# STAT 626: Outline of Lectures 16 and 17
# ARMA Model Estimation (§3.6)

1. **Method of Moments (1900): Yule-Wakler Estimators (1927, 1933)**

   Example: For causal AR(1), $x_t = \phi x_{t-1} + w_t$, there are two parameters $(\phi, \sigma_w^2)$ to be estimated. satisfy

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

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

   The two equations together are referred to as the Yule-Walker (YW) equations for the AR(1) model.

   Replacing $\gamma(h), \quad h = 0, 1 \ldots$, by their estimators $\widehat{\gamma}(h), \quad h = 0, 1 \ldots$, one obtains the sample YW equations. Solve them to obtain the YW estimates of $(\phi, \sigma_w^2)$.

   See Definition 3.10 for the AR(p) Model,

   Property 3.8,

   Property 3.9

   Examples 3.26-3.28.

2. Lecture 17: **Maximum Likelihood (1920's) and LS (1805) Estimators,**
   **The likelihood function: The probability of seeing what you have seen!**

   One would like to maximize this as function of the parameters.
   Usually, it is nonlinear in the parameters.

   How does one maximize such a function?

   Grid Search, The Newton-Raphson and Scoring Algorithms, Asymptotic Property 3.10,

Example 3.33.

## Wiki-Review of PACF and AR Models:

The correlation coefficient between $x_t$ and $x_{t+h}$ after removing the linear effects of the intervening variables $\{x_{t+1}, \ldots, x_{t+h-1}\}$ is called the lag-$h$ partial autocorrelation of a stationary time series and denoted by $\phi_{hh}, h = 1, 2, \ldots,$. The plot of $\phi_{hh}$ vs $h = 1, 2, \ldots$ is call the partial correlogram.

**Given the Time Series Data $x_1, \ldots, x_n$ from a Causal AR($p$) Model:**

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

with known order $p$ and parameters $(\phi_1, \ldots, \phi_p, \sigma_w^2)$. The goal is to estimate these parameters as ”good” as one can.

How does one do this? What does ”good” mean here?

AR(p) looks like a regression model, one idea is to use LS estimation, i.e. estimate the $\phi_i$,s by minimizing the SEE,

$$\sum_{t=p+1}^{n} (x_t - \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p})^2,$$

which is a quadratic function of the parameters $(\phi_1, \ldots, \phi_p)$. Then, one proceeds as in estimation in the linear regression setup.

How does one fine the LSE of $\theta$ in the MA(1) model

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

Invert the model, truncate the resulting AR($\infty$), proceed as in LS estimation of AR($p$).

## 3.6 Estimation

Throughout this section, we assume we have $n$ observations, $x_1, \dots, x_n$, from a causal and invertible Gaussian ARMA$(p, q)$ process in which, initially, the order parameters, $p$ and $q$, are known. Our goal is to estimate the parameters, $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, and $\sigma_w^2$. We will discuss the problem of determining $p$ and $q$ later in this section.

We begin with method of moments estimators. The idea behind these estimators is that of equating population moments to sample moments and then solving for the parameters in terms of the sample moments. We immediately see that, if $E(x_t) = \mu$, then the method of moments estimator of $\mu$ is the sample average, $\bar{x}$. 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.

When the process is AR$(p)$,

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

the first $p+1$ equations of (3.47) and (3.48) lead to the following:

**Definition 3.10** *The* **Yule–Walker equations** *are given by*

$$\gamma(h) = \phi_1 \gamma(h-1) + \cdots + \phi_p \gamma(h-p), \quad h = 1, 2, \dots, p, \tag{3.98}$$

$$\sigma_w^2 = \gamma(0) - \phi_1 \gamma(1) - \cdots - \phi_p \gamma(p). \tag{3.99}$$

