STAT 626: Outline of Lecture 14
ARMA Models (Chapter 4)

1. **Autoregressive Models** (§4.1)

2. **Estimation** (§4.3)

   Yule-Walker Equations.

3. **Correlation Functions (§4.2)**

   ACF and PACF

4. **Forecasting/Prediction (§4.4)**

Review of Stationarity, Overview of TS Models (Chapter 4)

1. **Autoregressive Models of order p or AR(p) Models:**

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + w_t, \quad \phi_p \neq 0.$$

QUESTION: Is a time series $\{x_t\}$ defined via an AR(p) model always stationary? If so, what is its autocovariance function?
To get a feel for the answer consider the AR(1):

$$x_t = \phi x_{t-1} + w_t,$$

what happens when $\phi = 1$?

See Examples 1.9, 1.10, 2.20 and Problem 2.4 for more details on AR models.

2. **Linear Processes:** $x_t = \mu + \sum_{j=-\infty}^{+\infty} \psi_j w_{t-j}$ is stationary with the *autocovariance function*

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

3. **MA(q) Models**: $x_t = w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}, \quad \theta_q \neq 0$, is stationary, its autocovariance is zero at lags greater than q.

4. **The Backshift Operator B:** $Bx_t = x_{t-1}$.

5. **MA(q) and B:**

$$x_t = w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q} = (1 + \theta_1 B + \ldots + \theta_q B^q) w_t = \theta(B) w_t.$$

6. **AR(p) and B:**

$$x_t - \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p} = w_t, \quad (1 - \phi_1 B - \ldots - \phi_p B^p) x_t = \phi(B) x_t = w_t.$$

7. **The ROOTS of the polynomial equation**

$$\phi(B) = 0,$$

hold the key to the question of stationarity of the solutions of AR models.

Chapter 4

ARMA Models

4.1 Autoregressive Moving Average Models

Linear regression models are often unsatisfactory for explaining all of the interesting dynamics of a time series. Instead, the introduction of correlation through lagged relationships leads to autoregressive (AR) and moving average (MA) models. These models are often combined to form autoregressive moving average (ARMA) models.

Autoregressive models are an obvious extension of linear regression models. An *autoregressive model* of order p, abbreviated AR(p), is of the form

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t, \tag{4.1}$$

where x_t is stationary and w_t is white noise. We note that (4.1) is similar to the regression model of Section 3.1, and hence the term auto (or self) regression. Some technical difficulties develop from applying that model because the regressors, $x_{t-1}, \ldots, x_{t-p}$, are random components, whereas in regression, the regressors are assumed to be fixed. For example, we will see that restrictions must be put on the AR parameters, as opposed to linear regression where there are no parameter restrictions.

Example 4.1. The AR(1) Model and Causality
Consider the first-order, zero-mean AR(1) model,

$$x_t = \phi x_{t-1} + w_t.$$

Because x_t must be stationary, we can rule out the case $\phi = 1$ because this would make x_t a random walk, which we know is not stationary. Similarly, we can rule out $\phi = -1$. In other words, the models

$$x_t = x_{t-1} + w_t\,, \quad \text{and} \quad x_t = -x_{t-1} + w_t\,,$$

are *not* AR models because they are not stationary.

As we saw in Example 2.20, if x_t is stationary, then

$$\text{var}(x_t) = \phi^2 \text{var}(x_{t-1}) + \text{var}(w_t)\,,$$

which, because $\text{var}(x_{t-1}) = \text{var}(x_t)$, implies

$$\text{var}(x_t) = \gamma(0) = \sigma_w^2 \frac{1}{(1-\phi^2)}\,.$$

Thus, we must have $|\phi| < 1$ for the process to have a positive (finite) variance. Similarly, in Example 2.20, we showed that ϕ is the correlation between x_t and x_{t-1}.

Provided that $|\phi| < 1$ we can represent an AR(1) model as a linear process given by

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}. \tag{4.2}$$

Representation (4.2) is called the *causal solution* of the model (see Section D.2 for details). The term causal refers to the fact that x_t does not depend on the future. In fact, by simple substitution,

$$\underbrace{\sum_{j=0}^{\infty} \phi^j w_{t-j}}_{x_t} = \phi \underbrace{\left(\sum_{k=0}^{\infty} \phi^k w_{t-1-k}\right)}_{x_{t-1}} + w_t \,.$$

As a check, the right-hand side is $w_t + \phi w_{t-1}\,[k=0] + \phi^2 w_{t-2}\,[k=1] + \cdots$. Using (4.2), it is easy to see that the AR(1) process is stationary with mean

$$E(x_t) = \sum_{j=0}^{\infty} \phi^j E(w_{t-j}) = 0,$$

and autocovariance function ($h \geq 0$),

$$\begin{aligned}
\gamma(h) &= \operatorname{cov}(x_{t+h}, x_t) = \operatorname{cov}\left(\sum_{j=0}^{\infty} \phi^j w_{t+h-j}, \sum_{k=0}^{\infty} \phi^k w_{t-k}\right) \\
&= \operatorname{cov}\left[w_{t+h} + \cdots + \phi^h w_t + \phi^{h+1} w_{t-1} + \cdots,\ \phi^0 w_t + \phi w_{t-1} + \cdots\right] \\
&= \sigma_w^2 \sum_{j=0}^{\infty} \phi^{h+j}\phi^j = \sigma_w^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma_w^2 \phi^h}{1-\phi^2}\,.
\end{aligned} \tag{4.3}$$

Recall that $\gamma(h) = \gamma(-h)$, so we will only exhibit the autocovariance function for $h \geq 0$. From (4.3), the ACF of an AR(1) is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h, \quad h \geq 0. \tag{4.4}$$

In addition, from the causal form (4.2) we see that, as required in Example 2.20, x_{t-1} and w_t are uncorrelated because $x_{t-1} = \sum_{j=0}^{\infty} \phi^j w_{t-1-j}$ is a linear filter of past shocks, $w_{t-1}, w_{t-2}, \ldots$, which are uncorrelated with w_t, the present shock. Also, the causal form of the model allows us to easily see that if we replace x_t by $x_t - \mu$, then

$$x_t = \mu + \sum_{j=0}^{\infty} \phi^j w_{t-j}\,,$$

so that the mean function is now $E(x_t) = \mu$. ◊

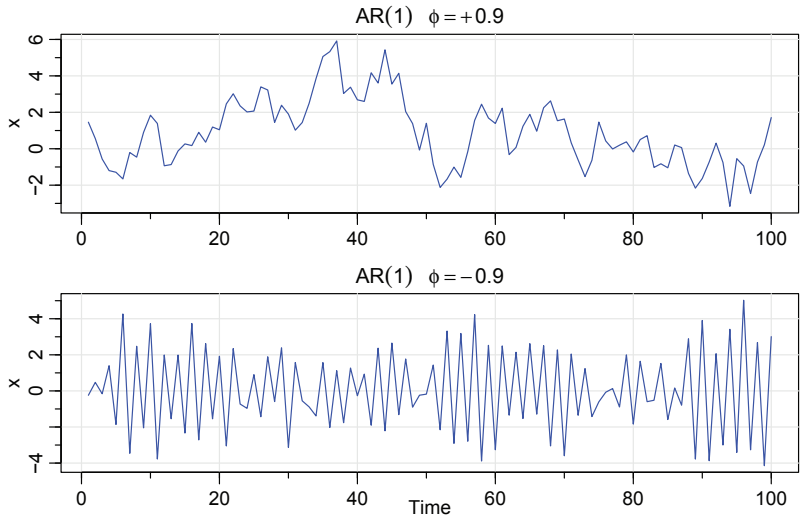

Figure 4.1 *Simulated AR(1) models: $\phi = .9$ (top); $\phi = -.9$ (bottom).*

Example 4.2. The Sample Path of an AR(1) Process

Figure 4.1 shows a time plot of two AR(1) processes, one with $\phi = .9$ and one with $\phi = -.9$; in both cases, $\sigma_w^2 = 1$. In the first case, $\rho(h) = .9^h$, for $h \geq 0$, so observations close together in time are positively correlated. Thus, observations at contiguous time points will tend to be close in value to each other; this fact shows up in the top of Figure 4.1 as a very smooth sample path for x_t. Now, contrast this with the case in which $\phi = -.9$, so that $\rho(h) = (-.9)^h$, for $h \geq 0$. This result means that observations at contiguous time points are negatively correlated but observations two time points apart are positively correlated. This fact shows up in the bottom of Figure 4.1, where, for example, if an observation, x_t, is positive, the next observation, x_{t+1}, is typically negative, and the next observation, x_{t+2}, is typically positive. Thus, in this case, the sample path is very choppy. The following R code can be used to obtain a figure similar to Figure 4.1:

```
par(mfrow=c(2,1))
tsplot(arima.sim(list(order=c(1,0,0), ar=.9), n=100), ylab="x", col=4,
           main=expression(AR(1)~~~phi==+.9))
tsplot(arima.sim(list(order=c(1,0,0),ar=-.9), n=100), ylab="x", col=4,
           main=expression(AR(1)~~~phi==-.9))
```

◊

Example 4.3. AR(p) and Causality

In Example 4.1, we saw that an AR(1) has as a causal representation; for example, the AR(1) model $x_t = .9x_{t-1} + w_t$ can also be written as $x_t = \sum_{j=0}^{\infty} .9^j w_{t-j}$. In the general case, it is more difficult to go from one version to another. It is, however, possible to use the R command `ARMAtoMA` to print some of the coefficients.

