In this chapter we shall examine the problem of finding an appropriate model for a given set of observations $\{x_1, ..., x_n\}$ that are not necessarily generated by a stationary time series.

6.1 ARIMA Models for Nonstationary Time Series

Go back to Table of Contents. Please click TOC

Definition 6.1. If d is a nonnegative integer, then $\{X_t\}$ is an **ARIMA**(p, d, q) **process** if $Y_t := (1 - B)^d X_t$ is a causal ARMA(p, q) process.

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

$$\phi^*(B)X_t \equiv \phi(B)(1 - B)^d X_t = \theta(B)Z_t,\ \{Z_t\} \sim WN(0, \sigma^2),$$

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

Notice that if $d \geq 1$, we can add an arbitrary polynomial trend of degree $(d-1)$ to $\{X_t\}$ without violating the differencing equation $\phi^*(B)X_t$. ARIMA models are therefore useful for representing data with trend (see Sections 1.5 and 6.2).

Example 6.2. $\{X_t\}$ is an ARIMA(1,1,0) process if for some $\phi \in (-1, 1)$,

$$(1 - \phi B)(1 - B)X_t = Z_t,\ \{Z_t\} \sim WN(0, \sigma^2).$$

We can then wrtie

$$X_t = X_0 + \sum_{j=1}^{t} Y_j,\ t \geq 1,$$

where

$$Y_t = (1 - B)X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$

A realization of $\{X_1, ..., X_{200}\}$ with $X_0 = 0$, $\phi = 0.8$, and $\sigma^2 = 1$ is shown in Figure 26, with the corresponding sample autocorrelation and partial autocorrelation functions in Figures 27 and 28, respectively.

□

Figure 26: 200 observations of the ARIMA(1,1,0) series X_t of Example 6.2.

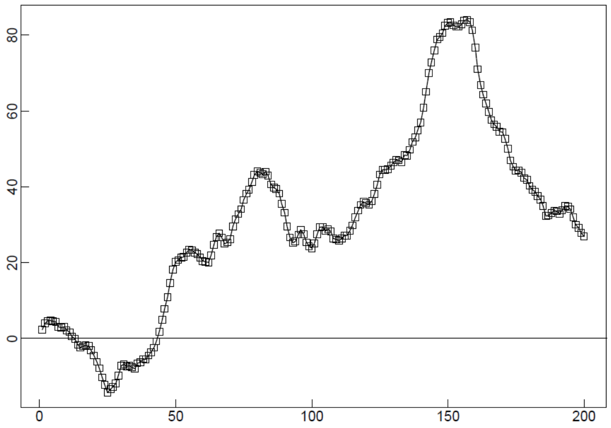


Figure 27: The sample ACF of the data in Figure 26.

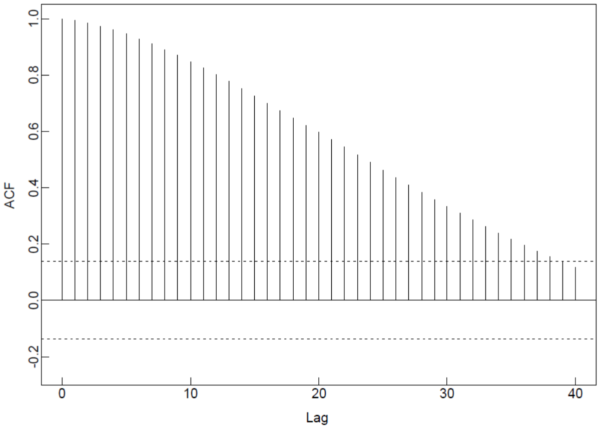
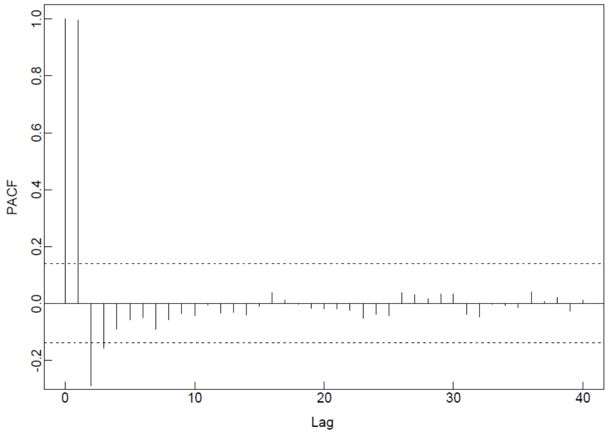


Figure 28: The sample PACF of the data in Figure 26.



6.2 Identification Techniques

Go back to Table of Contents. Please click TOC

The estimation methods of Chapter 5 enable us to find, for given values of p and q, an ARMA(p, q) model to fit a given series of data. For this procedure to be meaningful it must be at least plausible that the data are in fact a realization of an ARMA proceess and in particular a realization of a stationary process.

6.3 Unit Roots in Time Series Models

Go back to Table of Contents. Please click TOC

The unit root problem in time series when either the autoregressive or moving-average polynomial of an ARMA model has a root on or near the unit circle. A unit root in either of these polynomials has important implications for modeling.

6.3.1 Unit Roots in Autoregressions

Go back to Table of Contents. Please click TOC

In section 6.1 we discussed the use of differencing to transform a non-stationary time series with a slowly decaying sample ACF and values near 1 at small lags into one with a rapidly decreasing sample ACF. The degree of differencing of a time series $\{X_t\}$ was largely determined by applying the difference operator repeatedly until the sample ACF of $\{\nabla^d X_t\}$ decays quickly. The differenced time series could then by modeled by a low-order ARMA(p, q) process, and hence the resulting ARIMA(p, d, q) model for the original data has an autoregressive polynomial $(1 - \phi_1 z - \cdots - \phi_p z^p)(1 - z)^d$ (see (6.1.1)) with d roots on the unit circle. In this subsection we discuss a more systematic approach to testing for the presence of a unit root of the autoregressive polynomial in order to decide whether or not a time series should be differenced. This approach was pioneered by Dickey and Fuller (1979) [12].

Let $X_1, ..., X_n$ be observations from the AR(1) model

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + Z_t,\ \{Z_t\} \sim WN(0, \sigma^2),$$

where $|\phi_1| < 1$ and $\mu = \mathbb{E}(X_t)$. For large n, the maximum likelihood estimator $\hat{\phi}_1$ of ϕ_1 is approximately $N(\phi_1, (1-\phi_1^2)/n)$. For the unit root case, this normal approximation is no longer applicable, even asymptotically, which precludes its use for testing the unit root hypothesis $H_0 : \phi_1 = 1$ vs. $H_1; \phi_1 < 1$. To construct a test of H_0, write the model $X_t - \mu$ as

$$\triangledown X_t = X_t - X_{t-1} = \phi_0^* + \phi_1^* X_{t-1} + Z_t,\ \{Z_t\} \sim WN(0, \sigma^2),$$

where $\phi_0^* = \mu(1 - \phi_1)$ and $\phi_1^* = \phi_1 - 1$. Now let $\hat{\phi}_1^*$ be the ordinary least squares (OLS) estimator of ϕ_1^* found by regressing $\triangledown X_t$ on 1 and X_{t-1}. The estimated standard error of $\hat{\phi}_1^*$ is

$$\hat{SE}(\hat{\phi}_1^*) = S\left(\sum_{t=2}^{n}(X_{t-1} - \bar{X})^2\right)^{-1/2},$$

where $S^2 = \sum_{t=2}^{n}(\triangledown X_t - \hat{\phi}_0^* - \hat{\phi}_1^* X_{t-1})^2/(n-3)$ and $\bar{X}$ is the sample mean of $X_1, ..., X_{n-1}$. Dickey and Fuller derived the limit distribution as $n \to \infty$ of the t-ratio

$$\hat{\tau}_\mu := \hat{\phi}_1^*/\hat{SE}(\hat{\phi}_1^*)$$

under the unit root assumption $\phi_1^* = 0$, from which a test of the null hypothesis $H_0 : \phi_1 = 1$ can be constructed. The 0.01, 0.05, and 0.10 quantiles of the limit distribution of $\hat{\tau}_\mu$. The augmented Dickery-Fuller test then rejects the null hypothesis of a unit root, at say, level 0.05 if $\hat{\tau}_\mu < -2.86$. Notice that the cutoff value for this test statistic is much smaller than the standard cutoff value of -1.645 obtained from the normal approximation to the t-distribution, so that the unit root hypothesis is less likely to be rejected using the correct limit distribution.

The above procedure can be extended to the case where $\{X_t\}$ follows the AR(p) model with mean μ given by

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

This model can be rewritten as (see Problem 6.2 of text [9])

$$\triangledown X_t = \phi_0^* + \phi_1^* X_{t-1} + \phi_2^* \triangledown X_{t-1} + \cdots + \phi_p^* \triangledown X_{t-p+1} + Z_t,$$

where $\phi_0 = \mu(1 - \phi_1 - \cdots - \phi_p)$, $\phi_1^* = \sum_{i=1}^{p}\phi_i - 1$, and $\phi_j^* = -\sum_{i=j}^{p}\phi_i$, $j = 2, ..., p$. If the autoregressive polynomial has a unit root at 1, then $0 = \phi(1) = -\phi_1^*$, and the differenced series $\{\triangledown X_t\}$ is an AR($p-1$) process. Consequently, testing the hypothesis of a unit root at 1 of the autoregressive polynomial is equivalent to testing $\phi_1^* = 0$.

6.3.2 Unit Roots in Moving Averages

Go back to Table of Contents. Please click TOC

A unit root in the moving-average polynomial can have a number of interpretations depending on the modeling application. For example, let $\{X_t\}$ be a causal and invertible ARMA(p,q) process astisfying the equations

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

Then the differenced series $Y_t := \triangledown X_t$ is a noninvertible ARMA$(p,q+1)$ process with moving-average polynomial $\theta(z)(1-z)$. Consequently, testing for a unit root in the moving-average polynomial is equivalent to testing that the time series has been overdifferenced.

As a second application, it is possible to distinguish between the competing models

$$\triangledown^k X_t = a + V_t$$

and

$$X_t = c_0 + c_1 t + \cdots + c_k t^k + W_t,$$

where $\{V_t\}$ and $\{W_t\}$ are invertible ARMA processes. For the former model the differenced series $\{\triangledown^k X_t\}$ has no moving-average unit roots, while for the latter model $\{\triangledown^k X_t\}$ has a multiple moving-average unit root of order k. We can therefore distinguish between the two models by using the observed values of $\{\triangledown^k X_t\}$ to test for the presence of a moving-average unit root.

6.4 Forecasting ARIMA Models

Go back to Table of Contents. Please click TOC

In this section we desmonstrate how the methods of Section 3.3 and 5.4 can be adapted to forecast the future values of an ARIMA(p,d,q) process $\{X_t\}$.

If $d \geq 1$, the first and second moments $\mathbb{E}(X_t)$ and $\mathbb{E}(X_{t+h}X_t)$ are not determined by the difference equations $\phi^*(B)X_t \equiv \phi(B)(1-B)^d X_t = \theta(B)Z_t$. We cannot expect, therefore, to determine best linear predictors for $\{X_t\}$ without further assumptions.

In general, we shall assume that our observed process $\{X_t\}$ satisfies the difference equations

$$(1-B)^d X_t = Y_t,\ t = 1,2,...,$$

where $\{Y_t\}$ is a causal ARMA(p,q) process, and that the random vector $(X_{1-d},...,X_0)$ is uncorrelated with $Y_t,\ t > 0$. The difference equations can be rewritten in the form

$$X_t = Y_t - \sum_{j=1}^{d} \binom{d}{j} (-1)^j X_{t-j},\ t = 1,2,...$$

It is convenient,m by relabeling the time axis if necessary, to assume that we observe X_{1-d}, X_{2-d}, ..., X_n or equivalently 1, $X_{1-d},...,X_0,Y_1,...,Y_n)$.

Our goal is to compute the best linear predictors $P_n X_{n+h}$. This can be done by applying the operator P_n to each side of $X_t = Y_t - \sum_{j=1}^{d} \binom{d}{j}(-1)^j X_{t-j}$ (with $t = n+h$) and using the linearity of P_n to obtain

$$P_n X_{n+h} = P_n Y_{n+h} - \sum_{j=1}^{d} \binom{d}{j} (-1)^j P_n X_{n+h-j}.$$

Now the assumption that $(X_{1-d}, ..., X_0)$ is uncorrelated with Y_t, $t > 0$, enables us to identify $P_n Y_{n+h}$ with the best linear predictor of Y_{n+h} in terms of $\{1, Y_1, ..., Y_n\}$, and this can be calculated as described in Section 3.3. The predictor $P_n X_{n+1}$ is obtained directly from $P_n X_{n+h}$ by noting that $P_n X_{n+1-j} = X_{n+1-j}$ for each $j \geq 1$. The predictor $P_n X_{n+2}$ can then be found from $P_n X_{n+h}$ using the previously calculated value of $P_n X_{n+1}$. The predictors $P_n X_{n+3}$, $P_n X_{n+4}$, ... can be computed recursively in the same way.

To find the mean squared error of prediction it is convenient to express $P_n Y_{n+h}$ in terms of $\{X_j\}$. For $n \geq 0$ we denote the one-step predictors by $\hat{Y}_{n+1} = P_n Y_{n+1}$ and $\hat{X}_{n+1} = P_n X_{n+1}$. Then from X_t and $P_n X_{n+h}$ we have

$$X_{n+1} - \hat{X}_{n+1} = Y_{n+1} - \hat{Y}_{n+1}, \ n \geq 1.$$

and hence, if $n > m = \max(p, q)$ and $h \geq 1$, we can write

$$P_n X_{n+h} = \sum_{i=1}^{p} \phi_i P_n Y_{n+h-i} + \sum_{j=h}^{q} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j}).$$

Setting $\phi^*(z) = (1-z)^d \phi 9z) = 1 - \phi_1^* z - \cdots - \phi_{p+d}^* z^{p+d}$, we find from $P_n X_{n+h}$ that

$$P_n X_{n+h} = \sum_{j=1}^{p+d} \phi_j^* P_n X_{n+h-j} + \sum_{j=h}^{q} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j}),$$

which is analogous to the h-step prediction formula $P_n X_{n+h} = \sum_{i=1}^{p} \phi_i P_n X_{n+h-i} + \sum_{j=h}^{q} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j})$ for an ARMA process. The mean squared error of the h-step predictor is

$$\sigma_n^2(h) = \mathbb{E}(X_{n+h} - P_n X_{n+h})^2 = \sum_{j=0}^{h-1} \left(\sum_{r=0}^{j} \chi_r \theta_{n+h-r-1,j-r} \right)^2 v_{n+h-j-1},$$

where $\theta_{n0} = 1$,

$$\chi(z) = \sum_{r=0}^{\infty} \chi_r z^r = (1 - \phi_1^* z - \cdots - \phi_{p+d}^* z^{p+d})^{-1},$$

and

$$v_{n+h-j-1} = \mathbb{E}(X_{n+h-j} - \hat{X}_{n+h-j})^2 = \mathbb{E}(Y_{n+h-j} - \hat{Y}_{n+h-j})^2.$$

The coefficients χ_j can be found from the recursions

$$\chi_j = \sum_{k=1}^{\min(p,j)} \phi_k \chi_{j-k}$$

with ϕ_j^* replacing ϕ_j. For large n we can approximate $\sigma_n^2(h)$, provided that $\theta(\cdot)$ is invertible, by

$$\sigma_n^2(h) = \sum_{j=0}^{j-1} \psi_j^2 \sigma^2,$$

where

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = (\phi^*(z))^{-1}\theta(z).$$

6.4.1 The Forecast Function

Go back to Table of Contents. Please click TOC
Inspection of equation $P_n X_{n+h}$ shows that for fixed $n > m = \max(p, q)$, the h-step predictors

$$g(h) := P_n X_{n+h},$$

satisfy the homogeneous linear difference equations

$$g(h) - \phi_1^* g(h-1) - \cdots - \phi_{p+d}^* g(h-p-d) = 0, \; h > q,$$

where $\phi_1^*, ..., \phi_{p+d}^*$ are the coefficients of $z, ..., z^{p+d}$ in $\phi(z)^* = (1-z)^d \phi(z)$. The solution is well known from the theory of linear difference equations. If we assume that the zeros of $\phi(z)$ (denoted by $\xi_1, ..., \xi_p$) are all distinct, then the solution is

$$g(h) = a_0 + a_1 h + \cdots + a_d h^{d-1} + b_1 \xi_1^{-h} + \cdots + b_p \xi_p^{-h}, \; h > q - p - d,$$

where the coefficients $a_1, ..., a_d$ and $b_1, ..., b_p$ can be determined from the $p+d$ equations obtained by equating the right-handside of $g(h)$ for $q-p-d < h \leq q$ with the corresponding value of $g(h)$ computed numerically (for $h \leq 0$, $P_n X_{n+h} = X_{n+h}$, and for $1 \leq h \leq q$, $P_n X_{n+h}$ can be computed from $P_n X_{n+h}$ as already described). Once the constants a_i and b_i have been evaualted, the algebraic expression $g(h)$ gives the predictors for all $a_0, ..., a_d, b_1, ..., b_p$ are simply the *observed* values $g(h) = X_{n=h}$, $-p-d \leq h \leq 0$, and the expression $\sigma^2(h)$ for the mean sqaured error is exact.

The calculation of the forecast function is easily generalized to deal with more complicated ARIMA processes.

6.5 Seasonal ARIMA Models

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

Definition 6.3. If d and D are nonnegative integers, then $\{X_t\}$ is a **sesonal ARIMA$(p, d, q) \times (P, D, Q)_s$ process with period** s if the differenced series $Y_t = (1-B)^d(1-B^s)^D X_t$ is a causal ARMA process defined by

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

where $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$, $\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_p z^p$, $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$, and $\Theta(z) = 1 + \Theta_1 z + \ldots \Theta_Q z^Q$.

Remark 6.4. Note that the process $\{Y_t\}$ is causal if and only if $\phi(z) \neq 0$ and $\Phi(z) \neq 0$ for $|z| \leq 1$. In applications D is rarely more than one, and P and Q are typically less than three.