In matrix notation, the Yule–Walker equations are

$$\mathit{\Gamma}_p \boldsymbol{\phi} = \boldsymbol{\gamma}_p, \quad \sigma_w^2 = \gamma(0) - \boldsymbol{\phi}' \boldsymbol{\gamma}_p, \tag{3.100}$$

where $\mathit{\Gamma}_p = \{\gamma(k-j)\}_{j,k=1}^p$ is a $p \times p$ matrix, $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)'$ is a $p \times 1$ vector, and $\boldsymbol{\gamma}_p = (\gamma(1), \dots, \gamma(p))'$ is a $p \times 1$ vector. Using the method of moments, we replace $\gamma(h)$ in (3.100) by $\widehat{\gamma}(h)$ [see equation (1.34)] and solve

$$\widehat{\boldsymbol{\phi}} = \widehat{\mathit{\Gamma}}_p^{-1} \widehat{\boldsymbol{\gamma}}_p, \quad \widehat{\sigma}_w^2 = \widehat{\gamma}(0) - \widehat{\boldsymbol{\gamma}}_p' \widehat{\mathit{\Gamma}}_p^{-1} \widehat{\boldsymbol{\gamma}}_p. \tag{3.101}$$

These estimators are typically called the Yule–Walker estimators. For calculation purposes, it is sometimes more convenient to work with the sample ACF. By factoring $\widehat{\gamma}(0)$ in (3.101), we can write the Yule–Walker estimates as

$$\widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{R}}_p^{-1} \widehat{\boldsymbol{\rho}}_p, \quad \widehat{\sigma}_w^2 = \widehat{\gamma}(0) \left[ 1 - \widehat{\boldsymbol{\rho}}_p' \widehat{\boldsymbol{R}}_p^{-1} \widehat{\boldsymbol{\rho}}_p \right], \tag{3.102}$$

where $\widehat{R}_p = \{\widehat{\rho}(k-j)\}_{j,k=1}^p$ is a $p \times p$ matrix and $\widehat{\boldsymbol{\rho}}_p = (\widehat{\rho}(1), \dots, \widehat{\rho}(p))'$ is a $p \times 1$ vector.

For AR$(p)$ models, if the sample size is large, the Yule–Walker estimators are approximately normally distributed, and $\widehat{\sigma}_w^2$ is close to the true value of $\sigma_w^2$. We state these results in Property 3.8; for details, see Appendix B, §B.3.

**Property 3.8 Large Sample Results for Yule–Walker Estimators**

*The asymptotic ($n \to \infty$) behavior of the Yule–Walker estimators in the case of causal AR(p) processes is as follows:*

$$\sqrt{n}\left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma_w^2 \Gamma_p^{-1}\right), \qquad \widehat{\sigma}_w^2 \xrightarrow{p} \sigma_w^2. \tag{3.103}$$

The Durbin–Levinson algorithm, (3.68)-(3.70), can be used to calculate $\widehat{\boldsymbol{\phi}}$ without inverting $\widehat{\Gamma}_p$ or $\widehat{R}_p$, by replacing $\gamma(h)$ by $\widehat{\gamma}(h)$ in the algorithm. In running the algorithm, we will iteratively calculate the $h \times 1$ vector, $\widehat{\boldsymbol{\phi}}_h = (\widehat{\phi}_{h1}, \dots, \widehat{\phi}_{hh})'$, for $h = 1, 2, \dots$. Thus, in addition to obtaining the desired forecasts, the Durbin–Levinson algorithm yields $\widehat{\phi}_{hh}$, the sample PACF. Using (3.103), we can show the following property.