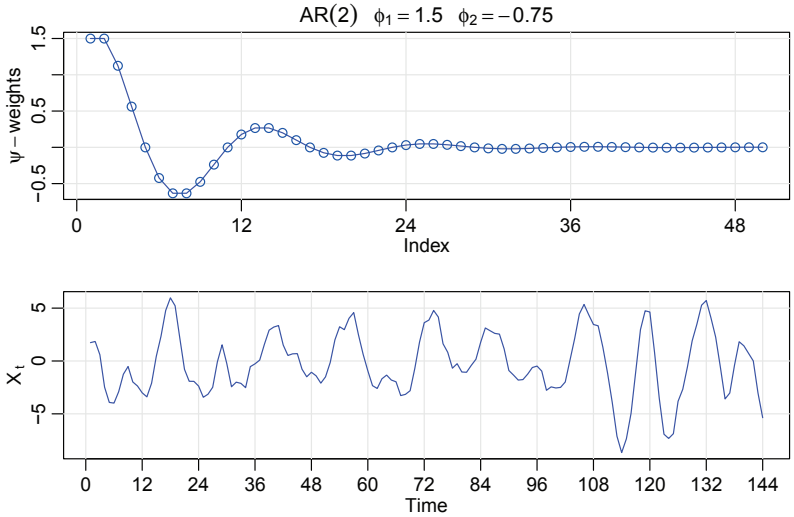

Figure 4.2 *ψ-weights and simulated data of an AR(2),* $x_t = 1.5x_{t-1} - .75x_{t-2} + w_t$.

For example, the AR(2) model

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t\,,$$

can be written in its *causal* form, $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}$, where $\psi_0 = 1$ and

$$\psi_j = 2(\tfrac{\sqrt{3}}{2})^j \cos(\tfrac{2\pi(j-2)}{12})\,, \qquad j = 1, 2, \ldots\,.$$

The ψ-weights were solved for using difference equation theory (see Shumway and Stoffer, 2017, §3.2). Notice that the coefficients are cyclic with a period of 12 (like monthly data), but they decrease exponentially fast to zero (because $\sqrt{3}/2 < 1$) indicating a short dependence on the past. Figure 4.2 shows a plot of the ψ_j for $j = 1, \ldots, 50$, as well as simulated data from the model. Both show the cyclic-type behavior of this particular model. It is evident that the linear process form of the model gives more insight into the model than the regression form of the model. Finally, we note that an AR(p) is also an MA(∞).

The following R code was used for this example.

```
psi = ARMAtoMA(ar = c(1.5, -.75), ma = 0, 50)
par(mfrow=c(2,1), mar=c(2,2.5,1,0)+.5, mgp=c(1.5,.6,0), cex.main=1.1)
plot(psi, xaxp=c(0,144,12), type="n", col=4,
           ylab=expression(psi-weights),
           main=expression(AR(2)~~~phi[1]==1.5~~~phi[2]==-.75))
abline(v=seq(0,48,by=12), h=seq(-.5,1.5,.5), col=gray(.9))
lines(psi, type="o", col=4)
set.seed(8675309)
simulation = arima.sim(list(order=c(2,0,0), ar=c(1.5,-.75)), n=144)
```

```
plot(simulation, xaxp=c(0,144,12), type="n", ylab=expression(X[~t]))
abline(v=seq(0,144,by=12), h=c(-5,0,5), col=gray(.9))
lines(simulation, col=4)
```

◊

We now formally define the concept of causality. The importance of this condition is to make sure that a time series model is not future-dependent. This allows us to be able to predict future values of a time series based on only the present and the past.

Definition 4.4. *A time series x_t is said to be* **causal** *if it can be written as*

$$x_t = \mu + \sum_{j=0}^{\infty} \psi_j w_{t-j}$$

for constants ψ_j satisfying $\sum_{j=0}^{\infty} \psi_j^2 < \infty$.

Remark. As stated in Property 2.21, any stationary (non-deterministic) time series has a causal representation.

As an alternative to autoregression, think of w_t as a "shock" to the process at time t. One can imagine that what happens today might be related to shocks from a few previous days. In this case, we have the moving average model of order q, abbreviated as MA(q). The *moving average model* of order q, is defined by[1]

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}, \tag{4.5}$$

where w_t is white noise. Unlike the autoregressive process, the moving average process is stationary for any values of the parameters $\theta_1, \ldots, \theta_q$. In addition, the MA($q$) is already in the form of Definition 4.4 with $\psi_j = \theta_j$ and $\theta_j = 0$ for $j > q$.

Example 4.5. The MA(1) Process

Consider the MA(1) model $x_t = w_t + \theta w_{t-1}$. Then, $E(x_t) = 0$, and if we replace x_t by $x_t - \mu$, then $E(x_t) = \mu$. The autocovariance function is

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma_w^2 & h = 0, \\ \theta\sigma_w^2 & |h| = 1, \\ 0 & |h| > 1, \end{cases}$$

and the ACF is

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

Note $|\rho(1)| \leq 1/2$ for all values of θ (Problem 4.1). Also, x_t is correlated with x_{t-1}, but not with $x_{t-2}, x_{t-3}, \ldots$. Contrast this with the case of the AR(1) model in which the correlation between x_t and x_{t-k} is never zero. When $\theta = .9$, for example,

[1] Some texts and software packages write the MA model with negative coefficients; that is, $x_t = w_t - \theta_1 w_{t-1} - \theta_2 w_{t-2} - \cdots - \theta_q w_{t-q}$.

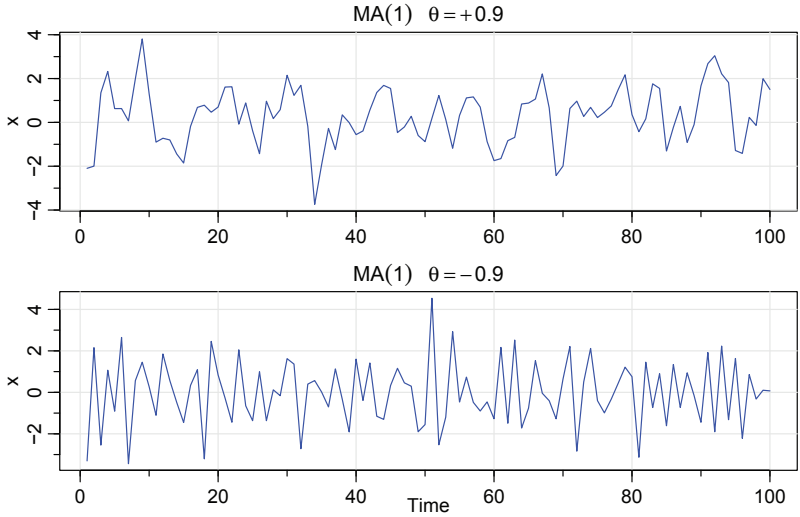

Figure 4.3 *Simulated MA(1) models: $\theta = .9$ (top); $\theta = -.9$ (bottom).*

x_t and x_{t-1} are positively correlated, and $\rho(1) = .497$. When $\theta = -.9$, x_t and x_{t-1} are negatively correlated, $\rho(1) = -.497$. Figure 4.3 shows a time plot of these two processes with $\sigma_w^2 = 1$. The series for which $\theta = .9$ is smoother than the series for which $\theta = -.9$. A figure similar to Figure 4.3 can be created in R as follows:

```
par(mfrow = c(2,1))
tsplot(arima.sim(list(order=c(0,0,1), ma=.9),  n=100), col=4,
           ylab="x", main=expression(MA(1)~~~theta==+.5))
tsplot(arima.sim(list(order=c(0,0,1), ma=-.9), n=100), col=4,
           ylab="x", main=expression(MA(1)~~~theta==-.5))
```

◊

Example 4.6. Non-uniqueness of MA Models and Invertibility
Using Example 4.5, we note that for an MA(1) model, the pair $\sigma_w^2 = 1$ and $\theta = 5$ yield the same autocovariance function as the pair $\sigma_w^2 = 25$ and $\theta = 1/5$, namely,

$$\gamma(h) = \begin{cases} 26 & h = 0, \\ 5 & |h| = 1, \\ 0 & |h| > 1. \end{cases}$$

Thus, the MA(1) processes

$$x_t = w_t + \tfrac{1}{5} w_{t-1}, \quad w_t \sim \text{iid } \mathrm{N}(0, 25)$$

and

$$y_t = v_t + 5 v_{t-1}, \quad v_t \sim \text{iid } \mathrm{N}(0, 1)$$

are stochastically the same. We can only observe the time series, x_t or y_t, and not the noise, w_t or v_t, so we cannot distinguish between the models. Hence, we will have to

choose only one of them. For convenience, by mimicking causality for AR models, we will choose the model with an infinite AR representation. Such a process is called an *invertible* process.

To discover which model is the invertible model, we can reverse the roles of x_t and w_t (because we are mimicking the AR case) and write the MA(1) model as

$$w_t = -\theta w_{t-1} + x_t\,.$$

As in (4.2), if $|\theta| < 1$, then $w_t = \sum_{j=0}^{\infty}(-\theta)^j x_{t-j}$, which is the desired infinite representation of the model. Hence, given a choice, we will choose the model with $\sigma_w^2 = 25$ and $\theta = 1/5$ because it is invertible. ◊

Henceforth, for uniqueness, we require that a moving average have an *invertible* representation:

Definition 4.7. *A time series x_t is said to be* **invertible** *if it can be written as*

$$w_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}\,.$$

for constants π_j satisfying $\sum_{j=0}^{\infty}\pi_j^2 < \infty$.

Remark. Aside from the uniqueness problem, invertibility is important because it gives a representation of a present shock, w_t, in terms of the present and past data. Consequently, the current shock to the system does not depend on future data. Also, note that an MA(q) is an AR(∞).

We now proceed with the general development of mixed *autoregressive moving average* (ARMA) models for stationary time series.

Definition 4.8. *A time series $\{x_t;\ t = 0, \pm1, \pm2, \ldots\}$ is* **ARMA(p, q)** *if*

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \tag{4.6}$$

with $\phi_p \neq 0$, $\theta_q \neq 0$, $\sigma_w^2 > 0$, and the model is causal and invertible. Henceforth, unless stated otherwise, w_t is a Gaussian white noise series with mean zero and variance σ_w^2. If $E(x_t) = \mu$, then $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$.

The ARMA model may be seen as a regression of the present outcome (x_t) on the past outcomes ($x_{t-1}, \ldots, x_{t-p}$), with correlated errors. That is,

$$x_t = \beta_0 + \beta_1 x_{t-1} + \cdots + \beta_p x_{t-p} + \epsilon_t,$$

where $\epsilon_t = w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$, although we call the regression parameters ϕ instead of β. As opposed to ordinary regression, the ϕ parameters are restricted to certain values in order to obtain causality and the θ parameters are restricted to certain values to obtain invertibility.