□

Remark 6.5. The equation $\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t$ satisfied by the differenced process $\{Y_t\}$ can be rewritten in the equivalent from

$$\phi^*(B)Y_t = \theta^*(B)Z_t,$$

where $\phi^*(\cdot)$, $\theta^*(\cdot)$ are polynomials of degree $p+sP$ and $q+sQ$, respectively, whose coefficients can all be expressed in terms of $\phi_1, ..., \phi_p, \Phi_1, ..., \Phi_p, \theta_1, ..., \theta_q$, and $\Theta_1, ..., \Theta_Q$. Provided that $p < s$ and $q < s$, the constraints on the coefficients of $\phi^*(\cdot)$ and $\theta^*(\cdot)$ can all be expressed as multiplicative relations

$$\phi^*_{is+j} = \phi^*_{is}\phi^*_j,\ i = 1, 2, ...; j = 1, ..., s-1,$$

and

$$\theta^*_{is+j} = \theta^*_{is}\theta^*_j,\ i = 1, 2, ...; j = 1, ..., s-1.$$

In Section 1.5 we discussed the classical decomposition model incorporating trend, seasonality, and random noise, namely, $X_t = m_t + s_t + Y_t$. In modeling real data it might not be reasonable to assume, as in the classical decomposition model, that the seasonal component s_t repeats itself precisely in the same way cycle after cycle. Seasonal ARIMA models allow for randomness in the seasonal pattern from one cycle to the next.

□

6.6 Regression with ARIMA Errors

Go back to Table of Contents. Please click TOC

6.6.1 OLS and GLS Estimation

Go back to Table of Contents. Please click TOC

In standard linear regression, the errors (or deviations of the observations from the regression function) are assumed to be independent are identically distributed. In many applications of regression analysis, however, this assumption is cleraly violated, as can be seen by examination of the residuals from the fitted regression and their sample autocorrelations. It is often more appropriate to assume that the errors are observations of a zero-mean second-order stationary process. Since many autocorrelation functions can be well approximated by the autocorrelation function of a suitably chosen ARMA(p, q) process, it is of particular interest to consider the model

$$Y_t = \mathbf{x}_t'\beta + W_t,\ t = 1, ..., n,$$

or in matrix notatio,

$$\mathbf{Y} = X\beta + \mathbf{W},$$

where $\mathbf{Y} = (Y_1, ..., Y_n)'$ is the vector of observations at times $t = 1, ..., n$, X is the design matrix whose tth row, $\mathbf{x}_t' = (x_{t1}, ..., x_{tk})$, consists of the values of the explanatory variables at time t, $\beta = (\beta_1, ..., \beta - k)'$ is the vector of regression coefficients, and the components of $\mathbf{W} = (W_1, ..., W_n)'$ are values of a causal zero-mean ARMA(p, q) process satisfying

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

The model Y_t arises naturally in trend estimation for time series data.

The **ordinary least squares** (OLS) estimator of β is the value, $\hat{\beta}_{OLS}$, which minimizes the sum of squares

$$(\mathbf{Y} - X\beta)'(\mathbf{Y} - X\beta) = \sum_{t=1}^{n}(Y_t - \mathbf{x}_t'\beta)^2.$$

Equating to zero the partial derivatives with respect to each component of β and assuming (as we shall) that $X'X$ is nonsingular, we find that

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'\mathbf{Y}.$$

(If $X'X$ is singular, $\hat{\beta}_{OLS}$ is not uniquely determined but still satisfies $\hat{\beta}_{OLS}$ with $(X'X)^{-1}$ any generalized inverse of $X'X$.) The OLS estimate also maximizes the likelihood of the observations when the errors $W_1, ..., W_n$ are iid and Gaussian. If the design matrix X is nonrandom, then even when the errors are non-Gaussian and dependent, the OLS estimator is unbiased (ie., $\mathbb{E}(\hat{\beta}_{OLS}) = \beta$) and its covariance matrix is

$$Cov(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\Gamma_n X(X'X)^{-1},$$

where $\Gamma_n = \mathbb{E}(\mathbf{W}\mathbf{W}')$ is the covariance matrix of $\mathbf{W}$.

The **generalized least squares** (GLS) estimator of β is the value $\hat{\beta}_{GLS}$ that minimizes the *weighted* sum of squares

$$(\mathbf{Y} - X\beta)'\Gamma_n^{-1}(\mathbf{Y} - X\beta).$$

Differentiating partially with respect to each component of β and setting the derivatives equal to zero, we find that

$$\hat{\beta}_{GLS} = (X'\Gamma_n^{-1}X)^{-1}X'\Gamma_n^{-1}\mathbf{Y}.$$

If the design matrix X is nonrandom, the GLS estimator is unbiased and has covariance matrix

$$Cov(\hat{\beta}_{GLS}) = (X'\Gamma_n^{-1}X)^{-1}.$$

It can be shown that the GLS estimator is the best linear unbiased estimator of β, i.e., for any k-dimensional vector **c**and for any unbiased estimator $\hat{\beta}$ of β that is a linear function of the observations $Y_1, ..., Y_n$,

$$Var(\mathbf{c}'\hat{\beta}_{GLS}) \le Var(\mathbf{c}'\hat{\beta}).$$

In this sense the GLS estimator is therefore superior to the OLS estimator. However, it can be computed only if ϕ and θ are known.

Let $V(\phi, \theta)$ denote the matrix $\sigma^{-2}\Gamma_n$ and let $T(\phi, \theta)$ be any square root of V^{-1} (i.e., a matrix such that $T'T = V^{-1}$). Then we can multiply each side of $\mathbf{Y} = X\beta + \mathbf{W}$, by T to obtain

$$T\mathbf{Y} = TX\beta + T\mathbf{W},$$

a regression equation with coefficient vector β, data vector $T\mathbf{Y}$, design matrix TX, and error vector $T\mathbf{W}$. Since the latter has uncorrelated, zero-mean components, each with variance σ^2, the best linear estimator of β interms of $T\mathbf{Y}$ (which is clearly the same as the best linear estimator of β in terms of $\mathbf{Y}$, i.e., $\hat{\beta}_{GLS}$) can be obtained by applying OLS estimaton to the transformed regression equation $T\mathbf{Y}$. This gives

$$\hat{\beta}_{GLS} = (X'T'TX)^{-1}X'T'T\mathbf{Y},$$

which is clearly the same as $\hat{\beta}_{GLS}$. Cochrane and Orcutt (1949) [11] pointed out that if $\{W_t\}$ is an AR(p) process satisfying

$$\phi(B)W_t = Z_t, \ \{Z_t\} \sim WN(0, \sigma^2),$$

then application of $\phi(B)$ to each side of the regression equations $Y_t = \mathbf{x}_t'\beta + W_t$transforms them into regression equations with uncorrelated, zero-mean, constant-variance errors, so that ordinary least squares can again be used to compute best linear unbiased estimates of the components of β in terms of $Y_t^* = \phi(B)Y_t$, $t = p+1, ..., n$. This approach eliminates the need to compute the matrix T but suffers from the drawback that $\mathbf{Y}^*$ does not contain all the information in $\mathbf{Y}$. Cochrane and Orcutt's transformation can be improved, and at the same generalized to ARMA errors, as follows.

Instead of applying the operator $\phi(B)$ to each side of the regression equations $Y_t = \mathbf{x}_t'\beta + W_t$, we multiply each side of equation $\mathbf{Y} = X\beta + \mathbf{W}$by the matrix $T(\phi, \theta)$ that maps $\{W_t\}$ into the residuals of $\{W_t\}$ from the ARMA model, $\phi(B)W_t = \theta(B)Z_t$. To see that T is a square root of the matrix V as defined in the previous paragraph, we simply recall that the residuals are uncorrelated with zero mean and variance σ^2, so that

$$Cov(T\mathbf{W}) = T\Gamma_n T' = \sigma^2 I,$$

where I is the $n \times n$ identify matrix. Hence

$$T'T = \sigma^2\Gamma_n^{-1} = V^{-1}.$$

GLS estimation of β can therefore be carried out by multiplying each side of $\mathbf{Y} = X\beta + \mathbf{W}$ by T and applying ordinary least squares to the transformed regression model. It remains only to compute $T\mathbf{Y}$ and TX.

6.6.2 ML Estimation

Go back to Table of Contents. Please click TOC

If (as is usually the case) the parameters of the ARMA(p, q) model for the errors are unknown, they can be estimated together with the regression coefficients by **maximizing the Gaussian likelihood**

$$L(\beta,\phi,\theta,\sigma^2) = (2\pi)^{-n/2}(\det\Gamma_n)^{-1/2}\exp\bigg\{-\frac{1}{2}(\mathbf{Y}-X\beta)'\Gamma_n^{-1}(\mathbf{Y}-X\beta)\bigg\},$$

where $\Gamma_n(\phi,\theta,\sigma^2)$ is the covariance matrix of $\mathbf{W}=\mathbf{Y}-X\beta$. Since $\{W_t\}$ is an ARMA(p,q) process with parameters ϕ,θ,σ^2), the maximum likelihood estimators $\hat\beta$, $\hat\phi$, $\hat\theta$ are found by minimizing

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

where

$$S(\beta,\phi,\theta) = \sum_{t=1}^{n}(W_t-\hat W_t)^2/r_{t-1},$$

$\hat W_t$ is the best one-step predictor of W_t, and $r_{t-1}\sigma^2$ is its mean squared error. The function $\ell(\beta,\phi,\theta)$ can be expressed in terms of the observations $\{Y_t\}$ and the parameters β, ϕ, and θ using the innovations algorithm (see Section 3.3) and minimized numerically to give the maximum likelihood estimators, $\hat\beta$, $\hat\phi$, and $\hat\theta$. The maximum likelihood estimator of σ^2 is then given, as in Section 5.2, by $\hat\sigma^2 = S(\hat\beta,\hat\phi,\hat\theta)/n$.

An extention of an iterative scheme, proposed by Cochrane and Orcutt (1949) for the case $q=0$, simplifies the minimization considerably [**?**]. It is based on the observation that for fixed ϕ and θ, the value of β that minimizes $\ell(\beta,\phi,\theta)$ is $\hat\beta_{GLS}(\phi,\theta)$, which can be computed algebraically from $\hat\beta_{GLS}$ instead of by searching numerically for the minimizing value. The scheme is as follows.

(i) Compute $\hat\beta_{OLS}$ and the estimated residuals $Y_t - \mathbf{t}'\hat\beta_{OLS}$, $t=1,...,n$.

(ii) Fit an ARMA(p,q) model by maximum Gaussian likelihood to the estimated residuals.

(iii) For the fitted ARMA model compute the corresponding estimator $\hat\beta_{GLS}$ from $\hat\beta_{GLS}$.

(iv) Compute the residuals $Y_t - \mathbf{x}_t'\hat\beta_{GLS}$, $t=1,...,n$, and return to (ii), stopping when the estimators have stablized.

If $\{W_t\}$ is a causal and invertible ARMA process, then under mild conditions on the explanatory variables $\mathbf{x}_t$, the maximum likelihood estimates are asymptotically independent of the estimated ARMA parameters.

The large-sample covariance matrix of the ARMA parameter estimators, suitably normalized, has a complicated form that involves both the regression variables $\mathbf{x}_t$ and the covariance function of $\{W_t\}$. It is therefore convenient to estimate the covariance matrix as $-H^{-1}$, where H is teh Hessian matrix of the observed log-likelihood evaluated at its maximum.

The OLS, GLS, and maximum likelihood estimators of the regression coefficients all have the same aymptotic covariance matrix, so in this sense the dependence does not play a major role. However, the asymptotic covariance of both the OLS and GLS estimators can be very inaccurate if the appropriate covariance matrix Γ_n is not used in the expressions

$$Cov(\hat\beta_{OLS}) = (X'X)^{-1}X'\Gamma_nX(X'X)^{-1},$$

and

$$Cov(\hat{\beta}_{GLS}) = (X'\Gamma_n^{-1}X)^{-1}.$$

7 Multivariate Time Series

Go back to Table of Contents. Please click TOC

Many time series arising in practice are best considered as components of some vector-valued (multivariate) time series $\{\mathbf{X}_t\}$ having not only serial dependence within each component series $\{X_{ti}\}$ but also interdependence between the difference component series $\{X_{ti}\}$ and $\{X_{tj}\}$, $i \neq j$. Much of the theory of univariate time series extends in a natural way to the multivariate case; however, new problems arise. in this chapter we introduce the basic properties of multivariate series and consider the multivariate extensions of some of the techniques developed earlier. In Section 7.1 we introduce two sets of bivariate time series data for which we develop multivariate models later in the chapter. in Section 7.2 we discuss the basic properties of stationary multivariate time series, namely, the mean vector $\mu = \mathbb{E}(\mathbf{X}_t)$ and the covariance matrices $\Gamma(h) = \mathbb{E}(\mathbf{X}_{t+h}\mathbf{X}_t') - \mu\mu'$, $h = 0, \pm1, \pm2, ...$, with reference to some simple examples, including multivariate white noise. Section 7.3 deals with estimation of μ and $\Gamma(\cdot)$ and the question of testing for serial independence on the basis of observations of $\mathbf{X}_1, ..., \mathbf{X}_n$. In Section 7.4 we introduce multivariate ARMA processes and illustrate the problem of multivariate model identification with an example of a multivariate AR(1) process that also has an MA(1) representation. The identification problem can be avoided by confining attention to multivariate autoregressive (or VAR) models. Forecasting multivariate time series with known second-order properties is discussed in Section 7.5, and in Section 7.6 we consider the modeling and forecasting of multivariate time series using the multivariate Yule-Walker equations and Whittle's generalization of the Durbin-Levinson algorithm. Section 7.7 contains a brief introduction to the notion of cointegrated time series.

7.1 Examples

Go back to Table of Contents. Please click TOC

In this section we introduce two examples of bivariate time series. A bivariate time series is a series of two-dimensional vectors $(X_{t1}, X_{t2})'$ observed at times t (usually $t = 1, 2, 3, ...$). The two component series $\{X_{t1}\}$ and $\{X_{t2}\}$ could be studied independently as univariate time series, each characterized, from a second-order point of view, by its own mean and autocovariance function. Such an approach, however, fails to take into account possible dependence *between* the two component series, and such cross-dependence may be of great importance, for example in predicting future values of the two component series.

We therefore consider the series of random vectors $\mathbf{X}_t = (X_{t1}, X_{t2})'$ and define the mean vector

$$\mu_t := \mathbb{E}(\mathbf{X}_t) = \begin{bmatrix} \mathbb{E}(X_{t1}) \\ \mathbb{E}(X_{t2}) \end{bmatrix}$$

and covariance matrices

$$\Gamma(t+h,t) := Cov(\mathbf{X}_{t+h}, \mathbf{X}_t) = \begin{bmatrix} cov(X_{t+h,1}, X_{t1}) & cov(X_{t+h,1}, X_{t2}) \\ cov(X_{t+h,2}, X_{t1}) & cov(X_{t+h,2}, X_{t2}) \end{bmatrix}.$$

The bivariate series $\{\mathbf{X}_t\}$ is said to be **(weakly) stationary** if the moments μ_t and $\Gamma(t+h,t)$ are both independent of t, in which case we use the notation

$$\mu = \mathbb{E}(\mathbf{X}_t) = \begin{bmatrix} \mathbb{E}(X_{t1}) \\ \mathbb{E}(X_{t2}) \end{bmatrix}$$

and

$$\Gamma(h) = Cov(\mathbf{X}_{t+h}, \mathbf{X}_t) = \begin{bmatrix} \gamma_{11}(h) & \gamma_{12}(h) \\ \gamma_{21}(h) & \gamma_{22}(h) \end{bmatrix}.$$

The diagonal elements are the autocovariance functions of the univariate series $\{X_{t1}\}$ and $\{X_{t2}\}$ as defined in Chapter 2, while the off-diagonal elements are the covariances between $X_{t+h,i}$ and X_{tj}, $i \neq j$. Notice that $\gamma_{12}(h) = \gamma_{21}(-h)$.

A natural estimator of the mean vector μ in terms of the observations $\mathbf{X}_1, ..., \mathbf{X}_n$ is the vector of sample means

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t,$$

and a natural estimator of $\Gamma(h)$ is

$$\hat{\Gamma}(h) = \begin{cases} n^{-1}\sum_{t=1}^{n-h}(\mathbf{X}_{t+h} - \bar{\mathbf{X}}_n)(\mathbf{X}_t - \bar{\mathbf{X}}_n)', & for\ 0 \leq h \leq n-1, \\ \hat{\Gamma}(-h)', & for\ -n+1 \leq h < 0. \end{cases}$$

The correlation $\rho_{ij}(h)$ between $X_{t+h,i}$ and $X_{t,j}$ is estimated by

$$\hat{\rho}_{ij}(h) = \hat{\gamma}_{ij}(h)(\hat{\gamma}_{ij}(0)\hat{\gamma}_{jj}(0))^{-1/2}.$$

If $i = j$, then $\hat{\rho}_{ij}$ reduces to the sample autocorrelation function of the ith series. These estimators will be discussed in more detail in Section 7.2.

7.2 Second-Order Properties of Multivariate Time Series

Go back to Table of Contents. Please click TOC

Consider m time series $\{X_{ti}, t = 0, \pm 1, ..., \}$ $i = 1, ..., m$, with $\mathbb{E}(X_{ti}^2 < \infty$ for all t and i. If all the finite-dimensional distributions of the random variables $\{X_{ti}\}$ were multivariate normal, then the distributional properties of $\{X_{ti}\}$ would be completely determined by the means

$$\mu_{ti} := \mathbb{E}(X_{ti})$$

and the covariances

$$\gamma_{ij}(t+h,t) := \mathbb{E}[(X_{t+h,i} - \mu_{ti})(X_{tj} - \mu_{tj})].$$

Even when the observations $\{X_{ti}\}$ do not have joint normal distributions, the quantities μ_{ti} and $\gamma_{ij}(t+h,t)$ spcify the second-order properties, the covariances providing us with a measure of the dependence, not only between observations in the same series, but also between the observations in different series.

It is more convenient in dealing with m interrelated series to use vector notation. Thus, we define

$$\mathbf{X}_t := \begin{bmatrix} X_{t1} \\ \vdots \\ X_{tm} \end{bmatrix}, t = 0, \pm 1, ...$$

The second-order properties of the multivariate time series $\{\mathbf{X}_t\}$ are then specified by the mean vectors

$$\mu_t := \mathbb{E}(\mathbf{X}_t) = \begin{bmatrix} \mu_{t1} \\ \vdots \\ \mu_{tm} \end{bmatrix}$$

and covariance matrices

$$\Gamma(t+h,t) := \begin{bmatrix} \gamma_{11}(t+h,t) & \dots & \gamma_{1m}(t+h,t) \\ \vdots & \ddots & \vdots \\ \gamma_{m1}(t+h,t) & \dots & \gamma_{mm}(t+h,t) \end{bmatrix},$$

where

$$\gamma_{ij}(t+h,t) := Cov(X_{t+h,i}, X_{t,j}).$$

Remark 7.1. The matrix $\Gamma(t+h,t)$ can also be expressed as

$$\Gamma(t+h,t) := \mathbb{E}[(\mathbf{X}_{t+h} - \mu_{t+h})(\mathbf{X}_t - \mu_t)'],$$

where as usual, the expected value of a random matrix A is the matrix whose components are the expected values of the components of A.