**Property 3.9 Large Sample Distribution of the PACF**

*For a causal AR(p) process, asymptotically ($n \to \infty$),*

$$\sqrt{n}\, \widehat{\phi}_{hh} \xrightarrow{d} N(0, 1), \quad \text{for} \quad h > p. \tag{3.104}$$

**Example 3.26 Yule–Walker Estimation for an AR(2) Process**

The data shown in Figure 3.3 were $n = 144$ simulated observations from the AR(2) model

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

where $w_t \sim \text{iid N}(0, 1)$. For these data, $\widehat{\gamma}(0) = 8.903$, $\widehat{\rho}(1) = .849$, and $\widehat{\rho}(2) = .519$. Thus,

$$\widehat{\boldsymbol{\phi}} = \begin{pmatrix} \widehat{\phi}_1 \\ \widehat{\phi}_2 \end{pmatrix} = \begin{bmatrix} 1 & .849 \\ .849 & 1 \end{bmatrix}^{-1} \begin{pmatrix} .849 \\ .519 \end{pmatrix} = \begin{pmatrix} 1.463 \\ -.723 \end{pmatrix}$$

and

$$\widehat{\sigma}_w^2 = 8.903 \left[1 - (.849, .519) \begin{pmatrix} 1.463 \\ -.723 \end{pmatrix}\right] = 1.187.$$

By Property 3.8, the asymptotic variance–covariance matrix of $\widehat{\boldsymbol{\phi}}$,

$$\frac{1}{144} \frac{1.187}{8.903} \begin{bmatrix} 1 & .849 \\ .849 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} .058^2 & -.003 \\ -.003 & .058^2 \end{bmatrix},$$

can be used to get confidence regions for, or make inferences about $\widehat{\boldsymbol{\phi}}$ and its components. For example, an approximate 95% confidence interval for $\phi_2$ is $-.723 \pm 2(.058)$, or $(-.838, -.608)$, which contains the true value of $\phi_2 = -.75$.

For these data, the first three sample partial autocorrelations are $\widehat{\phi}_{11} = \widehat{\rho}(1) = .849$, $\widehat{\phi}_{22} = \widehat{\phi}_2 = -.721$, and $\widehat{\phi}_{33} = -.085$. According to Property 3.9, the asymptotic standard error of $\widehat{\phi}_{33}$ is $1/\sqrt{144} = .083$, and the observed value, $-.085$, is about only one standard deviation from $\phi_{33} = 0$.

**Example 3.27 Yule–Walker Estimation of the Recruitment Series**

In Example 3.17 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 in R, which are nearly identical to the values in Example 3.17.

```
rec.yw = ar.yw(rec, order=2)
rec.yw$x.mean   # = 62.26 (mean estimate)
rec.yw$ar       # =  1.33, -.44  (parameter estimates)
sqrt(diag(rec.yw$asy.var.coef))  # = .04, .04  (standard errors)
rec.yw$var.pred  # = 94.80 (error variance estimate)
```

To obtain the 24 month ahead predictions and their standard errors, and then plot the results as in Example 3.24, use the R commands:

```
rec.pr = predict(rec.yw, n.ahead=24)
U = rec.pr$pred + rec.pr$se
L = rec.pr$pred - rec.pr$se
minx = min(rec,L); maxx = max(rec,U)
ts.plot(rec, rec.pr$pred, xlim=c(1980,1990), ylim=c(minx,maxx))
lines(rec.pr$pred, col="red", type="o")
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
```

In the case of AR($p$) models, the Yule–Walker estimators given in (3.102) are optimal in the sense that the asymptotic distribution, (3.103), is the best asymptotic normal distribution. This is because, given initial conditions, AR($p$) models are linear models, and the Yule–Walker estimators are essentially least squares estimators. If we use method of moments for MA or ARMA models, we will not get optimal estimators because such processes are nonlinear in the parameters.