When $q = 0$, the model is called an autoregressive model of order p, AR(p), and when $p = 0$, the model is called a moving average model of order q, MA(q). Before

proceeding, we establish some notation based on the backshift operator defined in Definition 3.8, $B^k x_t = x_{t-k}$. Using the backshift operator, we can write the AR(p) model as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) x_t = w_t \,.$$

Thus, it is convenient to define the **autoregressive operator** as

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p. \tag{4.7}$$

so that the AR model is $\phi(B)x_t = w_t$. As in the AR(p) case, the MA(q) model may be written as

$$x_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) w_t,$$

so we define the **moving average operator** as

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{4.8}$$

and write an MA(q) model as $x_t = \theta(B)w_t$. Consequently, an ARMA(p, q) model can be written as concisely as

$$\phi(B)(x_t - \mu) = \theta(B) w_t \,, \tag{4.9}$$

where the orders of $\phi(B)$ and $\theta(B)$ are understood to be p and q, respectively.

In addition to restricted values of the ϕs and θs, there are complications where the autoregressive side of the model can cancel the moving average side of the model. This is called over-parameterization or parameter redundancy. That is, given an ARMA(p, q) model, we can unnecessarily complicate the model by multiplying both sides by another operator, say

$$\eta(B)\phi(B)(x_t - \mu) = \eta(B)\theta(B) w_t \,,$$

without changing the dynamics. Consider the following example.

Example 4.9. Parameter Redundancy
Consider a white noise process $x_t = w_t$. Now multiply both sides of the equation by $(1 - .9B)$ to get

$$x_t - .9x_{t-1} = w_t - .9w_{t-1} \,,$$

or

$$x_t = .9x_{t-1} - .9w_{t-1} + w_t, \tag{4.10}$$

which looks like an ARMA$(1,1)$ model. Of course, x_t is still white noise; nothing has changed in this regard [i.e., $x_t = w_t$ is the solution to (4.10)], but we have hidden the fact that x_t is white noise because of the *parameter redundancy* or over-parameterization. ♢

Example 4.9 points out the need to be careful when fitting ARMA models to data. Unfortunately, *it is easy to fit an overly complex ARMA model to data*. For example, if a process is truly white noise, it is possible to fit a significant ARMA(k, k) model to the data. Consider the following example.

Example 4.10. Parameter Redundancy and Estimation
Although we have not discussed estimation yet, we present the following demonstration of the problem. We generated 150 iid normals with $\mu = 5$ and $\sigma = 1$, and then fit an ARMA$(1,1)$ to the data. Note that $\hat{\phi} = -.96$ and $\hat{\theta} = .95$, and both are significant. Below is the R code (note that the estimate called "intercept" is really the estimate of the mean).

```
set.seed(8675309)          # Jenny, I got your number
x = rnorm(150, mean=5)     # generate iid N(5,1)s
arima(x, order=c(1,0,1))   # estimation
  Coefficients:
          ar1      ma1  intercept <= misnomer
       -0.96    0.95       5.05
  s.e.  0.17    0.17       0.07
```

Of course the data are independent, but the estimation implies a seemingly different result that the data are highly dependent. ♢

Henceforth, we will require an ARMA model to be reduced to its simplest form. A simple way to discover if this problem exists with a model is to write the model with the AR part on the left and the MA part on the right, and then compare each side.

Example 4.11. Checking for Parameter Redundancy
In the previous example, it was easy to see that the left-hand and right-hand sides are nearly the same. For more complicated models, we can use R to compare each side. For example, consider the model

$$x_t = .3x_{t-1} + .4x_{t-2} + w_t + .5w_{t-1}\,,$$

which looks like an ARMA$(2,1)$. Now write the model as

$$(1 - .3B - .4B^2)x_t = (1 + .5B)w_t\,,$$

or

$$(1 + .5B)(1 - .8B)x_t = (1 + .5B)w_t\,.$$

We can cancel the $(1 + .5B)$ on each side, so the model is really an AR(1),

$$x_t = .8x_{t-1} + w_t\,.$$

These situations can be checked easily in R by looking at the roots of the polynomials in B corresponding to each side. If the roots are close, then there may be parameter redundancy:

```
AR = c(1, -.3, -.4)  # original AR coefs on the left
polyroot(AR)
 [1]  1.25-0i -2.00+0i
MA = c(1, .5)        # original MA coefs on the right
polyroot(MA)
 [1] -2+0i
```

This indicates there is one common factor (with root -2) and hence the model is over-parameterized and can be reduced. ◊

Example 4.12. Causal and Invertible ARMA
It might be useful at times to write an ARMA model in its causal or invertible forms. For example, consider the model

$$x_t = .8x_{t-1} + w_t - .5w_{t-1}\,.$$

Using R, we can list some of the causal and invertible coefficients of our ARMA(1, 1) model as follows:

```
round( ARMAtoMA(ar=.8, ma=-.5, 10), 2)   # first 10 ψ-weights
  [1] 0.30 0.24 0.19 0.15 0.12 0.10 0.08 0.06 0.05 0.04
round( ARMAtoAR(ar=.8, ma=-.5, 10), 2)  # first 10 π-weights
  [1] -0.30 -0.15 -0.08 -0.04 -0.02 -0.01  0.00  0.00  0.00  0.00
```

Thus, the causal form looks like,

$$x_t = w_t + .3w_{t-1} + .24w_{t-2} + .19w_{t-3} + \cdots + .05w_{t-9} + .04w_{t-10} + \cdots\,,$$

whereas the invertible form looks like,

$$w_t = x_t - .3x_{t-1} - .15x_{t-2} - .08x_{t-3} - .04x_{t-4} - .02x_{t-5} - .01x_{t-6} + \cdots\,.$$

If a model is not causal or invertible, the scripts will work, but the coefficients will not converge to zero. For a random walk, $x_t = x_{t-1} + w_t$, or $x_t = \sum_{j=1}^{t} w_j$, for example:

```
ARMAtoMA(ar=1, ma=0, 20)
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

◊

4.2 Correlation Functions

*Autocorrelation Function (*ACF)

Example 4.13. ACF of an MA(q)
Write the model as $x_t = \sum_{j=0}^{q} \theta_j w_{t-j}$ with $\theta_0 = 1$ for ease. Because x_t is a finite linear combination of white noise terms, the process is stationary with autocovariance function

$$\gamma(h) = \operatorname{cov}(x_{t+h}, x_t) = \operatorname{cov}\Big(\sum_{j=0}^{q} \theta_j w_{t+h-j},\ \sum_{k=0}^{q} \theta_k w_{t-k}\Big)$$

$$= \begin{cases} \sigma_w^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h}, & 0 \le h \le q \\ 0 & h > q\,, \end{cases} \tag{4.11}$$

which is similar to the calculation in (2.16). The cutting off of $\gamma(h)$ after q lags is the signature of the MA(q) model. Dividing (4.11) by $\gamma(0)$ yields the ACF of an MA(q):

$$\rho(h) = \begin{cases} \dfrac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{1+\theta_1^2+\cdots+\theta_q^2} & 1 \le h \le q \\ 0 & h > q. \end{cases} \tag{4.12}$$

In addition, we note that $\rho(q) \neq 0$ because $\theta_q \neq 0$. ♢

Example 4.14. ACF of an AR(p) and ARMA(p,q)
For an AR(p) or ARMA(p,q) model, write the model in its causal MA(∞) form,

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}. \tag{4.13}$$

It follows immediately that the autocovariance function of x_t can be written as

$$\gamma(h) = \operatorname{cov}(x_{t+h}, x_t) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_{j+h}\psi_j, \quad h \ge 0, \tag{4.14}$$

as was calculated in (2.16). The ACF is given by

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\sum_{j=0}^{\infty} \psi_{j+h}\psi_j}{\sum_{j=0}^{\infty} \psi_j^2}, \quad h \ge 0. \tag{4.15}$$

Unlike the MA(q), the ACF of an AR(p) or an ARMA(p,q) does not cut off at any lag, so using the ACF to help identify the order of an AR or ARMA is difficult. ♢

Result (4.15) is not appealing in that it provides little information about the appearance of the ACF of various models. We can, however, look at what happens for some specific models.

Example 4.15. ACF of an AR(2)
Figure 4.2 shows $n = 144$ observations from the AR(2) model

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t,$$

with $\sigma_w^2 = 1$. We examined this model in Example 4.3 where we noted that the process exhibits pseudo-cyclic behavior at the rate of one cycle every 12 time points. Because the ψ-weights are cyclic, the ACF of the model will also be cyclic with a period of 12. The R code to calculate and display the ACF for this model as shown on the left side of Figure 4.4 is:

```
ACF = ARMAacf(ar=c(1.5,-.75), ma=0, 50)
plot(ACF, type="h", xlab="lag", panel.first=Grid())
abline(h=0)
```

♢

The general behavior of the ACF of an AR(p) or an ARMA(p,q) is controlled by the AR part because the MA part has only finite influence.

Example 4.16. The ACF of an ARMA(1,1)
Consider the ARMA$(1,1)$ process $x_t = \phi x_{t-1} + \theta w_{t-1} + w_t$. Using the theory of difference equations, we can show that the ACF is given by

$$\rho(h) = \frac{(1+\theta\phi)(\phi+\theta)}{\phi(1+2\theta\phi+\theta^2)}\,\phi^h, \quad h \geq 1. \tag{4.16}$$

Notice that the general pattern of $\rho(h)$ in (4.16) is not different from that of an AR(1) given in (4.4), $\rho(h) = \phi^h$. Hence, it is unlikely that we will be able to tell the difference between an ARMA(1,1) and an AR(1) based solely on an ACF estimated from a sample. This consideration will lead us to the partial autocorrelation function. ◊

*Partial Autocorrelation Function (*PACF)

In (4.12), we saw that for MA(q) models, the ACF will be zero for lags greater than q. Moreover, because $\theta_q \neq 0$, the ACF will not be zero at lag q. Thus, the ACF provides a considerable amount of information about the order of the dependence when the process is a moving average process.

If the process, however, is ARMA or AR, the ACF alone tells us little about the orders of dependence. Hence, it is worthwhile pursuing a function that will behave like the ACF of MA models, but for AR models, namely, the *partial autocorrelation function (PACF)*.

Recall that if X, Y, and Z are random variables, then the partial correlation between X and Y given Z is obtained by regressing X on Z to obtain the predictor $\hat{X}$, regressing Y on Z to obtain $\hat{Y}$, and then calculating

$$\rho_{XY|Z} = \text{corr}\{X - \hat{X},\, Y - \hat{Y}\}.$$

The idea is that $\rho_{XY|Z}$ measures the correlation between X and Y with the linear effect of Z removed (or partialled out). If the variables are multivariate normal, then this definition coincides with $\rho_{XY|Z} = \text{corr}(X, Y \mid Z)$.

To motivate the idea of partial *auto*correlation, consider a causal AR(1) model, $x_t = \phi x_{t-1} + w_t$. Then,

$$\begin{aligned}\gamma_x(2) &= \text{cov}(x_t, x_{t-2}) = \text{cov}(\phi x_{t-1} + w_t, x_{t-2}) \\ &= \text{cov}(\phi x_{t-1}, x_{t-2}) = \phi\gamma_x(1).\end{aligned}$$

Note that $\text{cov}(w_t, x_{t-2}) = 0$ from causality because x_{t-2} involves $\{w_{t-2}, w_{t-3}, \ldots\}$, which are all uncorrelated with w_t. The correlation between x_t and x_{t-2} is not zero as it would be for an MA(1) because x_t is dependent on x_{t-2} through x_{t-1}. Suppose we break this chain of dependence by removing (or partialling out) the effect of x_{t-1}. That is, we consider the correlation between $x_t - \phi x_{t-1}$ and $x_{t-2} - \phi x_{t-1}$, because it is the correlation between x_t and x_{t-2} with the linear dependence of each on x_{t-1} removed. In this way, we have broken the dependence chain between x_t and x_{t-2},

$$\text{cov}(x_t - \phi x_{t-1}, x_{t-2} - \phi x_{t-1}) = \text{cov}(w_t, x_{t-2} - \phi x_{t-1}) = 0.$$

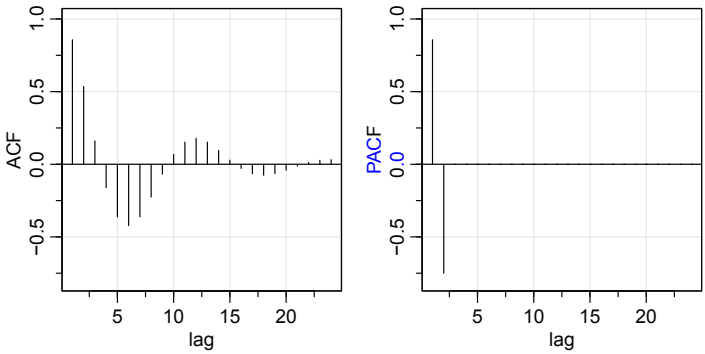

Figure 4.4 *The ACF and PACF of an AR(2) model with* $\phi_1 = 1.5$ *and* $\phi_2 = -.75$.

Hence, the tool we need is partial autocorrelation, which is the correlation between x_s and x_t with the linear effect of everything "in the middle" removed.

Definition 4.17. *The* **partial autocorrelation function (PACF)** *of a stationary process,* x_t*, denoted* ϕ_{hh}*, for* $h = 1, 2, \ldots,$ *is*

$$\phi_{11} = \text{corr}(x_1, x_0) = \rho(1) \tag{4.17}$$

and

$$\phi_{hh} = \text{corr}(x_h - \hat{x}_h,\ x_0 - \hat{x}_0), \quad h \geq 2, \tag{4.18}$$

where $\hat{x}_h$ *is the regression of* x_h *on* $\{x_1, x_2, \ldots, x_{h-1}\}$ *and* $\hat{x}_0$ *is the regression of* x_0 *on* $\{x_1, x_2, \ldots, x_{h-1}\}$.

Thus, due to the stationarity, the PACF, ϕ_{hh}, is the correlation between x_{t+h} and x_t with the linear dependence of everything between them, namely $\{x_{t+1}, \ldots, x_{t+h-1}\}$, on each, removed.

It is not necessary to actually run regressions to compute the PACF because the values can be computed recursively based on what is known as the Durbin–Levinson algorithm due to Levinson (1947) and Durbin (1960).

Example 4.18. The PACF of an AR(p)

The PACF of an AR(p) model will be zero for all lags larger than p and the PACF at lag p will not be zero because it can be shown that $\phi_{pp} = \phi_p$ (the last parameter in the model).

In Example 4.15 we looked at the AR(2) model

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t\,.$$

In this case, $\phi_{11} = \rho(1) = \phi_1/(1-\phi_2) = 1.5/1.75 \approx .86$, $\phi_{22} = \phi_2 = -.75$, and $\phi_{hh} = 0$ for $h > 2$. Figure 4.4 shows the ACF and the PACF of this AR(2) model. To reproduce Figure 4.4 in R, use the following commands:

Table 4.1 *Behavior of the ACF and PACF for ARMA Models*

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

```
ACF = ARMAacf(ar=c(1.5,-.75), ma=0, 24)[-1]
PACF = ARMAacf(ar=c(1.5,-.75), ma=0, 24, pacf=TRUE)
par(mfrow=1:2)
tsplot(ACF, type="h", xlab="lag", ylim=c(-.8,1))
abline(h=0)
tsplot(PACF, type="h", xlab="lag", ylim=c(-.8,1))
abline(h=0)
```

◊

We also have the following large sample result for the PACF, which may be compared to the similar result for the ACF given in Property 2.28.

Property 4.19 (Large Sample Distribution of the PACF). *If a time series is an AR(p) and the sample size n is large, then for $h > p$, the $\hat{\phi}_{hh}$ are approximately independent normal with mean 0 and standard deviation $1/\sqrt{n}$. This result also holds for $p = 0$, wherein the process is white noise.*

Example 4.20. The PACF of an MA(q)
An MA(q) is invertible, so it has an AR(∞) representation,

$$x_t = -\sum_{j=1}^{\infty} \pi_j x_{t-j} + w_t\,.$$

Moreover, no finite representation exists. From this result, it should be apparent that the PACF will never cut off, as in the case of an AR(p). For an MA(1), $x_t = w_t + \theta w_{t-1}$, with $|\theta| < 1$, it can be shown that

$$\phi_{hh} = -\frac{(-\theta)^h(1-\theta^2)}{1-\theta^{2(h+1)}}, \quad h \geq 1.$$

◊

The PACF for MA models behaves much like the ACF for AR models. Also, the PACF for AR models behaves much like the ACF for MA models. Because an invertible ARMA model has an infinite AR representation, the PACF will not cut off. We may summarize these results in Table 4.1.

Example 4.21. Preliminary Analysis of the Recruitment Series
We consider the problem of modeling the Recruitment series shown in Figure 1.5. There are 453 months of observed recruitment ranging over the years 1950–1987. The ACF and the PACF given in Figure 4.5 are consistent with the behavior of

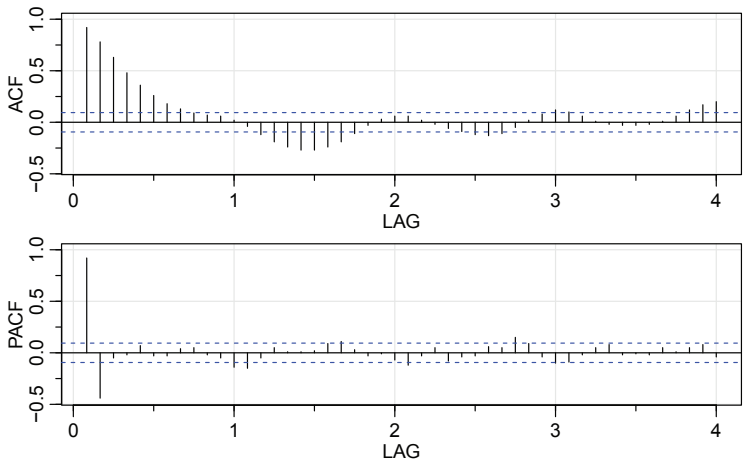

Figure 4.5 *ACF and PACF of the Recruitment series. Note that the lag axes are in terms of season (12 months in this case).*

an AR(2). The ACF has cycles corresponding roughly to a 12-month period, and the PACF has large values for $h = 1, 2$ and then is essentially zero for higher-order lags. Based on Table 4.1, these results suggest that a second-order ($p = 2$) autoregressive model might provide a good fit. Although we will discuss estimation in detail in Section 4.3, we ran a regression (OLS) using the data triplets $\{(x; z_1, z_2) : (x_3; x_2, x_1), (x_4; x_3, x_2), \ldots, (x_{453}; x_{452}, x_{451})\}$ to fit the model

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

for $t = 3, 4, \ldots, 453$. The values of the estimates were $\hat{\phi}_0 = 6.74_{(1.11)}$, $\hat{\phi}_1 = 1.35_{(.04)}$, $\hat{\phi}_2 = -.46_{(.04)}$, and $\hat{\sigma}_w^2 = 89.72$, where the estimated standard errors are in parentheses.

The following R code can be used for this analysis. We use the script `acf2` from `astsa` to print and plot the ACF and PACF.