□

As in the univariate case, a particularly important role is played by the class of **multivariate stationary time series**, defined as follows.

Definition 7.2. The m-variate series $\{\mathbf{X}_t\}$ is (**weakly**) **stationary** if

(i) $\mu_X(t)$ is independent of t, and

(ii) $\Gamma_X(t+h,t)$ is independent of t for each h.

For a stationary time series we shall use the notation

$$\mu := \mathbb{E}(\mathbf{X}_t) = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{bmatrix}$$

and

$$\Gamma(h) := \mathbb{E}[(\mathbf{X}_{t+h} - \mu)(\mathbf{X}_t - \mu)'] = \begin{bmatrix} \gamma_{11}(h) & \dots & \gamma_{1m}(h) \\ \vdots & \ddots & \vdots \\ \gamma_{m1}(h) & \dots & \gamma_{mm}(h) \end{bmatrix}.$$

We shall refer to μ as the mean of the series and to $\Gamma(h)$ as the covariance matrix at lag h. Notice that if $\{\mathbf{X}_t\}$ is stationary with covariance matrix function $\Gamma(\cdot)$, then for each i, $\{X_{ti}\}$ is stationary with covariance function $\gamma_{ii}(\cdot)$. The function $\gamma_{ij}(\cdot)$, $i \neq j$, is called the cross-covariance function of the two series $\{X_{ti}\}$ and $\{X_{tj}\}$. It should be noted that $\gamma_{ij}(\cdot)$ is not in general the same as $\gamma_{ji}(\cdot)$. The correlation matrix function $R(\cdot)$ is defined by

$$R(h) := \begin{bmatrix} \rho_{11}(h) & \dots & \rho_{1m}(h) \\ \vdots & \ddots & \vdots \\ \rho_{m1}(h) & \dots & \rho_{mm}(h) \end{bmatrix},$$

where $\rho_{ij}(h) = \gamma_{ij}(h)/[\gamma_{ii}(0)\gamma_{jj}(0)]^{1/2}$. The function $R(\cdot)$ is the covariance matrix function of the normalized series obtained by subtracting μ from $\mathbf{X}_t$ and then dividing each component by its standard deviation.

Proposition 7.3. ***Basic Properties of*** $\Gamma(\cdot)$***:***

(1) $\Gamma(h) = \Gamma'(-h)$,

(2) $|\gamma_{ij}(h)| \leq [\gamma_{ii}(0)\gamma_{jj}(0)]^{1/2}$, $i, j, = 1, ..., m$,

(3) $\gamma_{ii}(\cdot)$ *is an autocovariance function,* $i = 1, ..., m$, *and*

(4) $\sum\limits_{j,k=1}^{n} \boldsymbol{a}_j' \Gamma(j-k) \boldsymbol{a}_k \geq 0$ *for all* $n \in \{1, 2, ...\}$ *and* $\boldsymbol{a}_1, ..., \boldsymbol{a}_n \in \mathbb{R}^m$.

Proof: The first property follows at once from the definition, the second from the fact that correlations cannot be greater than one in absolute value, and the third from the observations that $\gamma_{ii}(\cdot)$ is the autocovariance function of the stationary series $\{X_{ti}, t = 0, \pm 1, ...,\}$. Property 4 is statement of the obvious fact that

$$\mathbb{E}\left(\sum_{j=1}^{n} \mathbf{a}_j' (\mathbf{X}_j - \mu)\right)^2 \geq 0.$$

Q.E.D.

Remark 7.4. The basic properties of the matrices $\Gamma(h)$ are shared also by the coresponding matrices of correlations $R(h) = [\rho_{ij}(h)]_{i,j=1}^{m}$, which have the additional property

$$\rho_{ii}(0) = 1, \ \forall i.$$

The correlation $\rho_{ij}(0)$ is the correlation between X_{ti} and X_{tj}, which is generally not equal to 1 if $i \neq j$. It is also possible that $|\gamma_{ij}(h)| > |\gamma_{ij}(0)|$ if $i \neq j$.

Definition 7.5. The m-variate series $\{\mathbf{Z}_t\}$ is called **white noise with mean 0 and covariance matrix** $\bar{\Sigma}$, written

$$\{\mathbf{Z}_t\} \sim WN(0, \bar{\Sigma}),$$

if $\{\mathbf{Z}_t\}$ is stationary with mean vector 0 and covariance matrix function

$$\Gamma(h) = \begin{cases} \bar{\Sigma}, & \textit{if } h = 0, \\ 0, & \textit{otherwise.} \end{cases}$$

Definition 7.6. The m-variate series $\{\mathbf{Z}_t\}$ is called **iid noise with mean 0 and covariance matrix** $\bar{\Sigma}$, written

$$\{\mathbf{Z}_t\} \sim iid(0, \bar{\Sigma}),$$

if the random vectors $\{\mathbf{Z}_t\}$ are independent are identically distributed with mean 0 and covariance matrix $\bar{\Sigma}$.

Definition 7.7. The m-variate series $\{\mathbf{X}_t\}$ is a **linear process** if it has the representation

$$\mathbf{X}_t = \sum_{j=-\infty}^{\infty} C_j \mathbf{Z}_{t-j}, \ \{\mathbf{Z}_t\} \sim WN(\mathbf{0}, \bar{\Sigma}),$$

where $\{C_j\}$ is a sequence of $m \times m$ matrices whose components are absolutely summable.

The linear process $\mathbf{X}_t$ is stationary with mean $\mathbf{0}$ and covariance function

$$\Gamma(h) = \sum_{j=-\infty}^{\infty} C_{j+h} \bar{\Sigma} C_j', \ h = 0, \pm 1, ...$$

An **MA(∞) proceess** is a linear process with $C_j = 0$ for $j < 0$. Thus $\{\mathbf{X}_t\}$ is an MA(∞) process if and only if there exists a white noise sequence $\{\mathbf{Z}_t\}$ and a sequence of matrices C_j with absolutely summable components such that

$$\mathbf{X}_t = \sum_{j=0}^{\infty} C_j \mathbf{Z}_{t-j}.$$

Multivariate ARMA processes will be discussed in Section 7.4, where it will be shown in particular that any causal ARMA(p, q) process can be expressed as an MA(∞) process, while any invertible ARMA(p, q) process can be expressed as an **AR(∞) process**, i.e., a process satisfying equations of the form

$$\mathbf{X}_t + \sum_{j=1}^{\infty} A_j \mathbf{X}_{t-j} = \mathbf{Z}_t,$$

in which the matrices A_j have absolutely summable components.

Second-Order Properties in the Frequency Domain

Provided that the components of the covariance matrix function $\Gamma(\cdot)$ have the property $\sum\limits_{h=-\infty}^{\infty} |\gamma_{ij}(h)| < \infty$, $i, j = 1, ..., m$, then Γ has a **matrix-valued spectral density function**

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h} \Gamma(h), \ -\pi \leq \lambda \leq \pi,$$

and Γ can be expressed in terms of f as

$$\Gamma(h) = \int_{-\pi}^{\pi} e^{i\lambda h} f(\lambda) d\lambda.$$

The second-order properties of the stationary process $\{\mathbf{X}_t\}$ can therefore be described equivalently in terms of $f(\cdot)$ rather than $\Gamma(\cdot)$. Similarly, $\{\mathbf{X}_t\}$ can therefore be described equivalently in terms of $f(\cdot)$ rather than $\Gamma(\cdot)$, Similarly, $\{\mathbf{X}_t\}$ has a **spectral representation**

$$\mathbf{X}_t = \int_{-\pi}^{\pi} e^{i\lambda t} d\mathbf{Z}(\lambda),$$

where $\{\mathbf{Z}(\lambda),\ -\pi \leq \lambda \leq \pi\}$ is process whose components are complex-valued processes satisfying

$$\mathbb{E}(dZ_j(\lambda)d\bar{Z}_k(\mu)) = \begin{cases} f_{jk}(\lambda)d\lambda, & if\ \lambda = \mu, \\ 0, & if\ \lambda \neq \mu, \end{cases}$$

and $\bar{Z}_k$ denotes the complex conjugate of Z_k.

7.3 Estimation of the Mean and Covariance Function

Go back to Table of Contents. Please click TOC

As in the univariate case, the estimation of the mean vector and covariances of a stationary multivariate time series plays an important role in describing and modeling the dependence structure of the component series. In this section we introduce estimators, for a stationary m-variate time series $\{\mathbf{X}_t\}$, of the comonents μ_j, $\gamma_{ij}(h)$, and $\rho_{ij}(h)$ of μ, $\Gamma(h)$, and $R(h)$, respectively. We also exmaine the large-sample properties of these estimators.

7.3.1 Estimation of μ

Go back to Table of Contents. Please click TOC

A natural unbiased estimator of the mean vector μ based on the observations $\mathbf{X}_1, ..., \mathbf{X}_n$ is the vector of sample means

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbf{X}_t.$$

The resulting estimate of the mean of the jth time series is then the univariate sample mean $(1/n)\sum_{t=1}^{n} X_{tj}$. If each of the univariate autocovariance functions $\gamma_{ii}(\cdot)$, $i = 1, ..., m$, satisfies the conditions of Proposition 2.15 (see remark), then the consistency of the estimator $\bar{\mathbf{X}}_n$ can be established by applying the proposition to each of the component time series $\{X_{ti}\}$. This immediately gives the following proposition.

Remark 7.8. Recall Proposition 2.15: If $\{X_t\}$ is a stationary time series with mean μ and autocovariance function $\gamma(\cdot)$, then as $n \to \infty$,

$$Var(\bar{X}_n) = \mathbb{E}(\bar{X}_n - \mu) \to 0\ if\ \gamma(n) \to 0,$$

$$n\mathbb{E}(\bar{X}_n - \mu)^2 \to \sum_{|h|<\infty} \gamma(h)\ if\ \sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty.$$

□

Proposition 7.9. *If $\{\boldsymbol{X}_t\}$ is a stationary multivariate time series with mean μ and covariance function $\Gamma(\cdot)$, then as $n \to \infty$,*

$$\mathbb{E}(\bar{\boldsymbol{X}}_n - \mu)'(\bar{\boldsymbol{X}}_n - \mu) \to 0 \ if \ \gamma_{ii}(n) \to 0, \ 1 \leq i \leq m,$$

and

$$n\mathbb{E}(\bar{\boldsymbol{X}}_n - \mu)'(\bar{\boldsymbol{X}}_n - \mu) \to \sum_{i=1}^{m} \sum_{h=-\infty}^{\infty} \gamma_{ii}(h) \ if \ \sum_{h=-\infty}^{\infty} |\gamma_{ii}(h)| < \infty, \ 1 \leq i \leq m.$$

Under more restrictive assumptions on the process $\{\mathbf{X}_t\}$ it can also be shown that $\bar{\mathbf{X}}_n$ is approximately normally distributed for large n. Determination of the covariance matrix of this distribution would allow us to obtain confidence regions for μ. However, this is quite complicated, and the following simple approximation is useful in practice.

For each i we construct a confidence interval for μ_i based on the sample mean $\bar{X}_i$ of the univariate series $X_{1i}, ..., X_{ti}$ and combine these to form a confidence region for μ. If $f_i(\omega)$ is the spectral density of the ith process $\{X_{ti}\}$ and if the sample size n is large, then we know, under the same conditions as in Section 2.4, that $\sqrt{n}(\bar{X}_i - \mu_i)$ is approximately normally distributed with mean zero and variance

$$2\pi f_i(0) = \sum_{k=-\infty}^{\infty} \gamma_{ii}(k).$$

t can also be shown (see, e.g., Anderson, 1971) [3] that

$$2\pi \hat{f}_i(0) := \sum_{|h| \leq r} \left(1 - \frac{|h|}{r}\right) \hat{\gamma}_{ii}(h)$$

is consistent estimator of $2\pi f_i(0)$, provided that $r = r_n$ is a sequence of numbers depending on n in such a way that $r_n \to \infty$ and $r_n/n \to 0$ as $n \to \infty$. Thus if $\bar{X}_i$ denotes the sample mean of the ith process and Φ_α is the α-quantile of the standard normal distribution, then the bounds

$$\bar{X}_i \pm \Phi_{1-\alpha/2}(2\pi \hat{f}_i(0)/n)^{1/2}$$

are asymptotic $(1 - \alpha)$ confidence bounds for μ_i. Hence

$$\begin{aligned} P\bigg(|\mu_i - \bar{X}_i| &\leq \Phi_{1-\alpha/2}(2\pi \hat{f}_i(0)/n)^{1/2}, \ i = 1, ..., m\bigg) \\ &\geq 1 - \sum_{i=1}^{m} P\bigg(|\mu_i - \bar{X}_i| > \Phi_{1-\alpha/2}(2\pi \hat{f}_i(0)/n)^{1/2}\bigg). \end{aligned}$$

where the right-handside converges to $1 - m\alpha$ as $n \to \infty$. Consequently, as $n \to \infty$, the set of m-dimensional vectors bounded by

$$\left\{ x_i = \bar{X}_i \pm \Phi_{1-(\alpha/(2m))}(2\pi \hat{f}_i(0)/n)^{1/2}, \ i = 1, ..., m \right\}$$

has a confidence coefficient that converges to a value *greater than or equal to* $1 - \alpha$ (and substantially greater if m is large). Neverthelss, the region defined above is easy to determine and is of reasonable size, provided that m is not too large.

7.3.2 Estimation of $\Gamma(h)$

Go back to Table of Contents. Please click TOC

As in the univariate case, a natural estimator of the covariance $\Gamma(h) = \mathbb{E}[(\mathbf{X}_{t+h} - \mu)(\mathbf{X}_t - \mu)']$ is

$$\hat{\Gamma}(h) = \begin{cases} n^{-1} \sum\limits_{t=1}^{n=h} (\mathbf{X}_{t+h} - \bar{X}_n)(\mathbf{X}_t - \bar{\mathbf{X}}_n)' & \textit{for } 0 \leq h \leq n-1, \\ \hat{\Gamma}'(-h) & \textit{for } -n+1 \leq h < 0. \end{cases}$$

Writing $\hat{\gamma}_{ij}(h)$ for the (i, j)-component of $\hat{\Gamma}(h)$, $i, j = 1, 2, ...$, we estimate the cross-correlations by

$$\hat{\rho}_{ij}(h) = \hat{\gamma}_{ij}(h)(\hat{\gamma}_{ii}(0)\hat{\gamma}_{jj}(0))^{-1/2}.$$

If $i = j$, then $\hat{\rho}_{ij}$ reduces to the sample autocorrelation function of the ith series.

Theorem 7.10. *Let $\{\boldsymbol{X}_t\}$ be the bivariate time series whose components are defined by*

$$X_{t1} = \sum_{k=-\infty}^{\infty} \alpha_k Z_{t-k,1}, \ \{Z_{t1}\} \sim IID(0, \sigma_1^2),$$

and

$$X_{t2} = \sum_{k=-\infty}^{\infty} \beta_k Z_{t-k,2}, \ \{Z_{t2}\} \sim IID(0, \sigma_2^2),$$

where the two sequences $\{Z_{t1}\}$ and $\{Z_{t2}\}$ are independent, $\sum_k |\alpha_k| < \infty$, and $\sum_k |\beta_k| < \infty$.

*Then for all integers h and k with $h \neq k$, the random variables $n^{1/2}\hat{\rho}_{12}(h)$ and $n^{1/2}\hat{\rho}_{12}(k)$ are approximately bivariate normal with mean **0**, variance $\sum\limits_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j)$, and covariance $\sum\limits_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j + k - h)$, for n large.*

[For a related result that does not require the independence of the two series $\{X_{t1}\}$ and $\{X_{t2}\}$ see Theorem 7.3.2 from text [9].]

Remark 7.11. Theorem 7.10 is useful in testing for correlation between two time series. If one of the two processes in the theorem is white noise, then it follows at once from the theorem that $\rho_{12}(h)$ is approximately normally distributed with mean 0 and variance $1/n$, in which case it is straightforward to test the hypothesis that $\rho_{12}(h) = 0$. However, if neither process is white noise, then a value of $\hat{\rho}_{12}(h)$ that is large relative to $n^{-1/2}$ does not necessarily indicate that $\rho_{12}(h)$ is different from zero.

□

7.3.3 Testing for Independence of Two Stationary Time Series

Go back to Table of Contents. Please click TOC

Since by Theorem 7.10 the large-sample distribution of $\hat{\rho}_{12}(h)$ depends on both $\rho_{11}(\cdot)$ and $\rho_{22}(\cdot)$, any test for independence of the two component series cannot be based solely on estimated values of $\rho_{12}(h)$, $h = 0, \pm 1, ...$, without taking into account the nature of the two component series.

This difficulty can be circumvented by "prewhitening" the two series before computing the cross-correlations $\hat{\rho}_{12}(h)$, i.e., by transforming the two series to white noise by application of suitable filters. If $\{X_{t1}\}$ and $\{X_{t2}\}$ are invertible ARMA(p,q) process, this can be achived by the transformations

$$Z_{ti} = \sum_{j=0}^{\infty} \pi_j^{(i)} X_{t-j,i},$$

where $\sum\limits_{j=0}^{\infty} \pi_j^{(i)} z_j = \phi^{(i)}(z)/\theta^{(i)}(z)$ and $\phi^{(i)}$, $\theta^{(i)}$ are the autoregressive and moving-average polynomials of the ith series, $i = 1, 2$.

To test the hypothesis H_0 that $\{X_{t1}\}$ and $\{X_{t2}\}$ are independent series, we observe that under H_0, the corresponding two prewhitened series $\{Z_{t1}\}$ and $\{X_{t2}\}$ are independent series, we observe that under H_0, the corresponding two prewhitened series $\{Z_{t1}\}$ and $\{Z_{t2}\}$ are also independent. Theorem 7.10 then implies that the sample cross-correlations $\hat{\rho}_{12}(h)$, $\hat{\rho}_{12}(k)$, $h \neq k$, of $\{Z_{t1}\}$ and $\{Z_{t2}\}$ are for large n approximately independent and normally distributed with means 0 and variances n^{-1}. An approximate test for independence can therefore be obtained by comparing the values of $|\hat{\rho}_{12}(h)|$ with $1.96n^{-1/2}$, exactly as in Section 5.3.2. If we prewhiten only one of the two original series, say $\{X_{t1}\}$, then under H_0 Theorem 7.10 implies that the sample cross-correlations $\tilde{\rho}_{12}(h)$, $\tilde{\rho}_{12}(k)$, $h \neq k$, of $\{Z_{t1}\}$ and $\{X_{t2}\}$ are for large n approximately normal with means 0, variances n^{-1} and covariance $n^{-1}\rho_{22}(k-h)$, where $\rho_{22}(\cdot)$ is the autocorrelation function of $\{X_{t2}\}$. Hence, for any fixed h, $\tilde{\rho}_{12}(h)$ also falls (under H_0) between the bounds $\pm 1.96n^{-1/2}$ with a probability of approximately 0.95.