**Example 3.28 Method of Moments Estimation for an MA(1)**

Consider the time series

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

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 $\theta$. 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 $\theta$ is found by solving:

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

Two solutions exist, so we would pick the invertible one. If $|\widehat{\rho}(1)| \le \frac{1}{2}$, 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 $|\widehat{\rho}(1)| \ge \frac{1}{2}$ because it is an estimator. For example, the following simulation in R produces a value of $\widehat{\rho}(1) = .507$ 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)
acf(ma1, plot=FALSE)[1]  # = .507 (lag 1 sample ACF)
```

When $|\widehat{\rho}(1)| < \frac{1}{2}$, the invertible estimate is

$$\widehat{\theta} = \frac{1 - \sqrt{1 - 4\widehat{\rho}(1)^2}}{2\widehat{\rho}(1)}.$$

It can be shown that[5]

$$\widehat{\theta} \sim \mathrm{AN}\left(\theta, \; \frac{1 + \theta^2 + 4\theta^4 + \theta^6 + \theta^8}{n(1 - \theta^2)^2}\right);$$

AN is read *asymptotically normal* and is defined in Definition A.5, page 515, of Appendix A. The maximum likelihood estimator (which we discuss next) of $\theta$, in this case, has an asymptotic variance of $(1 - \theta^2)/n$. When $\theta = .5$, for example, the ratio of the asymptotic variance of the method of moments estimator to the maximum likelihood estimator of $\theta$ is about 3.5. That is, for large samples, the variance of the method of moments estimator is about 3.5 times larger than the variance of the MLE of $\theta$ when $\theta = .5$.

## Maximum Likelihood and Least Squares Estimation

To fix ideas, we first focus on the causal AR(1) case. Let

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t \tag{3.105}$$

where $|\phi| < 1$ and $w_t \sim \text{iid } \mathrm{N}(0, \sigma_w^2)$. Given data $x_1, x_2, \dots, x_n$, we seek the likelihood

$$L(\mu, \phi, \sigma_w^2) = f\left(x_1, x_2, \dots, x_n \;\middle|\; \mu, \phi, \sigma_w^2\right).$$

In the case of an AR(1), we may write the likelihood as

$$L(\mu, \phi, \sigma_w^2) = f(x_1) f(x_2 \mid x_1) \cdots f(x_n \mid x_{n-1}),$$

where we have dropped the parameters in the densities, $f(\cdot)$, to ease the notation. Because $x_t \mid x_{t-1} \sim \mathrm{N}\left(\mu + \phi(x_{t-1} - \mu), \, \sigma_w^2\right)$, we have

$$f(x_t \mid x_{t-1}) = f_w[(x_t - \mu) - \phi(x_{t-1} - \mu)],$$

where $f_w(\cdot)$ is the density of $w_t$, that is, the normal density with mean zero and variance $\sigma_w^2$. We may then write the likelihood as

$$L(\mu, \phi, \sigma_w) = f(x_1) \prod_{t=2}^{n} f_w\left[(x_t - \mu) - \phi(x_{t-1} - \mu)\right].$$

[5] The result follows from Theorem A.7 given in Appendix A and the delta method. See the proof of Theorem A.7 for details on the delta method.

To find $f(x_1)$, we can use the causal representation

$$x_1 = \mu + \sum_{j=0}^{\infty} \phi^j w_{1-j}$$

to see that $x_1$ is normal, with mean $\mu$ and variance $\sigma_w^2/(1-\phi^2)$. Finally, for an AR(1), the likelihood is

$$L(\mu, \phi, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2}(1-\phi^2)^{1/2} \exp\left[-\frac{S(\mu,\phi)}{2\sigma_w^2}\right], \tag{3.106}$$