```
acf2(rec, 48)      # will produce values and a graphic
(regr = ar.ols(rec, order=2, demean=FALSE, intercept=TRUE))
   Coefficients:
         1        2
   1.3541  -0.4632
 Intercept: 6.737 (1.111)
 sigma^2 estimated as  89.72
regr$asy.se.coef  # standard errors of the estimates
 $ar
   [1] 0.04178901 0.04187942
```

We could have used `lm()` to do the regression, however using `ar.ols()` is much simpler for pure AR models. Also, the term `intercept` is used correctly here. ◊

4.3 Estimation

Throughout this section, we assume we have n observations, $x_1, \ldots, x_n$, from an ARMA(p, q) process in which, initially, the order parameters, p and q, are known. Our goal is to estimate the parameters, $\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$, and σ_w^2.

We begin with *method of moments* estimators. The idea behind these estimators is that of equating population moments, $E(x_t^k)$, to sample moments, $\frac{1}{n}\sum_{t=1}^n x_t^k$, for $k = 1, 2, \ldots$, and then solving for the parameters in terms of the sample moments. We immediately see that if $E(x_t) = \mu$, the method of moments estimator of μ is the sample average, $\bar{x}$ ($k = 1$). Thus, while discussing method of moments, we will assume $\mu = 0$. Although the method of moments can produce good estimators, they can sometimes lead to suboptimal estimators. We first consider the case in which the method leads to optimal (efficient) estimators, that is, AR(p) models,

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t\,.$$

If we multiply each side of the AR equation by x_{t-h} for $h = 0, 1, \ldots, p$ and take expectation, we obtain the following result.

Definition 4.22. *The* **Yule–Walker equations** *are given by*

$$\rho(h) = \phi_1\rho(h-1) + \cdots + \phi_p\rho(h-p), \quad h = 1, 2, \ldots, p, \tag{4.19}$$

$$\sigma_w^2 = \gamma(0)\,[1 - \phi_1\rho(1) - \cdots - \phi_p\rho(p)]. \tag{4.20}$$

The estimators obtained by replacing $\gamma(0)$ with its estimate, $\hat{\gamma}(0)$ and $\rho(h)$ with its estimate, $\hat{\rho}(h)$, are called the *Yule–Walker estimators*. For AR(p) models, if the sample size is large, the Yule–Walker estimators are approximately normally distributed, and $\hat{\sigma}_w^2$ is close to the true value of σ_w^2. In addition, the estimates are close to the OLS estimates discussed in Example 4.21.

Example 4.23. Yule–Walker Estimation for an AR(1)
For an AR(1), $(x_t - \mu) = \phi(x_{t-1} - \mu) + w_t$, the mean estimate is $\hat{\mu} = \bar{x}$, and (4.19) is

$$\rho(1) = \phi\rho(0) = \phi\,,$$

so

$$\hat{\phi} = \hat{\rho}(1) = \frac{\sum_{t=1}^{n-1}(x_{t+1} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}\,,$$

as expected. The estimate of the error variance is then

$$\hat{\sigma}_w^2 = \hat{\gamma}(0)\,[1 - \hat{\phi}^2]\,;$$

recall $\gamma(0) = \sigma_w^2/(1 - \phi^2)$ from (4.3). ♢

Example 4.24. Yule–Walker Estimation of the Recruitment Series
In Example 4.21 we fit an AR(2) model to the Recruitment series using regression. Below are the results of fitting the same model using Yule–Walker estimation, which are close to the regression values in Example 4.21.

```
rec.yw = ar.yw(rec, order=2)
rec.yw$x.mean     # mean estimate
 [1] 62.26278
rec.yw$ar         # phi parameter estimates
 [1]  1.3315874  -0.4445447
sqrt(diag(rec.yw$asy.var.coef))  # their standard errors
 [1] 0.04222637 0.04222637
rec.yw$var.pred  # error variance estimate
 [1] 94.79912
```

◊

In the case of AR(p) models, the Yule–Walker estimators are optimal estimators, but this is not true for MA(q) or ARMA(p,q) models. AR(p) models are basically linear models, and the Yule–Walker estimators are essentially least squares estimators. MA or ARMA models are nonlinear models, so this technique does not give optimal estimators.

Example 4.25. Method of Moments Estimation for an MA(1)
Consider the MA(1) model, $x_t = w_t + \theta w_{t-1}$, where $|\theta| < 1$. The model can then be written as

$$x_t = -\sum_{j=1}^{\infty}(-\theta)^j x_{t-j} + w_t,$$

which is nonlinear in θ. The first two population autocovariances are $\gamma(0) = \sigma_w^2(1+\theta^2)$ and $\gamma(1) = \sigma_w^2\theta$, so the estimate of θ is found by solving

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

Two solutions exist, so we would pick the invertible one. If $|\hat{\rho}(1)| \leq \frac{1}{2}$, the solutions are real, otherwise, a real solution does not exist. Even though $|\rho(1)| < \frac{1}{2}$ for an invertible MA(1), it may happen that $|\hat{\rho}(1)| \geq \frac{1}{2}$ because it is an estimator. For example, the following simulation in R produces a value of $\hat{\rho}(1) = .51$ when the true value is $\rho(1) = .9/(1+.9^2) = .497$.

```
set.seed(2)
ma1 = arima.sim(list(order = c(0,0,1), ma = 0.9), n = 50)
acf1(ma1, plot=FALSE)[1]
 [1] 0.51
```

◊

The preferred method of estimation is maximum likelihood estimation (MLE), which determines the values of the parameters that are most *likely* to have produced the observations. MLE for an AR(1) is discussed in detail in Section D.1. For normal models, this is the same as weighted least squares. For ease, we first discuss conditional least squares.

Conditional Least Squares

Recall from Chapter 3, in simple linear regression, $x_t = \beta_0 + \beta_1 z_t + w_t$, we minimize

$$S(\beta) = \sum_{t=1}^{n} w_t^2(\beta) = \sum_{t=1}^{n} (x_t - [\beta_0 + \beta_1 z_t])^2$$

with respect to the βs. This is a simple problem because we have all the data pairs, (z_t, x_t) for $t = 1, \ldots, n$. For ARMA models, we do not have this luxury.

Consider a simple AR(1) model, $x_t = \phi x_{t-1} + w_t$. In this case, the error sum of squares is

$$S(\phi) = \sum_{t=1}^{n} w_t^2(\phi) = \sum_{t=1}^{n} (x_t - \phi x_{t-1})^2 .$$

We have a problem because we didn't observe x_0. Let's make life easier by forgetting the problem and dropping the first term. That is, let's perform least squares using the (conditional) sum of squares,

$$S_c(\phi) = \sum_{t=2}^{n} w_t^2(\phi) = \sum_{t=2}^{n} (x_t - \phi x_{t-1})^2$$

because that's easy (it's just OLS) and if n is large, it shouldn't matter much. We know from regression that the solution is

$$\hat{\phi} = \frac{\sum_{t=2}^{n} x_t x_{t-1}}{\sum_{t=2}^{n} x_{t-1}^2} ,$$

which is nearly the Yule–Walker estimate in Example 4.23 (replace x_t by $x_t - \bar{x}$ if the mean is not zero).

Now we focus on conditional least squares for ARMA(p, q) models via *Gauss–Newton*. Write the model parameters as $\beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$, and for the ease of discussion, we will put $\mu = 0$. Write the ARMA model in terms of the errors

$$w_t(\beta) = x_t - \sum_{j=1}^{p} \phi_j x_{t-j} - \sum_{k=1}^{q} \theta_k w_{t-k}(\beta), \tag{4.21}$$

emphasizing the dependence of the errors on the parameters (recall that $w_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}$ by invertibility, and the π_j are complicated functions of β).

Again we have the problem that we don't observe the x_t for $t \leq 0$, nor the errors w_t. For *conditional least squares*, we condition on $x_1, \ldots, x_p$ (if $p > 0$) and set $w_p = w_{p-1} = w_{p-2} = \cdots = w_{p+1-q} = 0$ (if $q > 0$), in which case, given β, we may evaluate (4.21) for $t = p + 1, \ldots, n$. For example, for an ARMA(1, 1),

$$x_t = \phi x_{t-1} + \theta w_{t-1} + w_t ,$$

we would start at $p+1=2$ and set $w_1=0$ so that

$$\begin{aligned} w_2 &= x_2 - \phi x_1 - \theta w_1 = x_2 - \phi x_1 \\ w_3 &= x_3 - \phi x_2 - \theta w_2 \\ &\vdots \\ w_n &= x_n - \phi x_{n-1} - \theta w_{n-1} \end{aligned}$$

Given data, we can evaluate these errors at any values of the parameters; e.g., $\phi = \theta = .5$. Using this conditioning argument, the conditional error sum of squares is

$$S_c(\beta) = \sum_{t=p+1}^{n} w_t^2(\beta). \tag{4.22}$$

Minimizing $S_c(\beta)$ with respect to β yields the conditional least squares estimates. We could use a brute force method where we evaluate $S_c(\beta)$ over a grid of possible values for the parameters and choose the values with the smallest error sum of squares, but this method becomes prohibitive if there are many parameters.

If $q=0$, the problem is linear regression as we saw in the case of the AR(1). If $q>0$, the problem becomes nonlinear regression and we will rely on numerical optimization. Gauss–Newton is an iterative method for solving the problem of minimizing (4.22). We demonstrate the method for an MA(1).

Example 4.26. Gauss–Newton for an MA(1)
Consider an MA(1) process, $x_t = w_t + \theta w_{t-1}$. Write the errors as

$$w_t(\theta) = x_t - \theta w_{t-1}(\theta), \quad t = 1, \ldots, n, \tag{4.23}$$

where we condition on $w_0(\theta) = 0$. Our goal is to find the value of θ that minimizes $S_c(\theta) = \sum_{t=1}^{n} w_t^2(\theta)$, which is a nonlinear function of θ.

Let $\theta_{(0)}$ be an initial estimate of θ, for example the method of moments estimate. Now we use a first-order Taylor approximation of $w_t(\theta)$ at $\theta_{(0)}$ to get

$$S_c(\theta) = \sum_{t=1}^{n} w_t^2(\theta) \approx \sum_{t=1}^{n} \big[w_t(\theta_{(0)}) - (\theta - \theta_{(0)}) z_t(\theta_{(0)})\big]^2, \tag{4.24}$$

where

$$z_t(\theta_{(0)}) = -\frac{\partial w_t(\theta)}{\partial \theta}\bigg|_{\theta=\theta_{(0)}},$$

(writing the derivative in the negative simplifies the algebra at the end). It turns out that the derivatives have a simple form that makes them easy to evaluate. Taking derivatives in (4.23),

$$\frac{\partial w_t(\theta)}{\partial \theta} = -w_{t-1}(\theta) - \theta \frac{\partial w_{t-1}(\theta)}{\partial \theta}, \quad t = 1, \ldots, n, \tag{4.25}$$

where we set $\partial w_0(\theta)/\partial\theta = 0$. We can also write (4.25) as

$$z_t(\theta) = w_{t-1}(\theta) - \theta z_{t-1}(\theta), \quad t = 1, \ldots, n, \tag{4.26}$$

where $z_0(\theta) = 0$. This implies that the derivative sequence is an AR process, which we may easily compute recursively given a value of θ.

We will write the right side of (4.24) as

$$Q(\theta) = \sum_{t=1}^{n} \Big[\underbrace{w_t(\theta_{(0)})}_{y_t} - \underbrace{(\theta - \theta_{(0)})}_{\beta} \underbrace{z_t(\theta_{(0)})}_{z_t} \Big]^2 \tag{4.27}$$

and this is the quantity that we will minimize. The problem is now simple linear regression ("$y_t = \beta z_t + \epsilon_t$"), so that

$$\widehat{(\theta - \theta_{(0)})} = \textstyle\sum_{t=1}^{n} z_t(\theta_{(0)}) w_t(\theta_{(0)}) \;/\; \sum_{t=1}^{n} z_t^2(\theta_{(0)}),$$

or

$$\hat{\theta} = \theta_{(0)} + \textstyle\sum_{t=1}^{n} z_t(\theta_{(0)}) w_t(\theta_{(0)}) \;/\; \sum_{t=1}^{n} z_t^2(\theta_{(0)}).$$

Consequently, the Gauss–Newton procedure in this case is, on iteration $j+1$, set

$$\theta_{(j+1)} = \theta_{(j)} + \frac{\sum_{t=1}^{n} z_t(\theta_{(j)}) w_t(\theta_{(j)})}{\sum_{t=1}^{n} z_t^2(\theta_{(j)})}, \quad j = 0, 1, 2, \ldots, \tag{4.28}$$

where the values in (4.28) are calculated recursively using (4.23) and (4.26). The calculations are stopped when $|\theta_{(j+1)} - \theta_{(j)}|$, or $|Q(\theta_{(j+1)}) - Q(\theta_{(j)})|$, are smaller than some preset amount. ♢

Example 4.27. Fitting the Glacial Varve Series

Consider the glacial varve series (say x_t) analyzed in Example 3.12 and in Problem 3.6, where it was argued that a first-order moving average model might fit the logarithmically transformed and differenced varve series, say,

$$\nabla \log(x_t) = \log(x_t) - \log(x_{t-1}).$$

The transformed series and the sample ACF and PACF are shown in Figure 4.6 and based on Table 4.1, confirm the tendency of $\nabla \log(x_t)$ to behave as a first-order moving average. The code to display the output of Figure 4.6 is:

```
tsplot(diff(log(varve)), col=4, ylab=expression(nabla~log~X[~t]),
          main="Transformed Glacial Varves")
acf2(diff(log(varve)))
```

We see $\hat{\rho}(1) = -.4$ and using method of moments for our initial estimate:

$$\theta_{(0)} = \frac{1 - \sqrt{1 - 4\hat{\rho}(1)^2}}{2\hat{\rho}(1)} = -.5$$

based on Example 4.25 and the quadratic formula. The R code to run the Gauss–Newton and the results are:

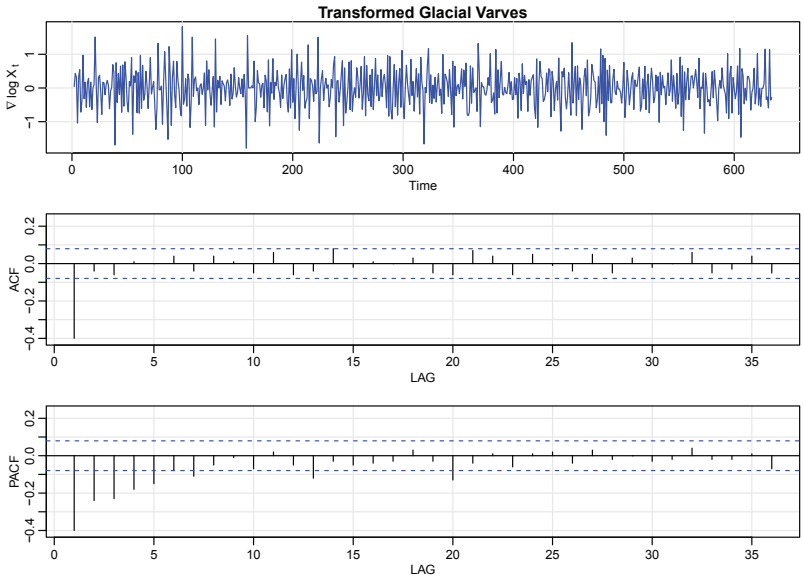

Figure 4.6 *Transformed glacial varves and corresponding sample ACF and PACF.*

```
x = diff(log(varve))                          # data
r = acf1(x, 1, plot=FALSE)                    # acf(1)
c(0) -> w -> z -> Sc -> Sz -> Szw -> para # initialize
num  = length(x)                              # = 633
## Estimation
para[1] = (1-sqrt(1-4*(r^2)))/(2*r)           # MME
niter   = 12
for (j in 1:niter){
  for  (i in 2:num){ w[i] = x[i]   - para[j]*w[i-1]
                     z[i] = w[i-1] - para[j]*z[i-1]
  }
  Sc[j]     = sum(w^2)
  Sz[j]     = sum(z^2)
  Szw[j]    = sum(z*w)
  para[j+1] = para[j] + Szw[j]/Sz[j]
}
## Results
cbind(iteration=1:niter-1, thetahat=para[1:niter], Sc, Sz)
iteration    thetahat        Sc        Sz
        0  -0.5000000  158.4258  172.1110
        1  -0.6704205  150.6786  236.8917
        2  -0.7340825  149.2539  301.6214
        3  -0.7566814  149.0291  337.3468
        4  -0.7656857  148.9893  354.4164
        5  -0.7695230  148.9817  362.2777
```

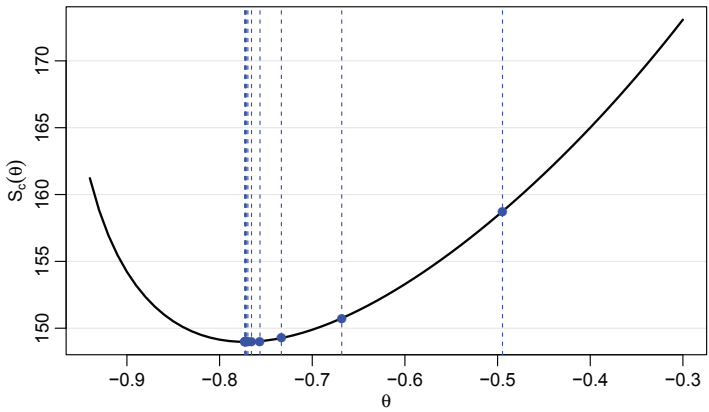

Figure 4.7 *Conditional sum of squares versus values of the moving average parameter for the glacial varve example, Example 4.27. Vertical lines indicate the values of the parameter obtained via Gauss–Newton.*

```
 6  -0.7712091  148.9802  365.8518
 7  -0.7719602  148.9799  367.4683
 8  -0.7722968  148.9799  368.1978
 9  -0.7724482  148.9799  368.5266
10  -0.7725162  148.9799  368.6748
11  -0.7725469  148.9799  368.7416
```

The estimate is

$$\hat{\theta} = \theta_{(11)} = -.773\,,$$

which results in the conditional sum of squares at convergence being

$$S_c(-.773) = 148.98\,.$$

The final estimate of the error variance is

$$\hat{\sigma}_w^2 = \frac{148.98}{632} = .236$$

with 632 degrees of freedom. The value of the sum of the squared derivatives at convergence is $\sum_{t=1}^{n} z_t^2(\theta_{(11)}) = 368.74$ and consequently, the estimated standard error of $\hat{\theta}$ is

$$\mathrm{SE}(\hat{\theta}) = \sqrt{.236/368.74} = .025$$

using the standard regression results as an approximation. This leads to a t-value of $-.773/.025 = -30.92$ with 632 degrees of freedom.

Figure 4.7 displays the conditional sum of squares, $S_c(\theta)$ as a function of θ, as well as indicating the values of each step of the Gauss–Newton algorithm. Note that the Gauss–Newton procedure takes large steps toward the minimum initially, and then takes very small steps as it gets close to the minimizing value.

```
## Plot conditional SS
c(0) -> w -> cSS
th = -seq(.3, .94, .01)
for (p in 1:length(th)){
  for (i in 2:num){ w[i] = x[i] - th[p]*w[i-1]
  }
  cSS[p] = sum(w^2)
}
tsplot(th, cSS, ylab=expression(S[c](theta)), xlab=expression(theta))
abline(v=para[1:12], lty=2, col=4)    #  add previous results to plot
points(para[1:12], Sc[1:12], pch=16, col=4)
```

◊

Unconditional Least Squares and MLE

Estimation of the parameters in an ARMA model is more like weighted least squares than ordinary least squares. Consider the normal regression model

$$x_t = \beta_0 + \beta_1 z_t + \epsilon_t\,,$$

where now, the errors have possibly different variances,

$$\epsilon_t \sim N(0, \sigma^2 h_t)\,.$$

In this case, we use weighted least squares to minimize

$$S(\beta) = \sum_{t=1}^{n} \frac{\epsilon_t^2(\beta)}{h_t} = \sum_{t=1}^{n} \frac{1}{h_t}\Big(x_t - [\beta_0 + \beta_1 z_t]\Big)^2$$

with respect to the βs. This problem is more difficult because the weights, $1/h_t$, are often unknown (the case $h_t = 1$ is ordinary least squares). For ARMA models, however, we do know the structure of these variances.

For ease, we'll concentrate on the full AR(1) model,

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t \tag{4.29}$$

where $|\phi| < 1$ and $w_t \sim$ iid $\mathrm{N}(0, \sigma_w^2)$. Given data $x_1, x_2, \ldots, x_n$, we cannot regress x_1 on x_0 because it is not observed. However, we know from Example 4.1 that

$$x_1 = \mu + \epsilon_1 \qquad \epsilon_1 \sim N(0, \sigma_w^2/(1-\phi^2))\,.$$

In this case, we have $h_1 = 1/(1-\phi^2)$. For $t = 2, \ldots, n$, the model is ordinary linear regression with w_t as the regression error, so that $h_t = 1$ for $t \geq 2$. Thus, the unconditional sum of squares is now

$$S(\mu, \phi) = (1-\phi^2)(x_1 - \mu)^2 + \sum_{t=2}^{n} [(x_t - \mu) - \phi(x_{t-1} - \mu)]^2\,. \tag{4.30}$$

In conditional least squares, we conditioned away the nasty part involving x_1 to make the problem easier. For unconditional least squares, we need to use numerical optimization even for the simple AR(1) case.

This problem generalizes in an obvious way to AR(p) models and in a not so obvious way to ARMA models. For us, unconditional least squares is equivalent to maximum likelihood estimation (MLE). MLE involves finding the "most likely" parameters given the data and is discussed further in Section D.1. In the general case of causal and invertible ARMA(p, q) models, maximum likelihood estimation, least squares estimation (conditional and unconditional), and Yule–Walker estimation in the case of AR models, all lead to optimal estimators for large sample sizes.

Example 4.28. Transformed Glacial Varves (cont)
In Example 4.27, we used Gauss–Newton to fit an MA(1) model to the transformed glacial varve series via conditional least squares. To use unconditional least squares (equivalently MLE), we can use the script `sarima` from `astsa` as follows. The script requires specification of the AR order (p), the MA order (q), and the order of differencing (d). In this case, we are already differencing the data, so we set $d = 0$; we will discuss this further in the next chapter. In addition, the transformed data appear to have a zero mean function so we do not fit a mean to the data. This is accomplished by specifying `no.constant=TRUE` in the call.