7.3.4 Barlett's Formula

Go back to Table of Contents. Please click TOC

The following theorem gives a large-sample approximation to the covariances of the sample cross-correlations $\hat{\rho}_{12}(h)$ and $\hat{\rho}_{12}(k)$ of the bivariate time series $\{\mathbf{X}_t\}$ under the assumption that $\{\mathbf{X}_t\}$ is Gaussian. However, it is not assumed (as in Theorem 7.10) that $\{X_{t1}\}$ is independent of $\{X_{t2}\}$.

Theorem 7.12. ***Bartlett's Formula:***

If $\{\boldsymbol{X}_t\}$ is a bivariate Gaussian time series with covariances satisfying $\sum\limits_{h=-\infty}^{\infty} |\gamma_{ij}(h)| < \infty$, $i, j = 1, 2$, then

$$\begin{aligned}\lim_{n\to\infty} nCov(\hat{\rho}_{12}(h), \hat{\rho}_{12}(k)) = \sum_{j=-\infty}^{\infty} \Big[& \rho_{11}(j)\rho_{22}(j+k-h) + \rho_{12}(j+k)\rho_{21}(j-h) \\ & -\rho_{12}(h)\{\rho_{11}(j)\rho_{12}(j+k) + \rho_{22}(j)\rho_{21}(j-k)\} \\ & -\rho_{12}(j)\{\rho_{11}(j)\rho_{12}(j+h) + \rho_{22}(j)\rho_{21}(j-h)\} \\ & +\rho_{12}(h)\rho_{12}(k)\Big\{\tfrac{1}{2}\rho_{11}^2(j) + \rho_{12}^2(j) + \tfrac{1}{2}\rho_{22}^2(j)\Big\}\Big]\end{aligned}$$

Corollary 7.13. *If* $\{\mathbf{X}_t\}$ *satisfies the conditions for Bartlett's formula, if either* $X_{t1}\}$ *or* $\{X_{t2}\}$ *is white noise, and if*

$$\rho_{12}(h) = 0,\ \ h \notin [a,b],$$

then

$$\lim_{n\to\infty} nVar(\hat{\rho}_{12}(h)) = 1,\ \ h \notin [a,b].$$

7.4 Multivariate ARMA Processes

Go back to Table of Contents. Please click TOC

Definition 7.14. $\{\mathbf{X}_t\}$ is an **ARMA(p,q) process** if $\{\mathbf{X}_t\}$ is stationary and if for every t,

$$\mathbf{X}_t - \Phi_1\mathbf{X}_{t-1} - \cdots - \Phi_p\mathbf{X}_{t=p} = \mathbf{Z}_t + \Theta_1\mathbf{Z}_{t-1} + \cdots + \Theta_q\mathbf{Z}_{t-q},$$

where $\{\mathbf{Z}_t\} \sim WN(0,\bar{\Sigma})$. ($\{\mathbf{X}_t\}$ is an **ARMA(p,q) process with mean** μ if $\{\mathbf{X}_t - \mu\}$ is an ARMA(p,q) process.)

The definition can be written in the more compact form

$$\Phi(B)\mathbf{X}_t = \Theta(B)\mathbf{Z}_t,\ \ \{\mathbf{Z}_t\} \sim WN(\mathbf{0},\bar{\Sigma}),$$

where $\Phi(Z) := I - \Phi_1 z - \cdots - \Phi_p z^p$ and $\Theta(z) := I + \Theta_1 z + \cdots + \Theta_q z^q$ are matrix-valued polynomials, I is the $m \times m$ identify matrix, and B as usual denotes the backward shift operator. (Each component of the matrices $\Phi(z)$, $\Theta(z)$ is a polynomial with real coefficients and degree less than or equal to p, q, respectively.)

Example 7.15. *The multivariate AR(1) process.*

Setting $p = 1$ and $q = 0$ in definition gives the defining equations

$$\mathbf{X}_t = \Phi\mathbf{X}_{t-1} + \mathbf{Z}_t,\ \ \{\mathbf{Z}_t\} \sim WN(0,\bar{\Sigma}),$$

for the multivariate AR(1) series $\{\mathbf{X}_t\}$. We can express $\mathbf{X}_t$ as

$$\mathbf{X}_t = \sum_{j=0}^{\infty} \Phi^j \mathbf{Z}_{t-j},$$

provided that all the eigenvalues of Φ are less than 1 in absolute value, i.e., provided that

$$\det(I - z\Phi) \neq 0\ \forall z \in \mathbb{C}\ \textit{such that}\ |z| \leq 1.$$

□

Definition 7.16. Causality:

An ARMA(p,q) process $\{\mathbf{X}_t\}$ is **causal**, or a **causal function of** $\{Z_t\}$, if there exist matrices $\{\Psi_j\}$ with absolutely summable components such that

$$\mathbf{X}_t = \sum_{j=0}^{\infty} \Psi_j \mathbf{Z}_{t-j},\ \ \forall t.$$

Causality is equivalent to the condition

$$\det \Phi(z) \neq 0 \; \forall z \in \mathbb{C} \; \textit{such that} \; |z| \leq 1.$$

The matrices Ψ_j are found recursively from the equations

$$\Psi_j = \Theta_j + \sum_{k=1}^{\infty} \Phi_k \Psi_{j-k}, \; j = 0, 1, ...,$$

where we define $\Theta_0 = I$, $\Theta_j = 0$ for $j > q$, $\Phi_j = 0$ for $j > p$, and $\Psi_j = 0$ for $j < 0$.

Definition 7.17. Invertibility:

An ARMA(p, q) process $\{\mathbf{X}_t\}$ is **invertible** if there exist matrices $\{\Pi_j\}$ with absolutely summable components such that

$$\mathbf{Z}_t = \sum_{j=0}^{\infty} \Pi_j \mathbf{X}_{t-j} \; \forall \; t.$$

Invertibility is equivalent to the condition

$$\det \Theta(z) \neq 0 \; \forall z \in \mathbb{C} \; \textit{such that} \; |z| \leq 1.$$

The matrices Π_j are found recursively from the equations

$$\Pi_j = -\Phi_j - \sum_{k=1}^{\infty} \Theta_k \Pi_{j-k}, \; j = 0, 1, ...,$$

where we defined $\Phi_0 = -I$, $\Phi_j = 0$ for $j > p$, $\Theta_j = 0$ for $j > q$, and $\Pi_j = 0$ for $j < 0$.

7.4.1 The Covariance Matrix Function of a Causal ARMA Process

Go back to Table of Contents. Please click TOC

From $\Gamma(h) = \sum\limits_{j=-\infty}^{\infty} C_{j+h} \bar{\Sigma} C_j'$, we can express the covariance matrix $\Gamma(h) = \mathbb{E}(\mathbf{X}_{t+h} \mathbf{X}_t')$ of the causal process $\mathbf{X}_t = \sum\limits_{j=0}^{\infty} \Psi_j \mathbf{Z}_{t=j}$ as

$$\Gamma(h) = \sum_{j=0}^{\infty} \Psi_{h+j} \bar{\Sigma} \Psi_j', \; h = 0, \pm 1, ...,$$

where the matrices Ψ_j are found from $\Psi_j = \Theta_j + \sum\limits_{k=1}^{\infty} \sum\limits_{k=1}^{\infty} \Phi_k \Psi_{j-k}$ and $\Psi_j := 0$ for $j < 0$.

The covariance matrices $\Gamma(h)$, $h = 0, \pm 1, ...,$ can also be found by solving the Yule-Walker equations

$$\Gamma(j) - \sum_{r=1}^{p} \Phi_r \Gamma(j-r) = \sum_{j \leq r \leq q} \Theta_r \bar{\Sigma} \Psi_{r-j}, \; j = 0, 1, 2, ...,$$

obtained by postmultiplying $\mathbf{X}_t - \Phi\mathbf{X}_{t-1} - \cdots - \Phi_p\mathbf{X}_{t-p} = \mathbf{Z}_t + \Theta_1\mathbf{Z}_{t-1} + \cdots + \Theta_q\mathbf{Z}_{t-q}$ by $\mathbf{X}'_{t-j}$ and taking expectations. The first $p+1$ of the equations $\Gamma(j) - \sum_{r=1}^{p}\Phi_r\Gamma(j-r) = \sum_{j\leq r\leq q}\Theta_r\bar{\Sigma}\Psi_{r-j}$ can be solved for the components of $\Gamma(0), ..., \Gamma(p)$ using the fact that $\Gamma(-h) = \Gamma'(h)$. The remaining equations then give $\Gamma(p+1)$, $\Gamma(p+2)$, ... recursively. An explicit form of the solution of these can be written down by making use of Kronecker products and the vec operator (see e.g., Lütkepohl, 1993) [23].

Remark 7.18. If z_0 is the root of $\det\Phi(z) = 0$ with smallest absolute value, then it can be shown from the recursions $\Psi_j = \Theta + \sum_{k=1}^{\infty}\Phi_k\Psi_{j-k}$ that $\Psi_j/r^j \to 0$ as $j\to\infty$ for all r such that $|z_0|^{-1} < r < 1$. Hence, there is a constant C such that each component of Ψ_j is smaller in absolute value by KR^{2j}. Provided that $|z_0|$ is not very close to 1, this means that the series $\Gamma(h) = \sum_{j=0}^{\infty}\Psi_{h+j}\bar{\Sigma}\Psi'_j$ converges rapidly, and the error incurred in each component by truncating the series after the term with $j = k-1$ is smaller in absolute value than $\sum_{j=k}^{\infty}Kr^{2k}/(1-r^2)$.

□

7.5 Best Linear Predictors of Second-Order Random Vectors

Go back to Table of Contents. Please click TOC Let $\{\mathbf{X}_t = (X_{t1}, ..., X_{tm})'\}$ be an m-variate time series with means $\mathbb{E}(\mathbf{X}_t) = \mu_t$ and covariance function given by the $m\times$m matrices

$$K(i,j) = \mathbb{E}(\mathbf{X}_j\mathbf{X}'_j) - \mu_i\mu'_j.$$

If $\mathbf{Y} = (Y_1, ..., Y_m)'$ is a random vector with finite second moments and $\mathbb{E}(\mathbf{Y}) = \mu$, we define

$$P_n(\mathbf{Y}) = (P_nY_1, ..., P_nY_m)',$$

where P_nY_j is the best linear predictor of the component Y_j of $\mathbf{Y}$ in terms of all of the components of the vectors $\mathbf{X}_t$, $t = 1, ..., n$, and the constant 1. It follows immediately from the properties of the prediction operator (Section 2.5) that

$$P_n(\mathbf{Y}) = \mu + A_1(\mathbf{X}_n - \mu_n) + \cdots + A_n(\mathbf{X}_1 - \mu_1)$$

for some matrices $A_1, ..., A_n$, and that

$$\mathbf{Y} - P_n(\mathbf{Y}) \perp \mathbf{X}_{n+1-i},\ i = 1, ..., n,$$

where we say that two m-dimensional random vectors $\mathbf{X}$ and $\mathbf{Y}$ are orthogonal (written $\mathbf{X}\perp\mathbf{Y}$) if $\mathbb{E}(\mathbf{XY}')$ is a matrix of zeros. The vector of best predictors $P_n() = (P_nY_1, ..., P_nY_m)'$ is uniquely determined by

$P_n(\mathbf{Y})$ and $Y - P_n(\mathbf{Y}) \perp \mathbf{X}_{n+1-i}$, although it is possible that there may be more than one possible choice for $A_1, ..., A_n$.

As a special case of the above, if $\{\mathbf{X}_t\}$ is a zero-mean time series, the best linear predictor $\hat{\mathbf{X}}_{n+1}$ of $\mathbf{X}_{n+1}$ in terms of $\mathbf{X}_1, ..., \mathbf{X}_n$ is obtained an replacing $\mathbf{Y}$ by $\mathbf{X}_{n+1}$ in $P_n(\mathbf{Y})$. Thus

$$\hat{[}\mathbf{X}_{n+1} = \begin{cases} \mathbf{0}, & if\ n = 0, \\ P_n(\mathbf{X}_{n+1}), & if\ n \geq 1. \end{cases}$$

Hence, we can write

$$\hat{\mathbf{X}}_{n+1} = \Phi_{n1}\mathbf{X}_n + \cdots + \Phi_{nn}\mathbf{X}_1,\ n = 1, 2, ...,$$

where, from $\mathbf{Y} - P_n(\mathbf{Y}) \perp \mathbf{X}_{n+1-i}$, the coefficients Φ_{nj}, $j = 1, ..., n$, are such that

$$\mathbb{E}(\hat{\mathbf{X}}_{n+1}\mathbf{X}'_{n+1-i}) = \mathbb{E}(\mathbf{X}_{n+1}\mathbf{X}'_{n+1-i}),\ i = 1, ..., n,$$

i.e.,

$$\sum_{j=1}^{n} \Phi_{nj} K(n+1-j, n+1-i) = K(n+1, n+1-i),\ i = 1, ..., n.$$

In the case where $\{\mathbf{X}_t\}$ is stationary with $K(i,j) = \Gamma(i-j)$, the prediction equations simplify to the m-dimensional analogues of $\Gamma_n \mathbf{a}_n = \gamma_n(h)$ (see Section 2.5), i.e.,

$$\sum_{j=1}^{n} \Phi_{nj}\Gamma(i-j) = \Gamma(i),\ i = 1, ...n.$$

Provided that the covariance matrix of the nm components of $\mathbf{X}_1, ..., \mathbf{X}_n$ is nonsingular for every $n \geq 1$, the coefficients $\{\Phi_{nj}\}$ can be determined recursively using a multivariate version of the Durbin-Levinson algorithm given by Whittle (1963) [36]. Whittle's recursions also dtermine the covariance matrices of the one-step prediction errors, namely, $V_0 = \Gamma(0)$ and, for $n \geq 1$,

$$\begin{aligned} V_n &= \mathbb{E}(\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1})(\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1})' \\ &= \Gamma(0) - \Phi_{n1}\Gamma(-1) - \cdots - \Phi_{nn}\Gamma(-n). \end{aligned}$$

7.6 Modeling and Forecasting with Multivariate AR Processes

Go back to Table of Contents. Please click TOC

If $\{\mathbf{X}_t\}$ is any zero-mean second-order multivariate time series, it is easy to show from the results of Section 7.5 (Problem 7.4 of text [9]) that the one-step prediction errors $\mathbf{X}_j - \hat{\mathbf{X}}_j$, $j = 1, ..., n$, have the property

$$\mathbb{E}(\mathbf{X}_j - \hat{\mathbf{X}}_j)(\mathbf{X}_k - \hat{\mathbf{X}}_k)' = 0\ for\ j \neq k.$$

Moreover, the matrix M such that

$$\begin{bmatrix} \mathbf{X}_1 - \hat{\mathbf{X}}_1 \\ \mathbf{X}_2 - \hat{\mathbf{X}}_2 \\ \mathbf{X}_3 - \hat{\mathbf{X}}_3 \\ \dots \\ \mathbf{X}_n - \hat{\mathbf{X}}_n \end{bmatrix} = M \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \\ \dots \\ \mathbf{X}_n \end{bmatrix}$$

is lower triangular with ones on the diagonal and therefore has determinant equal to 1.

If the series $\{\mathbf{X}_t\}$ is also Gaussian, then $\mathbb{E}(\mathbf{X}_j - \hat{\mathbf{X}}_j)(\mathbf{X}_k - \hat{\mathbf{X}}_k)' = 0$ implies that the prediction errors $\mathbf{I}_j = \mathbf{X}_j - \hat{\mathbf{X}}_j$, $j = 1, ..., n$, are independent with covariance matrices $V_0, ..., V_{n-1}$, respectively (as in $V_n = \mathbb{E}(\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1})(\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1})' = \Gamma(0) - \Phi_{n1}\Gamma(-1) - \cdots - \Phi_{nn}\Gamma(-n)$). Consequently, the joint density of the prediction errors is the product

$$f(\mathbf{u}_1, ..., \mathbf{u}_n) = (2\pi)^{-nm/2}\left(\prod_{j=1}^{n} \det V_{j-1}\right)^{-1/2} \exp\left[-\frac{1}{2}\sum_{j=1}^{n} \mathbf{u}_j' V_{j-1}^{-1} \mathbf{u}_j\right].$$

Since the determinant of the matrix M is equal to 1, the joint density of the observations $\mathbf{X}_1, ..., \mathbf{X}_n$ at $\mathbf{x}_1, ..., \mathbf{x}_n$ is obtained on replacing $\mathbf{u}_1, ..., \mathbf{u}_n$ in the last expression by the values of $\mathbf{X}_j - \hat{\mathbf{X}}_j$ corresponding to the observations $\mathbf{x}_1, ..., \mathbf{x}_n$.

If we suppose that $\{\mathbf{X}_t\}$ is a zero-mean m-variate AR(p) process with coefficient matrices $\Phi = \{\Phi_1, ..., \Phi_p\}$ and white noise covariance matrix $\bar{\Sigma}$, we can therefore express the likelihood of the observations $\mathbf{X}_1, ..., \mathbf{X}_n$ as

$$L(\Phi, \bar{\Sigma}) = (2\pi)^{-nm/2}\left(\prod_{j=1}^{n} \det V_{j-1}\right)^{-1/2} \exp\left[-\frac{1}{2}\sum_{j=1}^{n} \mathbf{U}_j' V_{j-1}^{-1} \mathbf{U}_j\right],$$

where $\mathbf{U}_j = \mathbf{X}_j - \hat{\mathbf{X}}_j$, $j = 1, ..., n$, and $\hat{\mathbf{X}}_j$ and V_j are found from $\hat{\mathbf{X}}_{n+1}$, $\sum_{j=1}^{n} \Phi_{nj}\Gamma(i-j) = \Gamma(i)$, and V_n.

7.6.1 Estimation for Autoregressive Processes Using Whittle's Algorithm

Go back to Table of Contents. Please click TOC

If $\{\mathbf{X}_t\}$ is the (causal) multivariate AR(p) process defined by the difference equations

$$\mathbf{X}_t = \Phi_1\mathbf{X}_{t-1} + \cdots + \Phi_p\mathbf{X}_{t-p} + \mathbf{Z}_t, \ \{\mathbf{Z}_t\} \sim WN(0, \bar{\Sigma}),$$

then multiplying by $\mathbf{X}_{t=j}'$, $j = 0, ..., p$, and taking expectations gives the equations

$$\bar{\Sigma} = \Gamma(0) - \sum_{j=1}^{p} \Phi_j\Gamma(-j)$$

and

$$\Gamma(i) = \sum_{j=1}^{n} \Phi_j\Gamma(i-j), \ \imath = 1, ..., p.$$

Given the matrices $\Gamma(0), ..., \Gamma(p)$, equations $\Gamma(i)$ can be used to determine the coefficient matrics $\Phi_1, ..., \Phi_p$. The white noise covariance matrix $\bar{\Sigma}$ can then be found from $\bar{\Sigma} = \Gamma(0) - \sum\limits_{j=1}^{p} \Phi_j \Gamma(-j)$. The solution of these equations $\Phi_1, ..., \Phi_p$, and $\bar{\Sigma}$ is identical to the solution $\sum\limits_{j=1}^{n} \Phi_{nj}\Gamma(i-j) = \Gamma(i)$, and V_n for the prediction coefficient matrics $\Phi_{p1}, ..., \Phi_{pp}$ and the corresponding prediction error covariance matrix V_p. Consequently, Whittle's algorithm can be used to carry out the algebra.

The Yule-Walker estimators $\hat{\Phi}_1, ..., \hat{\Phi}_p$, and $\hat{\bar{\Sigma}}$ for the model $\mathbf{X}_t$ fitted to the data $\mathbf{X}_1, ..., \mathbf{X}_n$ are obtained by replacing $\Gamma(j)$ by $\hat{\Gamma}(j)$, $j = 0, ..., p$, and solving the resulting equations for $\Phi_1, ..., \Phi_p$, and $\bar{\Sigma}$. The solution can be found from programs.

7.6.2 Forecasting Multivariate Autoregressive Processes

Go back to Table of Contents. Please click TOC

The technique developed in Section 7.5 allows us to compute the minimum mean squared error one-step linear predictors $\hat{\mathbf{X}}_{n+1}$ for any multivariate stationary time series from the mean μ and autocovariance matrices $\Gamma(h)$ by recursively determining the coefficients Φ_{ni}, $i = 1, ..., n$, and evaluating

$$\hat{\mathbf{X}}_{n+1} = \mu + \Phi_{n1}(\mathbf{X}_n - \mu) + \cdots + \Phi_{nn}(\mathbf{X}_1 - \mu_{)}.$$

The situation is simplified when $\{\mathbf{X}\}$ is the causal AR(p) process defined by $\mathbf{X}_t = \Phi_1\mathbf{X}_{t-1} + \cdots + \Phi_p\mathbf{X}_{t-p} + \mathbf{Z}_t$, since for $n \geq p$

$$\hat{\mathbf{X}}_{n+1} = \Phi_1\mathbf{X}_n + \cdots + \Phi_p\mathbf{X}_{n+1-p}.$$

To verify this, it suffices to observe that the right-hand side has the required form $P_n(\mathbf{Y}) = \mu + A_1(\mathbf{X}_n - \mu_n) + \cdots + A_n(\mathbf{X}_1 - \mu_1)$ and that the prediction error

$$\mathbf{X}_{n+1} - \Phi_1\mathbf{X}_n - \cdots - \Phi_p\mathbf{X}_{n+1-p} = \mathbf{Z}_{n+1}$$

is orthogonal to $\mathbf{X}_1, ..., \mathbf{X}_n$ in the sense of $\mathbf{Y} - P_n(\mathbf{Y} \perp \mathbf{X}_{n+1-i}$. (In fact, the prediction error is orthogonal to *all* $\mathbf{X}_j$, $-\infty < j \leq n$, showing that if $n \geq p$, then $\hat{\mathbf{X}}_{n+1}$ is also the best linear predictor of $\mathbf{X}_{n+1}$ in terms of all components of $\mathbf{X}_j$, $-\infty < j \leq n$.) The covariance matrix of the one-step prediction error is clearly $\mathbb{E}(\mathbf{Z}_{n+1}\mathbf{Z}'_{n+1}) = \bar{\Sigma}$.

To compute the best h-step linear predictor $P_n\mathbf{X}_{n+h}$ based on all the components of $\mathbf{X}_1, ..., \mathbf{X}_n$ we apply the linear operator P_n to $\mathbf{X}_t = \Phi_1\mathbf{X}_{t-1} + \cdots + \Phi_p\mathbf{X}_{t-p} + \mathbf{Z}_t$ to obtain the recursions

$$P_n\mathbf{X}_{n+h} = \Phi_1 P_n\mathbf{X}_{n+h-1} + \cdots + \Phi_p P_n\mathbf{X}_{n+h-p}.$$

These equations are easily solved resurcively, first for $P_n\mathbf{X}_{n+1}$, then for $P_n\mathbf{X}_{n+2}$, $P_n\mathbf{X}_{n+3}$, ..., etc. also satisfy $P_n\mathbf{X}_{n+h}$ and are therefore the same as the h-step predictors based on $\mathbf{X}_1, ..., \mathbf{X}_n$.

To compute the h-step error covariance matrices, recall from $\mathbf{X}_t = \sum\limits_{j=0}^{\infty}\Psi_j\mathbf{Z}_{t-j}$ that

$$\mathbf{X}_{n+h} = \sum_{j=0}^{\infty}\Psi_j\mathbf{Z}_{n+h-j},$$

where the coefficients matrics Ψ_j are found from the recursions from $\Psi_j = \Theta_j + \sum\limits_{k=1}^{\infty} \Phi_k \Psi_{j-k}$ with $q = 0$. From $\mathbf{X}_{n+h}$ we find that for $n \geq p$,

$$P_n \mathbf{X}_{n+h} = \sum_{j=h}^{\infty} \Psi_j \mathbf{Z}_{n+h-j}.$$

Subtracting $P_n \mathbf{X}_{n+h}$ from $\mathbf{X}_{n+h}$ gives the h-step prediction error

$$\mathbf{X}_{n+h} - P_n \mathbf{X}_{n+h} = \sum_{j=0}^{h-1} \Psi_j \mathbf{Z}_{n+h-j},$$

with covariance matrix

$$\mathbb{E}[(\mathbf{X}_{n+h} - P +_n \mathbf{X}_{n+h})(X_{n+h} - P_n \mathbf{X}_{n+h})'] = \sum_{j=0}^{h-1} \Psi_j \bar{\Sigma} \Psi_j', \ n \geq p.$$

For the (not necessarily zero-mean) causal AR(p) process defined by

$$\mathbf{X}_t = \Phi_0 + \Phi_1 \mathbf{X}_{t-1} + \cdots + \Phi_p \mathbf{X}_{t-p} + \mathbf{Z}_t, \ \{\mathbf{Z}_t\} \sim WN(\mathbf{0}\bar{\Sigma}),$$

equations $\hat{\mathbf{X}}_{n+1} = \Phi_1 \mathbf{X}_n + \cdots + \Phi_p \mathbf{X}_{n+1-p}$ and $P_n \mathbf{X}_{n+h} = \sum\limits_{j=0}^{\infty} \Psi_j \mathbf{Z}_{n+h-j}$ remain valid, provided that μ_0 is added to each of their right-hand sides. The error covariance matrices are the same as in the case $\phi_0 = \mathbf{0}$.

The above calculations are all based on the assumption that the AR(p) model for the series is known. However, in practice, the parameters of the model are usually estimated from the data, and the uncertainty in the predicted values of the series will be larger than indicated by $\mathbb{E}[(\mathbf{X}_{n+h} - P_n \mathbf{X}_{n+h})(\mathbf{X}_{n+h} - P_n \mathbf{X}_{n+h})'] = \sum\limits_{j=0}^{h-1} \Psi_j \bar{\Sigma} \Psi_j'$ because of parameter estimation errors (see Lütkepohl, 1993) [23].

7.7 Cointegration

Go back to Table of Contents. Please click TOC

We have seen that nonstationary univariate time series can frequently be made stationary by applying the differencing operator $\nabla = 1 - B$ respeatedly. If $\{\nabla^d X_t\}$ is stationary for some positive integer d but $\{\nabla^{d-1} X_t\}$ is nonstationary, we say that $\{X_t\}$ is **integrated of order** d, or some concisely, $\{X_t\} \sim I(d)$. many macroeconomic time series are found to be integrated of order 1.

If $\{\mathbf{X}_t\}$ is a k-variate time series, we define $\{\nabla^d \mathbf{X}_t\}$ to be the series whose jth component is obtained by applying the operator $(1_B)^d$ to the jth component of $\{\mathbf{X}_t\}$, $j = 1, ..., k$. The idea of a cointegrated multivariate time serries was introduced by Granger (1981) [15] and developed by Engle and Granger (1987) [14]. Here we use the slightly different definition of Lütkepohl (1993) [23]. We say that the k-dimensional time series $\{\mathbf{X}_t\}$ is integrated or order d (or $\{\mathbf{X}_t\} \sim I(d)$) if d is a positive integer, $\{\nabla^d \mathbf{X}_t\}$ is stationary, and $\{\nabla^{d-1} \mathbf{X}_t\}$ is nonstationary. The $I(d)$ process $\{\mathbf{X}_t\}$ is said to be **cointegrated with cointegration vector** α if α is a $k \times 1$ vector such that $\{\alpha' \mathbf{X}_t\}$ is of order less than d.

8 Forecasting Techniques

Go back to Table of Contents. Please click TOC

In this chapter we discuss three forecasting techniques that have less emphasis on the explicit construction of a model for the data. Each of the three selects, from a limited class of algorithms, the one that is optimal according to specified criteria.

The three techniques have been found in practice to be effective on wide ranges of real data sets.

The ARAR algorithm described in Section 9.1 is an adaptation of the ARARMA algorithm (Newton and Parzen, 1984; Parzen, 1982) [26] [27] in which the idea is to apply automatically selected "memory-shortening" transformations (if necessary) to the data and then to fit an ARMA model to the transformed series. The ARAR algorithm we describe is a version of this in which the ARMA fitting step is replaced by the fitting of a subset AR model to the transformed data.

The Holt-Winters (HW) algorithm described in Section 9.2 uses a set of simple recursions that generalizes the exponential smoothing recursions of Section 1.5.1 to generate forecasts of series containing a locally linear trend.

The Holt-Winters seasonal (HWS) algorithm extends the HW algorithm to handle data in which there are both trend and seasonal variation of known period. It is described in section 9.3.

8.1 The ARAR Algorithm

Go back to Table of Contents. Please click TOC

8.1.1 Memory Shortening

Go back to Table of Contents. Please click TOC

Given a data set $\{Y_t,\ t = 1, 2, ..., n\}$, the first step is to decide whether the underlying process is "long-memory," and if so to apply a memory-shortening transformation before attempting to fit an autoregressive model. There are two types allowed:

$$\tilde{Y}_t = Y_t - \hat{\Phi}(\hat{\tau})Y_{t-\hat{\tau}}$$

and

$$\tilde{Y}_t = Y_t - \hat{\phi}_1 Y_{t-1} - \hat{\phi}_2 Y_{t-2}.$$

With the aid of the five-step algorithm described below, we classify $\{Y_t\}$ and take one of the following three courses of action:

- **L**. Declare $\{Y_t\}$ to be long-memory and form $\{\tilde{Y}_t\}$ using $\tilde{Y}_t = Y_t - \hat{\phi}(\hat{\tau})Y_{t-\hat{\tau}})$.
- **M**. Declare $\{Y_t\}$ to be moderately long-memory and form $\{\tilde{Y}_t\}$ using $\tilde{Y}_t = Y - t - \hat{\phi}_1 Y_{t-1} - \hat{\phi}_2 Y_{t-2}$.
- **S**. $\{Y_t\}$ to be short-memory.

If the alternative L or M is chosen, then the transformed series $\{\tilde{Y}_t\}$ is again checked. If it is found to be long-memory or moderately long-memory, then a further transformation is performed. The process continues until the transformed series is classified as short-memoery. At most three memory-hsortening transformations are performed, but it is very rare to require more than two. The algorithm for deciding among L, M, and S can be as follows:

(1) For each $\tau = 1, 2, ..., 15$, we find the value $\hat{\phi}(\tau)$ of ϕ that minimizes

$$ERR(\phi, \tau) = \frac{\sum\limits_{t=\tau+1}^{n} [Y_t - \phi Y_{t-\tau}]^2}{\sum\limits_{t=\tau+1}^{n} Y_t^2}.$$

We then define

$$Err(\tau) = ERR(\hat{\phi}(\tau), \tau)$$

and choose the lag $\hat{\tau}$ to be the value of τ that minimizes $Err(\tau)$.

(2) If $Err(\hat{\tau}) \leq 8/n$, go to L.

(3) If $\hat{\phi}(\hat{\tau}) \geq .93$ and $\hat{\tau} > 2$, go to L.

(4) If $\hat{\phi}(\hat{\tau}) \geq .93$ and $\hat{\tau} = 1$ or 2, determine the values $\hat{\phi}_1$ and $\hat{\phi}_2$ of ϕ_1 and ϕ_2 that minimize $\sum\limits_{t=3}^{n} [Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}]^2$; then go to M.

(5) If $\hat{\phi}(\hat{\tau}) < .93$, go to S.

8.1.2 The Holt-Winters Algorithm

Go back to Table of Contents. Please click TOC

Let $\{S_t, t = k+1, ..., n\}$ denote the memory-shortened series derived from $\{Y_t\}$ by the algorithm of the previous section and let $\bar{S}$ denote the sample mean of $S_{k+1}, ..., S_n$.

The next step in the modeling procedure is to fit an autoregressive process to the mean-corrected series

$$X_t = S_t - \bar{S}, \ t = k+1, ..., n.$$

The fitted model has the form

$$X - t = \phi X_{t-1} + \phi_{l_1} X_{t-l_1} + \phi_{l_2} X_{t-l_2} + \phi_{l_3} X_{t-l_3} + Z_t,$$

where $\{Z_t\} \sim WN(0, \sigma^2)$, and for given lags l_1, l_2, and l_3, the coefficients ϕ_j and the white noise variance σ^2 are found from the Yule-Walker Equations

$$\begin{bmatrix} 1 & \hat{\rho}(l_1 - 1) & \hat{\rho}(l_2 - 1) & \hat{\rho}(l_3 - 1) \\ \hat{\rho}(l_1 - 1) & 1 & \hat{\rho}(l_2 - l_1) & \hat{\rho}(l_3 - l_1) \\ \hat{\rho}(l_2 - 1) & \hat{\rho}(l_2 - l_1) & 1 & \hat{\rho}(l_3 - l_2) \\ \hat{\rho}(l_3 - 1) & \hat{\rho}(l_3 - l_1) & \hat{\rho}(l_3 - l_2) & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_{l_1} \\ \phi_{l_2} \\ \phi_{l_3} \end{bmatrix} = \begin{bmatrix} \phi_\rho(1) \\ \phi_\rho(l_1) \\ \phi_\rho(l_2) \\ \phi_\rho(l_3) \end{bmatrix}$$

and

$$\sigma^2 = \hat{\gamma}(0)[1 - \phi_1 \hat{\rho}(1) - \phi_{l_1} \hat{\rho}(l_1) - \phi_{l_2} \hat{\rho}(l_2) - \phi_{l_3} \hat{\rho}(l_3)],$$

where $\hat{\gamma}(j)$ and $\rho\rho(j)$, $j = 0, 1, 2, ...,$ are the sample autocovariances and autocorrelations of the series $\{X_t\}$.

The program computes the coefficients ϕ_j for each set of lags such that

$$1 < l_1 < l_2 < l_3 \leq m,$$

where m can be chosen to be either 13 or 26. If then selects the model for which the Yule-Walker estimate σ^2 is minimal and prints out the lags, coefficients, and white noise variance for the fitted model.

A slower procedure chooses the lags and coefficients (computed from the Yule-Walker equations as above) that maximize the Gaussian likelihood of the observations. For this option the maximum lag m is 13.

8.1.3 Forecasting

Go back to Table of Contents. Please click TOC

If the memory-shortening filter found in the first step has coefficients $\psi_0(= 1)$, $\psi_1, ..., \psi_k$ $(k \geq 0)$, then the memory-shortened series can be expressed as

$$S_t = \psi(B)Y_t = Y_t + \psi_1 Y_{t-1} + \cdots + \psi_k Y_{t-k},$$

where $\psi(B)$ is the polynomial in the backward shift operator,

$$\psi(B) = 1 + \psi_1 B + \cdots + \psi_k B^k.$$

Similarly, if the coefficients of the subset autoregression found in the second step are ϕ_1, ϕ_{l_1}, ϕ_{l_2}, and ϕ_{l_3}, then the subset AR model for the mean-corrected series $\{X_t = S_t - \bar{S}\}$ is

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

where $\{Z_t\} \sim WN(0, \sigma^2)$ and

$$\phi(B) = 1 - \phi_1 B - \phi_{l_1} B^{l_1} - \phi_{l_2} B^{l_2} - \phi_{l_3} B^{l_3}.$$

From S_t and $\phi(B)X_t = Z_t$ we obtain the equations

$$\xi(B)Y_t = \phi(1)\bar{S} + Z_t,$$

where

$$\xi(B) = \psi(B)\phi(B) = 1 + \xi_1 B + \cdots + \xi_{k+l_3} B^{k+l_3}.$$

Assuming that the fitted model $\xi(B)Y_t = \phi(1)\bar{S} + Z_t$ is appropriate and that the white noise term Z_t is uncorrelated with $\{Y_j, j < t\}$ for each t, we can determine the minimum mean squared error linear predictors $P_n Y_{n+h}$ of Y_{n+h} in terms of $\{1, Y_1, ..., Y_n\}$, for $n > k + l_3$, from the recursions

$$P_n Y_{n+h} = -\sum_{j=1}^{k+l_3} \xi_j P_n Y_{n+h-j} + \phi(1)\bar{S}, \;\; h \geq 1,$$

with the initial conditions

$$P_n Y_{n+h} = Y_{n+h}, \;\; for \; h \leq 0.$$

The mean squared error of the predictor P_nY_{n+h} is found to be (see Problem 9.1 in text [9])

$$\mathbb{E}[(Y_{n+h} - P_nY_{n+h})^2] = \sum_{j=1}^{h-1} \tau_j^2 \sigma^2,$$

where $\sum_{j=0}^{\infty} \tau_j z^j$ is the Taylor expansion of $1/\xi(z)$ in a neighborhood of $z = 0$. Equivalently the sequence $\{\tau_j\}$ can be found from the recursion

$$\tau_0 = 1, \ \sum_{j=0}^{n} \tau_j \xi_{n-j} = 0, \ n = 1, 2, ...$$

8.1.4 Application of the ARAR Algorithm

Go back to Table of Contents. Please click TOC

To determine an ARAR model for a given data set $\{Y_t\}$, choose the appropriate options in the resulting dialog box. These include specification of the number of forecasts required, whether or not you wish to include the memory-shortening step, whether you require prediction bounds, and which of the optimality criteria is to be used. Once you have made these selections, the forecasts will be plotted with the original data. Then you want to check the coefficients $1, \psi_1, ..., \psi_k$ of the memory-shortening filter $\psi(B)$, the lags and coefficients of the subset autoregression

$$X - t - \phi_1 X_{t-1} - \phi_{l_1} X_{t-l_1} - \phi_{l_2} X_{t-l_2} - \phi_{l_3} X_{t-l_3} = Z_t,$$

and the coefficients ξ_j of B^j in the overall whitening filter

$$\xi(B) = (1 + \psi_1 B + \cdots + \psi_k B^k)(1 - \phi_1 B - \phi_{l_1} B^{l_1} - \phi_{l_2} B^{l_2} - \phi_{l_3} B^{l_3}).$$

The numerical values of the predictors, their root mean squared errors, and the prediction bounds are also printed.