```
sarima(diff(log(varve)), p=0, d=0, q=1, no.constant=TRUE)
# partial output
 initial  value -0.551778
 iter   2 value -0.671626
 iter   3 value -0.705973
 iter   4 value -0.707314
 iter   5 value -0.722372
 iter   6 value -0.722738   # conditional SS
 iter   7 value -0.723187
 iter   8 value -0.723194
 iter   9 value -0.723195
 final  value -0.723195
 converged
 initial  value -0.722700
 iter   2 value -0.722702   # unconditional SS (MLE)
 iter   3 value -0.722702
 final  value -0.722702
 converged
 ---
 Coefficients:
          ma1
      -0.7705
 s.e.   0.0341
 sigma^2 estimated as 0.2353:  log likelihood = -440.72,  aic = 885.44
```

The script starts by using the data to pick initial values of the estimates that are

within the causal and invertible region of the parameter space. Then, the script uses conditional least squares as in Example 4.27. Once that process has converged, the next step is to use the conditional estimates to find the unconditional least squares estimates (or MLEs).

The output shows only the iteration number and the value of the sum of squares. It is a good idea to look at the results of the numerical optimization to make sure it converges and that there are no warnings. If there is trouble converging or there are warnings, it usually means that the proposed model is not even close to reality.

The final estimates are $\hat{\theta} = -.7705_{(.034)}$ and $\hat{\sigma}_w^2 = .2353$. These are nearly the values obtained in Example 4.27, which were $\hat{\theta} = -.771_{(.025)}$ and $\hat{\sigma}_w^2 = .236$. ◊

Most packages use large sample theory to evaluate the estimated standard errors (standard deviation of an estimate). We give a few examples in the following proposition.

Property 4.29 (Some Specific Large Sample Distributions). *In the following, read* AN *as "approximately normal for large sample size".*

AR(1):

$$\hat{\phi}_1 \sim \text{AN}\left[\phi_1, n^{-1}(1-\phi_1^2)\right] \tag{4.31}$$

Thus, an approximate $100(1-\alpha)\%$ *confidence interval for* ϕ_1 *is*

$$\hat{\phi}_1 \pm z_{\alpha/2}\sqrt{\tfrac{1-\hat{\phi}_1^2}{n}}\,.$$

AR(2):

$$\hat{\phi}_1 \sim \text{AN}\left[\phi_1, n^{-1}(1-\phi_2^2)\right] \quad \text{and} \quad \hat{\phi}_2 \sim \text{AN}\left[\phi_2, n^{-1}(1-\phi_2^2)\right] \tag{4.32}$$

Thus, approximate $100(1-\alpha)\%$ *confidence intervals for* ϕ_1 *and* ϕ_2 *are*

$$\hat{\phi}_1 \pm z_{\alpha/2}\sqrt{\tfrac{1-\hat{\phi}_2^2}{n}} \quad \textit{and} \quad \hat{\phi}_2 \pm z_{\alpha/2}\sqrt{\tfrac{1-\hat{\phi}_2^2}{n}}\,.$$

MA(1):

$$\hat{\theta}_1 \sim \text{AN}\left[\theta_1, n^{-1}(1-\theta_1^2)\right] \tag{4.33}$$

Confidence intervals for the MA examples are similar to the AR examples.

MA(2):

$$\hat{\theta}_1 \sim \text{AN}\left[\theta_1, n^{-1}(1-\theta_2^2)\right] \quad \text{and} \quad \hat{\theta}_2 \sim \text{AN}\left[\theta, n^{-1}(1-\theta_2^2)\right] \tag{4.34}$$

Example 4.30. Overfitting Caveat

The large sample behavior of the parameter estimators gives us an additional insight into the problem of fitting ARMA models to data. For example, suppose a time series follows an AR(1) process and we decide to fit an AR(2) to the data. Do any problems occur in doing this? More generally, why not simply fit large-order

AR models to make sure that we capture the dynamics of the process? After all, if the process is truly an AR(1), the other autoregressive parameters will not be significant. The answer is that if we *overfit*, we obtain less efficient, or less precise parameter estimates. For example, if we fit an AR(1) to an AR(1) process, for large n, $\mathrm{var}(\hat{\phi}_1) \approx n^{-1}(1-\phi_1^2)$. But, if we fit an AR(2) to the AR(1) process, for large n, $\mathrm{var}(\hat{\phi}_1) \approx n^{-1}(1-\phi_2^2) = n^{-1}$ because $\phi_2 = 0$. Thus, the variance of ϕ_1 has been inflated, making the estimator less precise.

We do want to mention, however, that overfitting can be used as a diagnostic tool. For example, if we fit an AR(1) model to the data and are satisfied with that model, then adding one more parameter and fitting an AR(2) should lead to approximately the same model as in the AR(1) fit. We will discuss model diagnostics in more detail in Section 5.2. ◊

4.4 Forecasting

In forecasting, the goal is to predict future values of a time series, x_{n+m}, $m = 1, 2, \ldots$, based on the data, $x_1, \ldots, x_n$, collected to the present. Throughout this section, we will assume that the model parameters are known. When the parameters are unknown, we replace them with their estimates.

To understand how to forecast an ARMA process, it is instructive to investigate forecasting an AR(1),

$$x_t = \phi x_{t-1} + w_t\,.$$

First, consider *one-step-ahead prediction*, that is, given data $x_1, \ldots, x_n$, we wish to forecast the value of the time series at the next time point, x_{n+1}. We will call the forecast x_{n+1}^n. In general, the notation x_t^n refers to what we can expect x_t to be given the data $x_1, \ldots, x_n$.[2] Since

$$x_{n+1} = \phi x_n + w_{n+1}\,,$$

we should have

$$x_{n+1}^n = \phi x_n^n + w_{n+1}^n\,.$$

But since we know x_n (it is one of our observations), $x_n^n = x_n$, and since w_{n+1} is a future error and independent of $x_1, \ldots, x_n$, we have $w_{n+1}^n = E(w_{n+1}) = 0$. Consequently, the *one-step-ahead forecast* is

$$x_{n+1}^n = \phi x_n\,. \tag{4.35}$$

The one-step-ahead *mean squared prediction error* (MSPE) is given by

$$P_{n+1}^n = E[x_{n+1} - x_{n+1}^n]^2 = E[x_{n+1} - \phi x_n]^2 = E w_{n+1}^2 = \sigma_w^2.$$

The two-step-ahead forecast is obtained similarly. Since the model is

$$x_{n+2} = \phi x_{n+1} + w_{n+2}\,,$$

[2] Formally $x_t^n = E(x_t \mid x_1, \ldots, x_n)$ is conditional expectation, which is discussed in Section B.4.

we should have

$$x^n_{n+2} = \phi x^n_{n+1} + w^n_{n+2}\,.$$

Again, w_{n+2} is a future error, so $w^n_{n+2} = 0$. Also, we already know $x^n_{n+1} = \phi x_n$, so the forecast is

$$x^n_{n+2} = \phi x^n_{n+1} = \phi^2 x_n\,. \tag{4.36}$$

The two-step-ahead MSPE is given by

$$\begin{aligned} P^n_{n+2} &= E[x_{n+2} - x^n_{n+2}]^2 = E[\phi x_{n+1} + w_{n+2} - \phi^2 x_n]^2 \\ &= E[w_{n+2} + \phi(x_{n+1} - \phi x_n)]^2 = E[w_{n+2} + \phi w_{n+1}]^2 = \sigma_w^2(1+\phi^2). \end{aligned}$$

Generalizing these results, it is easy to see that the m-step-ahead forecast is.

$$x^n_{n+m} = \phi^m x_n\,, \tag{4.37}$$

with MSPE

$$P^n_{n+m} = E[x_{n+m} - x^n_{n+m}]^2 = \sigma_w^2(1 + \phi^2 + \cdots + \phi^{2(m-1)})\,. \tag{4.38}$$

for $m = 1, 2, \ldots$.

Note that since $|\phi| < 1$, we will have $\phi^m \to 0$ fast as $m \to \infty$. Thus the forecasts in (4.37) will soon go to zero (or the mean) and become useless. In addition, the MSPE will converge to $\sigma_w^2 \sum_{j=0}^{\infty} \phi^{2j} = \sigma_w^2/(1-\phi^2)$, which is the variance of the process x_t; recall (4.3).

Forecasting an AR(p) model is basically the same as forecasting an AR(1) provided the sample size n is larger than the order p, which it is most of the time. Since MA(q) and ARMA(p,q) are AR(∞) by invertibility, the same basic techniques can be used. Because ARMA models are invertible; i.e., $w_t = x_t + \sum_{j=1}^{\infty} \pi_j x_{t-j}$, we may write

$$x_{n+m} = -\sum_{j=1}^{\infty} \pi_j x_{n+m-j} + w_{n+m}.$$

If we had the infinite history $\{x_n, x_{n-1}, \ldots, x_1, x_0, x_{-1}, \ldots\}$, of the data available, we would predict x_{n+m} by

$$x^n_{n+m} = -\sum_{j=1}^{\infty} \pi_j x^n_{n+m-j}$$

successively for $m = 1, 2, \ldots$. In this case, $x^n_t = x_t$ for $t = n, n-1, \ldots$. We only have the actual data $\{x_n, x_{n-1}, \ldots, x_1\}$ available, but a practical solution is to truncate the forecasts as

$$x^n_{n+m} = -\sum_{j=1}^{n+m-1} \pi_j x^n_{n+m-j},$$

with $x^n_t = x_t$ for $1 \le t \le n$. For ARMA models in general, as long as n is large,

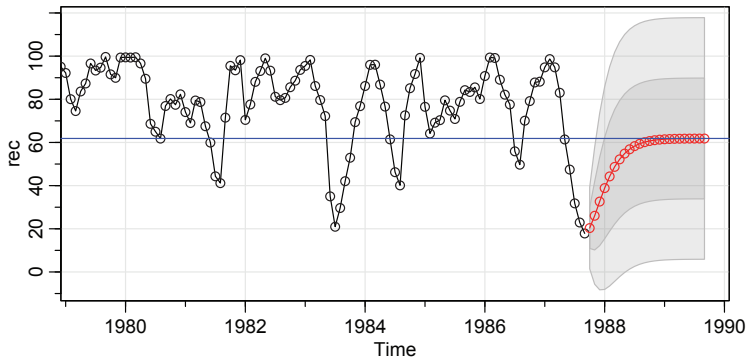

Figure 4.8 *Twenty-four-month forecasts for the Recruitment series. The actual data shown are from about January 1979 to September 1987, and then the forecasts plus and minus one and two standard error are displayed. The solid horizontal line is the estimated mean function.*

the approximation works well because the π-weights are going to zero exponentially fast. For large n, it can be shown (see Problem 4.10) that the mean squared prediction error for ARMA(p,q) models is approximately (exact if $q=0$)

$$P_{n+m}^{n} = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2 . \tag{4.39}$$

We saw this result in (4.38) for the AR(1) because in that case, $\psi_j^2 = \phi^{2j}$.

Example 4.31. Forecasting the Recruitment Series
In Example 4.21 we fit an AR(2) model to the Recruitment series using OLS. Here, we use maximum likelihood estimation (MLE), which is similar to unconditional least squares for ARMA models:

```
sarima(rec, p=2, d=0, q=0)    # fit the model
       Estimate     SE  t.value p.value
 ar1     1.3512 0.0416  32.4933       0
 ar2    -0.4612 0.0417 -11.0687       0
 xmean  61.8585 4.0039  15.4494       0
```

The results are nearly the same as using OLS. Using the parameter estimates as the actual parameter values, the forecasts and root MSPEs can be calculated in a similar fashion to the introduction to this section.

Figure 4.8 shows the result of forecasting the Recruitment series over a 24-month horizon, $m = 1, 2, \ldots, 24$, obtained in R as

```
sarima.for(rec, n.ahead=24, p=2, d=0, q=0)
abline(h=61.8585, col=4)   # display estimated mean
```

Note how the forecast levels off to the mean quickly and the prediction intervals are wide and become constant. That is, because of the short memory, the forecasts settle

to the estimated mean, 61.86, and the root MSPE becomes quite large (and eventually settles at the standard deviation of all the data). ◊

Problems

4.1. For an MA(1), $x_t = w_t + \theta w_{t-1}$, show that $|\rho_x(1)| \leq 1/2$ for any number θ. For which values of θ does $\rho_x(1)$ attain its maximum and minimum?

4.2. Let $\{w_t; t = 0, 1, \dots\}$ be a white noise process with variance σ_w^2 and let $|\phi| < 1$ be a constant. Consider the process $x_0 = w_0$, and

$$x_t = \phi x_{t-1} + w_t, \quad t = 1, 2, \dots .$$

We might use this method to simulate an AR(1) process from simulated white noise.

(a) Show that $x_t = \sum_{j=0}^{t} \phi^j w_{t-j}$ for any $t = 0, 1, \dots$.

(b) Find the $E(x_t)$.

(c) Show that, for $t = 0, 1, \dots$,

$$\text{var}(x_t) = \frac{\sigma_w^2}{1-\phi^2}(1 - \phi^{2(t+1)})$$

(d) Show that, for $h \geq 0$,

$$\text{cov}(x_{t+h}, x_t) = \phi^h \text{var}(x_t)$$

(e) Is x_t stationary?

(f) Argue that, as $t \to \infty$, the process becomes stationary, so in a sense, x_t is "asymptotically stationary."

(g) Comment on how you could use these results to simulate n observations of a stationary Gaussian AR(1) model from simulated iid N(0,1) values.

(h) Now suppose $x_0 = w_0/\sqrt{1-\phi^2}$. Is this process stationary? *Hint*: Show $\text{var}(x_t)$ is constant.

4.3. Consider the following two models:

(i) $x_t = .80x_{t-1} - .15x_{t-2} + w_t - .30w_{t-1}$.

(ii) $x_t = x_{t-1} - .50x_{t-2} + w_t - w_{t-1}$.

(a) Using Example 4.10 as a guide, check the models for parameter redundancy. If a model has redundancy, find the reduced form of the model.

(b) A way to tell if an ARMA model is causal is to examine the roots of AR term $\phi(B)$ to see if there are no roots less than or equal to one in magnitude. Likewise, to determine invertibility of a model, the roots of the MA term $\theta(B)$ must not be less than or equal to one in magnitude. Use Example 4.11 as a guide to determine if the reduced (if appropriate) models (i) and (ii), are causal and/or invertible.

(c) In Example 4.3 and Example 4.12, we used `ARMAtoMA` and `ARMAtoAR` to exhibit some of the coefficients of the causal [MA(∞)] and invertible [AR(∞)] representations of a model. If the model is in fact causal or invertible, the coefficients must converge to zero fast. For each of the reduced (if appropriate) models (i) and (ii), find the first 50 coefficients and comment.

4.4.

(a) Compare the *theoretical* ACF and PACF of an ARMA$(1,1)$, an ARMA$(1,0)$, and an ARMA$(0,1)$ series by plotting the ACFs and PACFs of the three series for $\phi = .6$, $\theta = .9$. Comment on the capability of the ACF and PACF to determine the order of the models. *Hint:* See the code for Example 4.18.

(b) Use `arima.sim` to generate $n = 100$ observations from each of the three models discussed in (a). Compute the sample ACFs and PACFs for each model and compare it to the theoretical values. How do the results compare with the general results given in Table 4.1?

(c) Repeat (b) but with $n = 500$. Comment.

4.5. Let c_t be the cardiovascular mortality series (`cmort`) discussed in Example 3.5 and let $x_t = \nabla c_t$ be the differenced data.

(a) Plot x_t and compare it to the actual data plotted in Figure 3.2. Why does differencing seem reasonable in this case?

(b) Calculate and plot the sample ACF and PACF of x_t and using Table 4.1, argue that an AR(1) is appropriate for x_t.

(c) Fit an AR(1) to x_t using maximum likelihood (basically unconditional least squares) as in Section 4.3. The easiest way to do this is to use `sarima` from `astsa`. Comment on the significance of the regression parameter estimates of the model. What is the estimate of the white noise variance?

(d) Examine the residuals and comment on whether or not you think the residuals are white.

(e) Assuming the fitted model is the true model, find the forecasts over a four-week horizon, x^n_{n+m}, for $m = 1, 2, 3, 4$, and the corresponding 95% prediction intervals; $n = 508$ here. The easiest way to do this is to use `sarima.for` from `astsa`.

(f) Show how the values obtained in part (e) were calculated.

(g) What is the one-step-ahead forecast of the actual value of cardiovascular mortality; i.e., what is c^n_{n+1}?

4.6. For an AR(1) model, determine the general form of the m-step-ahead forecast x^n_{n+m} and show

$$E[(x_{n+m} - x^n_{n+m})^2] = \sigma_w^2 \frac{1 - \phi^{2m}}{1 - \phi^2}.$$

4.7. Repeat the following numerical exercise five times. Generate $n = 100$ iid

N(0, 1) observations. Fit an ARMA(1, 1) model to the data. Compare the parameter estimates in each case and explain the results.

4.8. Generate 10 realizations of length $n = 200$ each of an ARMA(1,1) process with $\phi = .9, \theta = .5$ and $\sigma^2 = 1$. Find the MLEs of the three parameters in each case and compare the estimators to the true values.

4.9. Using Example 4.26 as your guide, find the Gauss–Newton procedure for estimating the autoregressive parameter, ϕ, from the AR(1) model, $x_t = \phi x_{t-1} + w_t$, given data $x_1, \ldots, x_n$. Does this procedure produce the unconditional or the conditional estimator?

4.10. (Forecast Errors) In (4.39), we stated without proof that, for large n, the mean squared prediction error for ARMA(p, q) models is approximately (exact if $q = 0$) $P^n_{n+m} = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2$. To establish (4.39), write a future observation in terms of its causal representation, $x_{n+m} = \sum_{j=0}^{\infty} \psi_j w_{m+n-j}$. Show that if an infinite history, $\{x_n, x_{n-1}, \ldots, x_1, x_0, x_{-1}, \ldots\}$, is available, then

$$x^n_{n+m} = \sum_{j=0}^{\infty} \psi_j w^n_{m+n-j} = \sum_{j=m}^{\infty} \psi_j w_{m+n-j}\,.$$

Now, use this result to show that

$$E[x_{n+m} - x^n_{n+m}]^2 = E\Big[\sum_{j=0}^{m-1} \psi_j w_{n+m-j}\Big]^2 = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2\,.$$