8.2 The Holt-Winters Algorithm

Go back to Table of Contents. Please click TOC

8.2.1 The Algorithm

Go back to Table of Contents. Please click TOC

Given observations $Y_1, Y_2, ..., Y - n$ from the “trend plus noise” model $X_t = m_t + Y_t$, the exponenail smoothing recursions $\hat{m}_t = \alpha X_t + (1 - \alpha)\hat{m}_{t-1}$ allowed us to compute estimates $\hat{m}_t$ of the trend at times $t = 1, 2, ..., n$.

If the data have a (nonconstant) trend, then a natural generalization of the forecast function $P_nY_{n+h} = \hat{m}_n$ that takes this into account is

$$P_nY_{n+h} = \hat{a}_n + \hat{b}_h, \ h = 1, 2, ...,$$

where $\hat{a}_n$ and $\hat{b}_n$ can be thought of as estimates of the "level" a_n and "slope" b_n of the trend function at time n. Holt (1957) [18] suggested a recursive scheme for computing the quantities $\hat{a}_n$ and $\hat{b}_n$ in $P_nY_{n+h}\hat{a}_n + \hat{b}_nh$. Denoting by $\hat{Y}_{n+1}$ the one-step forecast P_nY_{n+1}, we have from P_nY_{n+1}

$$\hat{Y}_{n+1} = \hat{a}_n + \hat{b}_n.$$

Now, as in exponential smoothing, we suppose that the estimated level at time $n+1$ is a linear combination of the observed value at time $n+1$ and the forecast value at time $n+1$. Thus,

$$\hat{a}_{n+1} = \alpha Y_{n+1} + (1-\alpha)(\hat{a}_n + \hat{b}_n).$$

We can then estimate the slope at time $n+1$ as a linear combination of $\hat{a}_{n+1} - \hat{a}_n$ and the estimated slope $\hat{b}_n$ at time n. Thus,

$$\hat{b}_{n+1} = \beta(\hat{a}_{n+1} - \hat{a}_n) + (1-\beta)\hat{b}_n.$$

In order to solve the recursions $\hat{a}_{n+1}$ and $\hat{b}_{n+1}$ we need initial conditions. A natural choise is to set

$$\hat{a}_2 = Y_2$$

and

$$\hat{b}_2 = Y_2 - Y_1.$$

Then $\hat{a}_{n+1}$ and $\hat{b}_{n+1}$ can be solved successively for $\hat{a}_i$ and $\hat{b}_i$, $i = 3, ..., n$, and the predictors P_nY_{n+h} found from P_nY_{n+h}.

The forecasts depend on the "smoothing parameters" α and β. These can either be prescribed arbitrarily (with values between 0 and 1) or chosen in a more systematic way to minimize the sum of squares of the one-step errors $\sum\limits_{i=3}^{n}(Y_i - P_{i-1}Y_i)^2$, obtained when the algorithm is applied to the already observed data.

Before illustrating the use of the Holt-Winters forecasting procedure, we discuss the connection between the recursions $\hat{a}_{n+1}$ and $\hat{b}_{n+1}$ and the steady-state solution of the Kalman filtering equations for a local linear trend model. Suppose $\{Y_t\}$ follows the local linear structural model with observation equation

$$Y_t = M_t + W_t$$

and the state equation

$$\begin{bmatrix} M_{t+1} \\ B_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} M_t \\ B_t \end{bmatrix} + \begin{bmatrix} V_t \\ U_t \end{bmatrix}.$$

Now define $\hat{a}_n$ and $\hat{b}_n$ to be the filtered estimates of M_n and B_n, respectively, i.e.,

$$\hat{a}_n = M_{n|n} := P_nM_n,$$

$$\hat{b}_n = B_{n|n} := P_nB_n.$$

Using Problem 8.17 from text [9] and Kalman recursion (see remark), we find that

$$\begin{bmatrix} \hat{a}_{n+1} \\ \hat{b}_{n+1} \end{bmatrix} = \begin{bmatrix} \hat{a}_n + \hat{b}_n \\ \hat{b}_n \end{bmatrix} + \triangle_n^{-1}\Omega_nG'(Y_n - \hat{a}_n - \hat{b}_n),$$

where $G = \begin{bmatrix}1 & 0\end{bmatrix}$. Assuming that $\Omega_n = \Omega_1 = [\Omega_{ij}]_{i,j=1}^2$ is the steady-state solution of $\Omega_{t+1} = F_t \Omega F_t' + Q_t - \Theta_t \triangle_t^{-1} \Theta_t'$ for this model, then $\triangle_n = \Omega_{11} + \sigma_\omega^2$ for all n, so that $\begin{bmatrix}\hat{a}_{n+1} \\ \hat{b}_{n+1}\end{bmatrix}$ simplifies to the equations

$$\hat{a}_{n+1} = \hat{a}_n + \hat{b}_n + \frac{\Omega_{11}}{\Omega_{11} + \sigma_\omega^2}(Y_n - \hat{a}_n - \hat{b}_n)$$

and

$$\hat{b}_{n+1} = \hat{b}_n + \frac{\Omega_{12}}{\Omega_{11} + \sigma_\omega^2}(Y_n - \hat{a}_n - \hat{b}_n).$$

Solving $\hat{a}_{n+1}$ for $(Y_n - \hat{a}_n - \hat{b}_n)$ and substituting into $\hat{b}_{n+1}$, we find that

$$\hat{a}_{n+1} = \alpha Y_{n+1} + (1-\alpha)(\hat{a}_n + \hat{b}_n),$$

$$\hat{b}_{n+1} = \beta(\hat{a}_{n+1} - \hat{a}_n) + (1-\beta)\hat{b}_n$$

with $\alpha = \Omega_{11}/(\Omega_{11} + \sigma_\omega^2)$ and $\beta = \Omega_{21}/\Omega_{11}$. These equations coincide with the Holt-Winters recursions. Equations relating to α and β to the variances σ_u^2, σ_v^2, and σ_ω^2 can be found in Harvey (1990) [17].

Remark 8.1. From page 273 of text [9], we have **Kalman Prediction:**

For the state-sapce model, the one-step predictors $\hat{\mathbf{X}}_t := P_{t-1}(\mathbf{X}_t)$ and their error covariance matrices $\Omega_t = \mathbb{E}[(\mathbf{X}_t - \hat{\mathbf{X}}_t)(\mathbf{X}_t - \hat{\mathbf{X}}_t)']$ are uniquely determined by the initial conditions

$$\hat{\mathbf{X}}_1 P(\mathbf{X}_1|\mathbf{Y}_0),\ \Omega_1 = \mathbb{E}[(\mathbf{X}_1 - \hat{\mathbf{X}}_1)(\mathbf{X}_1 - \hat{\mathbf{X}}_1)']$$

and the recursions, for $t = 1, ...,$

$$\hat{\mathbf{X}}_{t+1} = F_t \hat{\mathbf{X}}_t + \Theta_t \nabla_t^{-1}(\mathbf{Y}_t - G_t \hat{\mathbf{X}}_t),$$

$$\Omega_{t+1} = F_t \Omega_t F_t' + Q_t - \Theta_t \nabla_t^{-1} \Theta_t',$$

where

$$\nabla_t = G_t \Omega_t G_t' + R_t,$$

$$\Theta_t = F_t \Omega_t G_t',$$

and $\triangle_t^{-1}$ is any generalized inverse of $\triangle_t$.

Kalman Fitlering:

The filtered estimates $\mathbf{X}_{t|t} = P_t(\mathbf{X}_t)$ and their error covariance matrices $\Omega_{t|t} = \mathbb{E}[(\mathbf{X}_t - \mathbf{X}_{t|t}(\mathbf{X}_t - \mathbf{X}_{t|t})']$ are determined by the relations

$$P_t \mathbf{X}_t = P_{t-1}\mathbf{X}_t + \Omega_t G_t' \triangle_t^{-1}(\mathbf{Y}_t - G_t \hat{\mathbf{X}}_t)$$

and $\Omega_{t|t} = \Omega_t - \Omega_t G_t' \triangle_t^{-1} G_t \Omega_t'$.

□

8.3 The Holt-Winters Seasonal Algorithm

Go back to Table of Contents. Please click TOC

8.3.1 The Holt-Winters Seasonal Algorithm

Go back to Table of Contents. Please click TOC

If the series $Y_1, Y_2, ..., Y_n$ contains not only trend, but also seasonality with period d (as in model $X_t = m_t + s_t + Y_t$ from Section 1.5.2), then a further generalization of the forecast function $P_n Y_{n+h} = \hat{m}_n$ that takes this into account is

$$P_n Y_{n+h} = \hat{a}_n + \hat{b}_n h + \hat{c}_{n+h}, \ h = 1, 2, ...,$$

where $\hat{a}_n$, $\hat{b}_n$, and $\hat{c}_n$ can be thought of as estimates of the "trend level" a_n, "trend slope" b_n, and "seasonal component" c_n at time n. If k is the smallest integer such that $n + h - kd \leq n$, then we set

$$\hat{c}_{n+h} = \hat{c}_{n+h-kd}, \ h = 1, 2, ...,$$

while the values of $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$, $i = d+2, ..., n$, are found from recursions analogous to $\hat{a}_{n+1} = \alpha Y_{n+1}(1-\alpha)(\hat{a}_n + \hat{b}_n)$ and $\hat{b}_{n+1} = \beta(\hat{a}_{n+1} - \hat{a}_n) + (1-\beta)\hat{b}_n$, namely,

$$\hat{a}_{n+1} = \alpha(Y_{n+1} - \hat{c}_{n+1-d}) + (1-\alpha)(\hat{a}_n + \hat{b}_n),$$

$$\hat{b}_{n+1} = \beta(\hat{a}_{n+1} - \hat{a}_n) + (1-\beta)\hat{b}_n,$$

and

$$\hat{c}_{n+1} = \gamma(Y_{n+1} - \hat{a}_{n+1}) + (1-\gamma)\hat{c}_{n+1-d},$$

with initial conditions

$$\hat{a}_{d+1} = Y_{d+1},$$

$$\hat{b}_{d+1} = (Y_{d+1} - Y_t)/d,$$

and

$$\hat{c}_i = Y_i - (Y_1 + \hat{b}_{d+1}(i-1)), \ i = 1, ..., d+1.$$

Then $\hat{a}_{n+1}$, $\hat{b}_{n+1}$, and $\hat{c}_{n+1}$ above can be solved successively for $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$, $i = d+1, ..., n$, and the predictors $P_n Y_{n+h}$ found from $P_n Y_{n+h} = \hat{a}_n + \hat{b}_n h + \hat{c}_{n+h}$.

8.3.2 Holt-Winters Seasonal and ARIMA Forecasting

Go back to Table of Contents. Please click TOC

As in Section 9.2.2, the Holt-Winters seasonal recursions with seasonal period d correspond to the large-sample forecast recursions of an ARIMA process, in this case defined by

$$\begin{aligned}(1-B)(1-B^d)Y_t \quad = \quad & Z_t + \cdots + Z_{t-d+1} + \gamma(1-\alpha)(Z_{t-d} - Z_{t-d-1}) \\ & -(2-\alpha-\alpha\beta)(Z_{t-1} + \cdots + Z_{t-d}) \\ & +(1-\alpha)(Z_{t-2} + \cdots + Z_{t-d-1}),\end{aligned}$$

where $\{Z_t\} \sim WN(0, \sigma^2)$. Holt-Winters seasonal forecasting with optimal α, β, and γ can therefore be viewed as fitting a member of this four-parameter family of ARIMA models and using the corresponding large-sample forecast recursions.

8.4 Choosing a Forecasting Algorithm

Go back to Table of Contents. Please click TOC
Real data are rarely if ever generated by a simple mathematical model such as an ARIMA process. Forecasting methods that are predicated on the assumption of such a model are therefore not necessarily the best, even in the mean squared error sense. Nor is the measurement of error in terms of mean squared error necessarily always the most appropriate one in spide of its mathematical convenience. Even within the framework of minimum mean squared-error forecasting, we may ask (for example) whether we wish to minimize the one-step, two-step, or twelve-step mean squared error.

The use of more heuristic algorithms such as those discussed in this chapter is therefore well worth serious consideration in practical forecasting problems. But how do we decide which method to use? A relatively simple solution to this problem, given the availability of a substantial historical record, is to choose among competing algorithms by comparing the relevant errors when the algorithms are applied to the data already observed (e.g., by comparing the mean absolute percentage errors of the twel-step predictors of the historical data if twelve-step prediction is of primary concern).

9 Further Topics

Go back to Table of Contents. Please click TOC

In this final chapter we touch on a variety of topics of special interest. In Section 10.1 we consider transfer function models, designed to exploit for predictive purposes the relationship between two time series when one acts as a leading indicator for the other. Section 10.2 deals with intervention analysis, which allows for possible changes in the mechanism generating a time series, causing it to have different properties over different time intervals. in Section 10.3 we introduce the very fast growing area of nonlinear time series analysis, and in Section 10.4 we briefly discuss continuous-time ARMA processes, which besides being of itnerest in their own right, are very useful also for modeling irregularly spaced data. In Section 10.5 we discuss fractionally integrated ARMA processes, sometimes called "long-memory" processes on account of the slow rate of convergence of their autocorrelation functions to zero as the lag increases.

9.1 Transfer Function Models

Go back to Table of Contents. Please click TOC In this section we consider the problem of estimating the transfer function of a linear filter when the output includes added uncorrelated noise. Suppose that $\{X_{t1}\}$ and $\{X_{t2}\}$ are, respectively, the input and output of the trasnfer function model

$$X_{t2} = \sum_{j=0}^{\infty} \tau_j X_{t-j,1} + N_t,$$

where $T = \{\tau_j, j = 0, 1, ...,\}$ is a causal time-invariant linear filter and $\{N_t\}$ is a zero-mean stationary process, uncorrelated with the input process $\{X_{t1}\}$. We further assume that $\{X_{t1}\}$ is a zero=-mean stationary time series. Then the bivariate process $\{(X_{t1}, X_{t2})'\}$ is also stationary. Multiplying each side of $X_{t2} = \sum_{j=0}^{\infty} \tau_j X_{t-j,1} + N_t$ by $X_{t-k,1}$ and then taking expectations gives the equation

$$\tau_{21}(k) = \sum_{j=0}^{\infty} \tau_j \gamma_{11}(k-j).$$

Equation $\gamma_{21}(k)$ simplifies a great deal if the input process happens to be white noise. For example, if $\{X_{t1}\} \sim WN(0, \sigma_1^2)$, then we can immediately identify t_k from $\gamma_{21}(k) = \sum_{j=0}^{\infty} \tau_j \gamma_{11}(k-j)$ as

$$\tau_k = \gamma_{21}(k)/\sigma_1^2.$$

This observation suggests that "prewhitening" of the input process might simplify the identification of an appropriate transfer function model and at the same time provide simple preliminary estimates of the coefficients t_k.

If $\{X_{t1}\}$ can be represented as an invertible ARMA(p,q) process

$$\phi(B)X_{t1} = \theta(B)Z_t, \ \{Z_t\} \sim WN(0, \sigma_Z^2),$$

then application of the filter $\pi(B) = \phi(B)\theta^{-1}(B)$ to $\{X_{t1}\}$ will produce the whitened series $\{Z_t\}$. Now applying the operator $\pi(B)$ to each side of $X_{t2} = \sum_{j=0}^{\infty} \tau_j X_{t-j,1} + N_t$ and letting $Y_t = \pi(B)X_{t2}$, we obtain the relation

$$Y_t = \sum_{j=0}^{\infty} \tau_j Z_{t-j} + N_t',$$

where

$$N_t' = \pi(B)N_t,$$

and $\{N_t'\}$ is a zero-mean stationary process, uncorrelated with $\{Z_t\}$. The same arguments that led to $\tau_k = \gamma_{21}(k)/\sigma_1^2$ therefore yield the equation

$$\tau_j = \rho_{YZ}(j)\sigma_Y/\sigma_Z,$$

where ρ_{YZ} is the cross-correlation function of $\{Y_t\}$ and $\{Z_t\}$, $\sigma_Z^2 = Var(Z_t)$, and $\sigma_Y^2 = Var(Y_t)$.

Given the observations $\{(X_{t1}, X_{t2})', t = 1, ..., n\}$, the results of the previous paragraph suggest the following procedure for estimating $\{\tau_j\}$ and analyzing the noise $\{N_t\}$ in the model $X_{t2} = \sum_{j=0}^{\infty} \tau_j X_{t-j,1} + N_t$:

(1) Fit an ARMA model to $\{X_{t1}\}$ and file the residuals $(\hat{Z}_1, ..., \hat{Z}_n)$. Let $\hat{\phi}$ and $\hat{\theta}$ denote the maximum likelihood estimates of the autoregressive and moving-average parameters and let σ_Z^2 be the maximum likelihood estimate of the $\{Z_t\}$.

(2) Apply the operator $\hat{\pi}(B) = \hat{\phi}(B)\hat{\theta}^{-1}(B)$ to $\{X_{t2}\}$ to obtain the series $(\hat{Y}_1, ..., \hat{Y}_n)$. (After fitting the ARMA model as in Step 1 above, highlight the window containing the graph of $\{X_t\}$ and replace $\{X_t\}$ by $\{Y_t\}$. The residuals are then automatically replaced by the residuals of $\{Y_t\}$ under the model already fitted to $\{X_t\}$.) Let $\hat{\sigma}_Y^2$ denote the sample variance of $\hat{Y}_t$.

(3) Compute the sample auto- and corss-correlation functions of $\{Z_t\}$ and $\{Y_t\}$. Comparison of $\hat{\rho}_{YZ}(h)$ with the bounds $\pm 1.96n^{-1/2}$ gives a preliminary indication of the lags h at which $\rho_{YZ}(h)$ is significantly different from zero. A more refined check can be carried out by using Bartlett's formula in Section 7.3.4 for the asymptotic variance of $\hat{\rho}_{YZ}(h)$. UNder the assumptions that $\{\hat{Z}_t\} \sim WN(0, \sigma_Z^2)$ and $\{(\hat{Y}_t, \hat{Z}_t)'\}$ is a stationary Gaussian process,

$$nVar(\hat{\rho}_{YZ}(h)) \sim 1 - \rho_{YZ}^2(h)\left[1.5 - \sum_{k=-\infty}^{\infty}(\rho_{YZ}^2(k) + \rho_{YY}^2(k)/2)\right]$$

$$+ \sum_{k=-\infty}^{\infty}[\rho_{YZ}(h+k)\rho_{YZ}(h-k) - 2\rho_{YZ}(h)\rho_{YZ}(k+h)\rho_{YY}^2(k)].$$

In order to check the hypothesis H_0 that $\rho_{YZ}(h) = 0, \ h \notin [a, b]$, where a and b are integers, we note from Corollary 7.13 (see remark, also Corolarry 7.3.1 from text [9]) that under H_0,

$$Var(\hat{\rho}_{YZ}(h)) \sim n^{-1} \ for \ h \notin [a, b].$$

We can therefore check the hypothesis H_0 by comparing $\hat{\rho}_{YZ}, \ h \notin [a, b]$, with the bounds $\pm 1.96n^{-1/2}$. Observe that $\rho_{YZ}(h)$ should be zero for $h > 0$ if the model $X_{t2} = \sum_{j=0}^{\infty} \tau_j X_{t-j,1} + N_t$ is valid.

Remark 9.1. Recall Corollary 7.13: If $\{\mathbf{X}_t\}$ satisfies the conditions for Bartlett's formula, if either $\{X_{t1}\}$ or $\{X_{t2}\}$ is white noise, and if

$$\rho_{12}(h) = 0, \ h \notin [a, b],$$

then

$$\lim_{n\to\infty} nVar(\hat{\rho}_{12}(h)) = 1, \ h \notin [a, b].$$

□

(4) Preliminary estimates of τ_h for the lags h at which $\hat{\rho}_{YZ}(h)$ is significantly different from zero are

$$\hat{\tau}_h = \hat{\rho}_{YZ}(h)\hat{\sigma}_Y/\hat{\sigma}_Z.$$

For other values of h the preliminary estimates are $\hat{\tau}_h = 0$. The numerical value sof the cross-correlations $\hat{\rho}_{YZ}(h)$ are found on the graphs of the sample correlations. The values of $\hat{\sigma}_Z$ and $\hat{\sigma}_Y$ are found similaryly. Let $m \geq 0$ be the largest value of j such that $\hat{\tau}_j$ is nonzero and let $b \geq 0$ be the smallest such value. Then b is known as the *delay parameter* of the filter $\{\hat{\tau}_j\}$. If m is very large and if the coefficients $\{\hat{\tau}_j\}$ are approximately related by difference equations of the form

$$\hat{\tau}_j - v_1\hat{\tau}_{j-1} - \cdots - v_p\hat{\tau}_{j-p} = 0,\ j \geq b+p,$$

then $\hat{T}(B) = \sum\limits_{j=b}^{m} \hat{\tau}_j B^j$ can be represented approximately, using fewer parameters, as

$$\hat{T}(B) = w_0(1 - v_1 B - \cdots - v_p B_p)^{-1} B^b.$$

In particular, if $\hat{\tau}_j = 0$, $j < b$, and $\hat{\tau}_j = w_0 v_1^{j-b}$, $j \geq b$, then

$$\hat{T}(B) = w_0(1 - v_1 B)^{-1} B^b.$$

Box and Jenkins (1976) [6] recommend choosing $\hat{T}(B)$ to be a ratio of two polynomials. However, the degrees of the polynomial are often difficult to estimate from $\{\hat{\tau}_j\}$. The primary objective at this stage is to find a parametric function that provides an adequate approximation to $\hat{T}(B)$ without introducing too large a number of parameters. If $\hat{T}(B)$ is represented as $\hat{T}(B) = B^b w(B) v^{-1}(B) = B^b(w_0 + w_1 B + \cdots + w_q B^q)(1 - v_1 B - \cdots - v_p B^p)^{-1}$ with $v(z) \neq 0$ for $|z| \leq 1$, then we define $m = \max(q+b, p)$.

(5) The noise sequence $\{N_t, t = m+1, ..., n\}$ is estimated as

$$\hat{N}_t = X_{t2} - \hat{T}(B) X_{t1}.$$

(6) Preliminary identification of a sutiable model for the noise sequence is carried out by fitting a causal invertible ARMA model

$$\phi^{(N)}(B) N_t = \theta^{(N)}(B) W_t,\ \{W_t\} \sim WN(0, \sigma_W^2),$$

to the estimated noise $\hat{N}_{m+1}, ..., \hat{N}_n$.

(7) At this state we have the preliminary model

$$\phi^{(N)}(B) v(B) X_{t2} = B^b \phi^{(N)}(B) w(B) X_{t1} + \theta^{(N)}(B) v(B) W_t,$$

where $\hat{T}(B) = B^b w(B) v^{-1}(B)$ as in step (4). For this model we can compute $\hat{W}_t(\mathbf{w}, \mathbf{v}, \phi^{(N)}, \theta^{(N)})$, $t > m^* = \max(p_2 + p, b + p_2 + q)$, by setting $\hat{W}_t = 0$ for $t \leq m^*$. The parameters $\mathbf{w}$, $\mathbf{v}$, $\phi^{(N)}$, and $\theta^{(N)}$ can then be reestimated (more efficiently) by minimizing the sum of squares

$$\sum_{t=m^*+1}^{n} \hat{W}_t^2(\mathbf{w}, \mathbf{v}\phi^{(N)}, \theta^{(N)}).$$

(8) To test for goodness of fit, the estimated residuals $\{\hat{W}_t, t > m^*\}$ and $\{\hat{Z}_t, t > m^*\}$ should be filed as a bivariate series and the auto- and crosscorrelations compared with the bounds $\pm 1.96/\sqrt{n}$ in order to check the hypothesis that the two series are uncorrelated white noise sequences. Alternative mdoels can be compared using the AICC value that is printed with the estimated parameters in step (7).

9.1.1 Prediction Based on a Transfer Function Model

Go back to Table of Contents. Please click TOC

When predicting $X_{n+h,2}$ on the basis of the transfer function model defined by $_{t2} = \sum_{j=0}^{\infty} \tau_j X_{t-j,1} + N_t$, $\phi(B)X_{t1} = \theta(B)Z_t$, and $\phi^{(N)}(B)N_t = \theta^{(N)}(B)W_t$, with observations of X_{t1} and X_{t2}, $t = 1, ..., n$, our aim is to find the linear combination of $1, X_{11}, ..., X_{n1}, X_{12}, ..., X_{n2}$ that predicts $X_{n+h,2}$ with minimum mean squared error.

In order to provide a little more insight, we give here the predictors $\tilde{P}_n X_{n+h}$ and mean squared error based on infinitely many past observations X_{t1} and X_{t2}, $-\infty < t \leq n$. These predictors and their mean squared errors will be close to those based on X_{t1} and X_{t2}, $1 \leq t \leq n$, if n is sufficiently large.

The transfer function model can be rewritten as

$$X_{t2} = T(B)X_{t1} + \beta(B)W_t,$$

$$X_{t1} = \theta(B)\phi^{-1}(B)Z_t,$$

when $\beta(B) = \theta^{(N)}(B)/\phi^{(N)}(B)$. Eliminating X_{t1} gives

$$X_{t2} = \sum_{j=0}^{\infty} \alpha_j Z_{t-j} + \sum_{j=0}^{\infty} \beta_j W_{t-j},$$

where $\alpha(B) = T(B)\theta(B)/\phi(B)$.

Noting that each limit of linear combinations of $\{X_{t1}, X_{t2}, -\infty < t \leq n\}$ is a limit of linear combinations of $\{Z_t, W_t, -\infty < t \leq n\}$ and conversely and that $\{Z_t\}$ and $\{W_t\}$ are uncorrelated, we see at once from $X_{t2} = \sum_{j=0}^{\infty} \alpha_j Z_{t-j} + \sum_{j=0}^{\infty} \beta_j W_{t-j}$ that

$$\tilde{P}_n X_{n+h,2} = \sum_{j=h}^{\infty} \alpha_j Z_{n+h-j} + \sum_{j=h}^{\infty} \beta_j W_{n+h-j}.$$

Setting $t = n + h$ in $X_{t2} = \sum_{j=0}^{\infty} \alpha_j Z_{t-j} + \sum_{j=0}^{\infty} \beta_j W_{t-j}$ and subtracting $\tilde{P}_n X_{n+h,2} = \sum_{j=h}^{\infty} \alpha_j Z_{n+h-j} + \sum_{j=h}^{\infty} \beta_j W_{n+h-j}$ gives the mean squared error

$$\mathbb{E}(X_{n+h,2} - \tilde{P}_n X_{n+h,2})^2 = \sigma_Z^2 \sum_{j=0}^{h-1} \alpha_j^2 + \sigma_W^2 \sum_{j=0}^{h-1} \beta_j^2.$$

To compute the predictors $\tilde{P}_n X_{n+h,2}$ we proceed as follows. Rewrite $X_{t2} = T(B)X_{t1} + \beta(B)W_t$ as

$$A(B)X_{t2} = B^b U(B)X_{t1} + V(B)W_t,$$

where A, U, and V are polynomials of the form

$$A(B) = 1 - A_1 B - \cdots - A_a B^a,$$

$$U(B) = U_0 + U_1 B + \cdots + U_u B^u,$$
$$V(B) = 1 + V_1 B + \cdots + V_v B^v.$$

Applying the operator $\tilde{P}_n$ to equation $A(B)X_{t2} = B^b U(B)X_{t1} + V(B)W_t$ with $t = n + h$, we obtain

$$\tilde{P}_n X_{n+h,2} = \sum_{j=1}^{a} A_j \tilde{P}_n X_{n+h-j,2} + \sum_{j=0}^{u} U_j \tilde{P}_n X_{n+h-b-j,1} + \sum_{j=h}^{v} V_j W_{n+h-j},$$

where the last sum is zero if $h > v$.

Since $\{X_t\}$ is uncorrelated with $\{W_t\}$, the predictors appearing in the second sum in $\tilde{P}_n X_{n+h,2}$ are therefore obtained by predicting the univariate series $\{X_{t1}\}$ as described in Section 3.3 using the model $X_{t1} = \theta(B)\phi^{-1}(B)Z_t$.

The model $\tilde{P}_n X_{n+h,2}$ can now be sovled recursively for the predictors $\tilde{P}_n X_{n+1,2}, \tilde{P}_n X_{n+2,2}, \tilde{P}_n X_{n+3,2}, \ldots$.

9.2 Intervention Analysis

Go back to Table of Contents. Please click TOC

During the period for which a time series is observed, it is sometimes the case that a change occurs that affects the level of the series. A change in the tax laws may, for example, have a continuing effect on the daily closing prices of shares on the stock market. In the same way construction of a dam on a river may have a dramatic effect on the time series of streamflows below the dam. In the following we shall assume that time T at which the change (or "intervention") occurs is known.

To account for such changes, Box and Tiao (1975) [7] introduced a model for intervention analysis that has the same form as the trasnfer function model

$$Y_t = \sum_{j=0}^{\infty} \tau_j X_{t-j} + N_t,$$

except that the input series $\{X_t\}$ is not a random series but a deterministic function of t. It is clear from Y_t that $\sum_{j=0}^{\infty} \tau_j X_{t-j}$ is then the mean of Y_t. The function $\{X_t\}$ and the coefficients $\{\tau_j\}$ are therefore chosen in such a way that the changing level of the observations of $\{Y_t\}$ is well represented by the sequence $\sum_{j=0}^{\infty} \tau_j X_{t-j}$. For a series $\{Y_t\}$ with $\mathbb{E}(Y)_t = 0$ for $t \leq T$ and $\mathbb{E}(Y_t) \to 0$ as $t \to \infty$, a suitable input series is

$$X_t = I_t(T) = \begin{cases} 1 & if\ t = T, \\ 0 & if\ t \neq T. \end{cases}$$

For a series $\{Y_t\}$ with $\mathbb{E}(Y_t) = 0$ for $t \leq T$ and $\mathbb{E}(Y_t) \to a \neq 0$ as $t \to \infty$, a suitable input series is

$$X_t = H_t(T) = \sum_{k=T}^{\infty} I_t(k) = \begin{cases} 1 & if\ t \geq T, \\ 0 & if\ t < t. \end{cases}$$

(Other deterministic input functions $\{X_t\}$ can also be used, for example when interventions occur at more than one time.) The function $\{X_t\}$ having been selected by inspection of the data, the determination of the coefficients $\{\tau_j\}$ in Y_t then reduces to a regression problem in which the errors $\{N_t\}$ constitue an ARMA process.

9.3 Nonlinear Models

Go back to Table of Contents. Please click TOC

A time series of the form

$$X_t = \sum_{j=0}^{\infty}\psi_j Z_{t-j},\ \{Z_t\} \sim IID(0,\sigma^2),$$

where Z_t is expressible as a mean square limit of linear combinations of $\{X_s, \infty < s \leq t\}$, has the property that the best mean square predictor $\mathbb{E}(X_s, -\infty < s \leq t)$ and the best linear predictor $\tilde{P}_t X_{t+h}$ in terms of $\{X_s, -\infty < s \leq t\}$ are identical. It can be shown that if iid is repaced by WN in $X_t = \sum\limits_{j=0}^{\infty}\psi_j Z_{t-j}$, then the two predictors are identical if and only if $\{Z_t\}$ is a **martingale difference sequence** relative to $\{X_t\}$, i.e., if and only if $\mathbb{E}(Z_t | X_s, -\infty < s \leq t) = 0$ for all t.

The Wold decomposition (Section 2.6) ensures that every purely nondeterministic stationary process can be expressed in the form $X_t = \sum\limits_{j=0}^{\infty}\psi_j Z_{t-j}$ with $\{Z_t\} \sim WN(0,\sigma^2)$. The process $\{Z_t\}$ in the Wold decompoistion, however, is generally not an iid sequence, and the best mean square predictor of X_{t+h} may be qutie different from the best linear predictor. In the case where $\{X_t\}$ is a purely nondeterministic Gaussian stationary process, the sequence $\{Z_t\}$ in the Wold decomposition is Gaussian and therefore iid. Every stationary purely nondeterministic Gaussian process can therefore be generated by applying a causal linear filter to an iid Gaussian sequence. We shall therefore refer to such a process as a **Gaussian linear process**.

In this section we shall use the term linear process to mean a process $\{X_t\}$ of the form $X_t = \sum\limits_{j=0}^{\infty}\psi_j Z_{t-j}$. This is a more restrictive use of the term than in Definition 2.10 (see remark, also Definition 2.2.1 from text [9]).

Remark 9.2. Recall Definition 2.10:
The time series $\{X_t\}$ is a **linear process** if it has the representation

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

for all t, where $\{Z_t\} \sim WN(0,\sigma^2)$ and $\{\psi_j\}$ is a sequence of constants with $\sum\limits_{j=-\infty}^{\infty}|\psi_j| < \infty$.

9.3.1 Deviations from Linearity

Go back to Table of Contents. Please click TOC

Many of the time series encountered in practice exhibit characteristics not shown by linear processes, and so to obtain good model and predictors it is necessary to look to model more general than those satisfying $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ with iid noise. As indicated above, this will mean that the minimum mean squared error predictors are not, in general, linear functions of the past observations.

Gaussian linear processes have a number of properties that are often found to be violated by observed time series. The former are reversible in the sense that $(X_{t_1}, ..., X_{t_n})'$ has the same distribution as $(X_{t_n}, ..., X_{t_1})'$. (Except in a few special cases, ARMA processes are reversible if and only if they are Gaussian (Breidt and Davis, 1992). [8]) Deviations from this property by observed time series are suggested by sample paths that rise to their maxima and fall away at different rates. Bursts of outlying values are frequently observed in practical time series and are seen also in the sample paths of nonlinear (and infinite-variance) models. They are rarely seen, however, in the sample paths of Gaussian linear processes. Other characteristics suggesting deviation from a Gaussian linear model are discussed by Tong (1990) [35].

9.3.2 Chaotic Deterministic Sequences

Go back to Table of Contents. Please click TOC

To distinguish between linear and nonlinear processes, we need to be able to decide in particular when a white noise sequence is also iid. Sequences generated by nonlinear deterministic difference equations can exhibit sample correlation functions that are very close to those of samples from white noise sequence. However, the deterministic nature of the recursions implies the strongest possible dependence between successive observations.

9.3.3 Distinguishing Between Whtie Noise and iid Sequences

Go back to Table of Contents. Please click TOC

If $\{X_t\} \sim WN(0, \sigma^2)$ and $\mathbb{E}|X_t|^4 < \infty$, a useful tool for deciding whether or not $\{X_t\}$ is iid is the ACF $\rho_{X^2}(h)$ of the process $\{X_t^2\}$. If $\{X_t\}$ is iid, then $\rho_{X^2}(h) = 0$ for all $h \neq 0$, whereas this is not necessarily the case otherwise. This is the basis for the test of mcLeod and Li described in section 1.6.

Now suppose that $\{X_t\}$ is a strictly stationary time series such that $\mathbb{E}|X_t|^k \leq K < \infty$ for some itneger $k \geq 3$. The kth-order cumulant $C_k(r_1, ..., r_{k-1})$ of $\{X_t\}$ is then defined as the joint cumulant of the random variables, X_t, $X_{t+r_1}, ..., X_{t+r_{k-1}}$, i.e., as the coefficient of $i^k z_1 z_2 \ldots z_k$ in the Taylor expansion about (0, ..., 0) of

$$\chi(z_1, ..., z_k) := \ln \mathbb{E}[\exp(iz_1 X_t + iz_2 X_{t+r_1} + \cdots + iz_k X_{t+r_{k-1}})].$$

(Since $\{X_t\}$ is strictly stationary, this quantity does not depend on t.) In particular, the third-order cumulant function C_3 of $\{X_t\}$ coincides with the third-order central moment function, i.e.,

$$C_3(r,s) = \mathbb{E}[(X_t-\mu)(X_{t+r}-\mu)(X_{t+s}-\mu)],\ r,s\in\{0,\pm 1,...\},$$

where $\mu = \mathbb{E}(X_t)$. If $\sum_r \sum_s |C_3(r,s)| < \infty$, we define the third-order polyspectral density (or bispectral density) of $\{X_t\}$ to be the Fourier trasnform

$$f_3(\omega_1,\omega_2) = \frac{1}{(2\pi)^2}\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty} C_3(r,s)e^{-ir\omega_1-is\omega_2}, -\pi\le\omega_1,\omega_2\le\pi,$$

in which case

$$C_3(r,s) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{ir\omega_1+is\omega_2} f_3(\omega_1,\omega_2)d\omega_1 d\omega_2.$$

[More generally, if the kth order cumulants $C_k(r_1,...,r_{k-1})$, of $\{X_t\}$ are absolutely summable, we define the kth order polyspectral density as the Fourier transform of C_k. For details see Rosenblatt (1985) [29] and Priestley (1988) [28].]

If $\{X_t\}$ is a Guassian linear process, it follows from Problem 10.3 in text [9] that the cumulant function C_3 of $\{X_t\}$ is identically zero. (The same is also true of all the cumulant functions C_k with $k > 3$.) Consequently, $f_3(\omega_1,\omega_2) = 0$ for all $\omega_1,\omega_2 \in [-\pi,\pi]$. Appropriateness of a Gaussian linear model for a given data set can therefore be checked by using the data to test the null hypothesis $f_3 = 0$. For details of such a test, see Subba-Rao and Gabr (1984) [34].

If $\{X_t\}$ is a linear process of the form $X_t = \sum\limits_{j=0}^{\infty}\psi_j Z_{t-j}$, $\{Z_t\} \sim IID(0,\sigma^2)$ with $\mathbb{E}(|Z_t|)^3 < \infty$, $\mathbb{E}(Z_t^3) = \eta$, and $\sum\limits_{j=0}^{\infty}|\psi_j| < \infty$, it can be shown from $\chi(z_1,...,z_k) := \ln\mathbb{E}[\exp(iz_1X_t + iz_2X_{t+r_1} + \cdots + iz_kX_{t+r_{k-1}})]$ that the third-order cumulant function of $\{X_t\}$ is given by

$$C_3(r,s) = \eta\sum_{i=-\infty}^{\infty}\psi_i\psi_{i+r}\psi_{k+s}$$

(with $\psi_j = 0$ for $j < 0$), and hence that $\{X_t\}$ has bispectral density

$$f_3(\omega_1,\omega_2) = \frac{\eta}{4\pi^2}\psi(e^{i(\omega_1+\omega_2)})\psi(e^{-i\omega_1})\psi(e^{-i\omega_2}),$$

where $\psi(z) := \sum\limits_{j=0}^{\infty}\psi_j z^j$. By Proposition 4.19 (see the remark below, also see Proposition 4.3.1 in text [9]), the spectral density of $\{X_t\}$ is

$$f(\omega) = \frac{\sigma^2}{2\pi}|\psi(e^{-i\omega})|^2.$$

Hence,

$$\phi(\omega_1,\omega_2) := \frac{|f_3(\omega_1,\omega_2)|^2}{f(\omega_1)(f(\omega_2)(f(\omega_1+\omega_2)} = \frac{\eta^2}{2\pi\sigma^6}.$$

Appropriateness of the linear process for modeling a given data set can therefore be checked by using the data to test for constancy of $\phi(\omega_1,\omega_2)$ (see Subba-Rao and Gabr, 1984) [34].

Remark 9.3. Recall the following proposition: Let $\{X_t\}$ be a stationary time series with mean zero and spectral density $f_X(\lambda)$. Suppose that $\Psi = \{\psi_j,\ j = 0, \pm 1, ...\}$ is an absolutely summable TLF (i.e., $\sum\limits_{j=-\infty}^{\infty} \Psi_j| < \infty$). Then the time series

$$Y_t = \sum_{j=-\infty}^{\infty} \Psi_j X_{t-j}$$

is stationary with mean zero and spectral density

$$f_Y(\lambda) = |\Psi(e^{-i\lambda})|^2 f_X(\lambda) = \Psi(e^{-i\lambda})\Psi(e^{i\lambda}) f_X(\lambda),$$

where $\Psi(e^{-i\lambda}) = \sum\limits_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda}$. (The function $\Psi(e^{-i\cdot})$ is called the **transfer function** of the filter, and the squared modulus $|\Psi(e^{-i\cdot})|^2$ is referred to as the **power transfer function** of the filter.)

□

9.3.4 Three Useful Classes Nonlinear Models

Go back to Table of Contents. Please click TOC

If a linear Gaussian model is not appropriate, there is a choice of several families of nonlinear processes that have been found useful for modeling purposes. These include bilinear models, autoregressive models with random coefficients, and threshold models.

The bilinear model of order (p, q, r, s) is defined by the equations

$$X_t = Z_t + \sum_{i=1}^{p} a_i X_{t-i} + \sum_{j=1}^{q} b_j Z_{t-j} + \sum_{i=1}^{r}\sum_{j=1}^{s} c_{ij} X_{t-i} Z_{t-i},$$

where $\{Z_t\} \sim iid(0, \sigma^2)$. A sufficient condition for the existence of strictly stationary solution of these equations is given by Liu and Brockwell (1988) [22].

A random coefficient autoregressive process $\{X_t\}$ of order p satisfies an equation of the form

$$X_t = \sum_{i=1}^{p} (\phi_i + U_t^{(i)}) X_{t-i} + Z_t,$$

where $\{Z_t\} \sim IID(0, \sigma^2)$, $\{U_t^{(i)}\} \sim IID(0, v^2)$, $\{Z_t\}$ is independent of $\{U_t\}$, and $\phi_1, ..., \phi_p \in \mathbb{R}$.

Threshold models can be regarded as piecewise linear models in which the linear relationship varies with the values of the process. For example, if $R^{(i)}$, $i = 1, ..., k$ is a partition of $\mathbb{R}^p$, and $\{Z_t\} \sim IID(0, 1)$, then the k difference equations

$$X_t = \sigma^{(i)} Z_t + \sum_{j=1}^{p} \phi_j^{(i)} X_{t-j},\ (X_{t-1}, ..., X_{t-p}) \in R^{(i)},\ i = 1, ..., k,$$

define a threshold AR(p) model. Model identification and parameter estimation for threshold models can be carried out in a manner similar to that for linear models using amximum likelihood and the IAC criterion.

9.3.5 Modeling Volatility

Go back to Table of Contents. Please click TOC
For modeling changing volatility as discussed above under deviations from linearity, Engle (1982) introduced the **ARCH(p) process** $\{X_t\}$ as a solution of the equations

$$Z_t = \sqrt{h_t} e_t, \ \{e_t\} \sim IID \ N(0,1),$$

where h_t is the (positive) function of $\{Z_t, \ s < t\}$, defined by

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Z_{t-i}^2,$$

with $\alpha_0 > 0$ and $\alpha_j \geq 0$, $j = 1, ..., p$. The name ARCH signifies autoregressive conditional heteroscedasticity with h_t the conditional variance of Z_t given $\{Z_s, s < t\}$.

The simplest such process is the ARCH(1) process. In this case the recursions of $Z_t = \sqrt{h_t} e_t$ and $h_t = \alpha_0 + \sum\limits_{i=1}^{p}$ give

$$\begin{aligned}
Z_t^2 &= \alpha_0 e_t^2 + \alpha_1 Z_{t-1}^2 e_t^2 \\
&= \alpha_0 e_t^2 + \alpha_1 \alpha_0 e_t^2 e_{t-1}^2 + \alpha_1 Z_{t-2}^2 e_t^2 e_{t-1}^2 \\
&= \dots \\
&= \alpha_0 \sum_{j=0}^{n} \alpha_1^j e_t^2 e_{t-1}^2 \dots e_{t-j}^2 + \alpha_1^{n+1} Z_{t-n-1}^2 e_t^2 e_{t-1}^2 \dots e_{t-n}^2.
\end{aligned}$$

If $|\alpha_1| < 1$ and $\{Z_t\}$ is stationary and causal (i.e., Z_t is a function of $\{e_s, s \leq t\}$), then the last term has expectation $\alpha^{n+1} \mathbb{E}(Z_t^2)$ and consequently (by the Borel-Cantelli lemma) converges to zero with probability one as $n \to \infty$. The first term converges with probability one, and hence

$$Z_t^2 = \alpha_0 / (1 - \alpha_1).$$

Since

$$Z_t = e_t \sqrt{\alpha) \left(1 + \sum_{j=1}^{\infty} \alpha_1^j e_{t-1}^2 \dots e_{t-j}^2 \right)},$$

it is clear that $\{Z_t\}$ is strictly stationary and hence, since $\mathbb{E}(Z_t^2) < \infty$, also stationary in the weak sense. We have now the solution of the **ARCH(1) equations**:

Theorem 9.4. *If $|\alpha_1| < 1$, the unique causal stationary solution of the ARCH(1) equations is given by $Z_t = e_t \sqrt{\alpha) \left(1 + \sum\limits_{j=1}^{\infty} \alpha_1^j e_{t-1}^2 \dots e_{t-j}^2 \right)}$. It has the properties*

$$\begin{aligned}
\mathbb{E}(Z_t) &= \mathbb{E}(\mathbb{E}(Z_t | e_s, \ s < t)) = 0, \\
Var(Z_t) &= \alpha_0 / (1 - \alpha_1), \\
\mathbb{E}(Z_{t+h} Z_t) &= \mathbb{E}(\mathbb{E}(Z_{t+H} Z_t | e_s, \ s < t + h)) = 0 \ for \ h > 0.
\end{aligned}$$

Thus the ARCH(1) process with $|\alpha_1| < 1$ is strictly stationary white noise. However, it is not an iid sequence, since from $Z_t = \sqrt{h_t}e_t$ and $h_t = \alpha_0 + \sum\limits_{i=1}^{p} \alpha_i Z_{t-i}^2$,

$$\mathbb{E}(Z_t^2|Z_{t-1}) = (\alpha_0 + \alpha_1 Z_{t-1}^2)\mathbb{E}(e_t^2|Z_{t-1}) = \alpha_0 + \alpha_1 Z_{t-1}^2.$$

This also shows that $\{Z_t\}$ is not Gaussian, since strictly stationary white noise is necessarily iid. From $Z_t = e_t\sqrt{\alpha_0\left(1 + \sum\limits_{j=1}^{\infty} \alpha_1^j e_{t-1}^2 \dots e_{t-j}^2\right)}$, it is clear that the distribution of Z_t is symmetric, i.e., that Z_t and $-Z_t$ have the same distribution. From $Z_t^2 = \alpha_0 \sum\limits_{j=0}^{\infty} \alpha_1^j e_t^2 e_{t-1}^2 \dots e_{t-j}^2$ it is easy to calculate $\mathbb{E}(X_t^4)$ (see Problem 10.4 in text [9]) and hence to show that $\mathbb{E}(Z_t^4)$ is finite if and only if $3\alpha_1^2 < 1$. More generally (see Engle, 1982 [13]), it can be shown that for every α_1 in the interval $(0, 1)$, $\mathbb{E}(Z^{2k}) = \infty$ for some positive integer k. This indicates the "heavy-tailed" nature of the margainl distribution of Z_t. If $\mathbb{E}(Z_t^4) < \infty$, the squared process $Y_t = Z_t^2$ has the samea ACF as the AR(1) process $W_t \alpha_1 W_{t-1} + e_t$, a result that extends also to ARCH(p) processes (see Problem 10.5 in text [9]).

The ARCH(p) process is conditionally Gaussian, in the sense that for given values of $\{Z_s, s = t-1, t-2, ..., t-p\}$, Z_t is Gaussian with known distribution. This makes it easy to wrtie down the likelihood of $Z_{p+1}, ..., Z_n$ conditional on $\{Z_1, ..., Z_p\}$ and hence, by numerical maximization, to compute conditional maximum likelihood estimates of the parameters. For example, the conditional likelihood of observations $\{z_2, ..., z_n\}$ of the ARCH(1) process given $Z_1 = z_1$ is

$$L = \prod_{t=2}^{n} \frac{1}{\sqrt{2\pi(\alpha_0 + \alpha_1 z_{t-1}^2}} \exp\left\{ - \frac{z_t^2}{2(\alpha_0 + \alpha_1 z_{t-1}^2} \right\}.$$

The **GARCH(p, q) process** (see Bollerslev, 1986 [5]) is a generalization of the ARCH(p) process in which the variance equation $h_t = \alpha_0 + \sum\limits_{i=1}^{p} \alpha_i Z_{t-i}^2$ is replaced by

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Z_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}^2,$$

with $\alpha_0 > 0$ and α_j, $\beta_j \geq 0$, $j = 1, 2, ...$

References

[1] Akaike, H. (1969), Fitting autoregressive models for prediction, *Annals of the Institute of Statistical Mathematics*, 21, 243247.

[2] Akaike, H. (1978), Time series analysis and control through parametric models, **Applied Time Series Analysis**, D.F. Findley (ed.), Academic Press, New York.

[3] Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, John Wiley, New York.

[4] Ansley, C.F. (1979), An algorithm for the exact likelihood of a mixed autoregressivemoving-average process, *Biometrika*, 66, 5965.

[5] Bollerslev, T. (1986), Generalized autoregressive conditional heteroskedasticity, *J. Econometrics*, 31, 307327.

[6] Box, G.E.P. and Jenkins, G.M. (1976), *Time Series Analysis: Forecasting and Control, Revised Edition*, Holden-Day, San Francisco.

[7] Box, G.E.P. and Tiao, G.C. (1975), Intervention analysis with applications to economic and environmental problems, *J. Amer. Stat. Assoc.* 70, 7079.

[8] Breidt, F.J. and Davis, R.A. (1992), Time reversibility, identifiability and independence of innovations for stationary time series, *J. Time Series Anal. 13*, 377390.

[9] Brockwell, J. Peter, and Davis, A. Richard, *Introduction to Time Series and Forecasting.*

[10] Brockwell, P.J. and Davis, R.A. (1988), Applications of innovation representations in time series analysis, *Probability and Statistics, Essays in Honor of Franklin A. Graybill*, J.N. Srivastava (ed.), Elsevier, Amsterdam, 6184.

[11] Cochrane, D. and Orcutt, G.H. (1949), Applications of least squares regression to relationships containing autocorrelated errors, *J. Amer. Stat. Assoc.*, 44, 3261.

[12] Dickey, D.A. and Fuller,W.A. (1979), Distribution of the estimators for autoregressive time series with a unit root, *J. Amer. Stat. Assoc.*, 74, 427431.

[13] Engle, R.F. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation, *Econometrica*, 50, 9871007.

[14] Engle, R.F. and Granger, C.W.J. (1987), Co-integration and error correction: representation, estimation and testing, *Econometrica*, 55, 251276.

[15] Granger, C.W.J. (1981), Some properties of time series data and their use in econometric model specification, *J. Econometrics*, 16, 121130.

[16] Hannan, E.J. (1980), The estimation of the order of an ARMA process, *Ann. Stat.*, 8, 10711081.

[17] Harvey, A.C. (1990), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

[18] Holt, C.C. (1957), Forecasting seasonals and trends by exponentially weighted moving averages, *ONRResearch Memorandum 52*, Carnegie Institute of Technology, Pittsburgh, Pennsylvania.

[19] Hurvich, C.M. and Tsai, C.L. (1989), Regression and time series model selection in small samples, *Biometrika*, 76, 297307.

[20] Jones, R.H. (1975), Fitting autoregressions, *J. Amer. Stat. Assoc.*, 70, 590592.

[21] Lehmann, E.L. (1983), *Theory of Point Estimation*, John Wiley, New York.

[22] Liu, J. and Brockwell, P.J. (1988), The general bilinear time series model, *J. Appl. Probability 25*, 553564.

[23] Lütkepohl, H. (1993), *Introduction to Multiple Time Series Analysis*, 2nd Edition, Springer-Verlag, Berlin.

[24] McLeod, A.I. and Li, W.K. (1983), Diagnostic checking ARMA time series models using squared-residual autocorrelations, *J. Time Series Anal., 4*, 269273.

[25] Mood, A.M., Graybill, F.A., and Boes, D.C. (1974), *Introduction to the Theory of Statistics*, McGraw-Hill, New York.

[26] Newton, H.J. and Parzen, E. (1984), Forecasting and time series model types of 111 economic time series, *The Forecasting Accuracy of Major Time Series Methods*, S. Makridakis et al. (eds.), John Wiley and Sons, Chichester.

[27] Parzen, E. (1982), ARARMA models for time series analysis and forecasting, *J. Forecasting*, 1, 6782.

[28] Priestley, M.B. (1988), *Non-linear and Non-stationary Time Series Analysis*, Academic Press, London.

[29] Rosenblatt, M. (1985), *Stationary Sequences and Random Fields*, Birkhäuser, Boston. Said, S.E. and Dickey, D.A. (1984), Testing for unit roots in autoregressive moving average models with unknown order, *Biometrika*, 71, 599607.

[30] Sargent, Thomas (1979), *Macroeconomic Theory.*

[31] Shibata, R. (1976), Selection of the order of an autoregressive model by Akaikes information criterion, *Biometrika*, 63, 117126.

[32] Shibata, R. (1980), Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *Ann. Stat.*, 8, 147164.

[33] Shapiro, S.S. and Francia, R.S. (1972), An approximate analysis of variance test for normality, *J. Amer. Stat. Assoc., 67*, 215216.

[34] Subba-Rao, T. and Gabr, M.M. (1984), *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, Springer Lecture Notes in Statistics, 24.

[35] Tong, H. (1990), *Non-linear Time Series: A Dynamical Systems Approach*, Oxford University Press, Oxford.

[36] Whittle, P. (1963), On the fitting of multivariate autoregressions and the approximate canonical factorization of a spectral density matrix, *Biometrika*, 40, 129134.