where

$$S(\mu,\phi) = (1-\phi^2)(x_1-\mu)^2 + \sum_{t=2}^{n} [(x_t-\mu) - \phi(x_{t-1}-\mu)]^2. \tag{3.107}$$

Typically, $S(\mu,\phi)$ is called the unconditional sum of squares. We could have also considered the estimation of $\mu$ and $\phi$ using unconditional least squares, that is, estimation by minimizing $S(\mu,\phi)$.

Taking the partial derivative of the log of (3.106) with respect to $\sigma_w^2$ and setting the result equal to zero, we see that for any given values of $\mu$ and $\phi$ in the parameter space, $\sigma_w^2 = n^{-1}S(\mu,\phi)$ maximizes the likelihood. Thus, the maximum likelihood estimate of $\sigma_w^2$ is

$$\widehat{\sigma}_w^2 = n^{-1} S(\widehat{\mu}, \widehat{\phi}), \tag{3.108}$$

where $\widehat{\mu}$ and $\widehat{\phi}$ are the MLEs of $\mu$ and $\phi$, respectively. If we replace $n$ in (3.108) by $n-2$, we would obtain the unconditional least squares estimate of $\sigma_w^2$.

If, in (3.106), we take logs, replace $\sigma_w^2$ by $\widehat{\sigma}_w^2$, and ignore constants, $\widehat{\mu}$ and $\widehat{\phi}$ are the values that minimize the criterion function

$$l(\mu,\phi) = \log\left[n^{-1}S(\mu,\phi)\right] - n^{-1}\log(1-\phi^2); \tag{3.109}$$

that is, $l(\mu,\phi) \propto -2\log L(\mu,\phi,\widehat{\sigma}_w^2)$.[6] Because (3.107) and (3.109) are complicated functions of the parameters, the minimization of $l(\mu,\phi)$ or $S(\mu,\phi)$ is accomplished numerically. In the case of AR models, we have the advantage that, conditional on initial values, they are linear models. That is, we can drop the term in the likelihood that causes the nonlinearity. Conditioning on $x_1$, the conditional likelihood becomes

$$\begin{aligned} L(\mu,\phi,\sigma_w^2 \mid x_1) &= \prod_{t=2}^{n} f_w\left[(x_t-\mu) - \phi(x_{t-1}-\mu)\right] \\ &= (2\pi\sigma_w^2)^{-(n-1)/2} \exp\left[-\frac{S_c(\mu,\phi)}{2\sigma_w^2}\right], \end{aligned} \tag{3.110}$$

[6] The criterion function is sometimes called the profile or concentrated likelihood.

where the conditional sum of squares is

$$S_c(\mu, \phi) = \sum_{t=2}^{n} \left[(x_t - \mu) - \phi(x_{t-1} - \mu)\right]^2 . \tag{3.111}$$

The conditional MLE of $\sigma_w^2$ is

$$\widehat{\sigma}_w^2 = S_c(\widehat{\mu}, \widehat{\phi})/(n-1), \tag{3.112}$$

and $\widehat{\mu}$ and $\widehat{\phi}$ are the values that minimize the conditional sum of squares, $S_c(\mu, \phi)$. Letting $\alpha = \mu(1-\phi)$, the conditional sum of squares can be written as

$$S_c(\mu, \phi) = \sum_{t=2}^{n} \left[x_t - (\alpha + \phi x_{t-1})\right]^2 . \tag{3.113}$$

The problem is now the linear regression problem stated in §2.2. Following the results from least squares estimation, we have $\widehat{\alpha} = \bar{x}_{(2)} - \widehat{\phi}\bar{x}_{(1)}$, where $\bar{x}_{(1)} = (n-1)^{-1}\sum_{t=1}^{n-1} x_t$, and $\bar{x}_{(2)} = (n-1)^{-1}\sum_{t=2}^{n} x_t$, and the conditional estimates are then

$$\widehat{\mu} = \frac{\bar{x}_{(2)} - \widehat{\phi}\bar{x}_{(1)}}{1 - \widehat{\phi}} \tag{3.114}$$

$$\widehat{\phi} = \frac{\sum_{t=2}^{n}(x_t - \bar{x}_{(2)})(x_{t-1} - \bar{x}_{(1)})}{\sum_{t=2}^{n}(x_{t-1} - \bar{x}_{(1)})^2}. \tag{3.115}$$

From (3.114) and (3.115), we see that $\widehat{\mu} \approx \bar{x}$ and $\widehat{\phi} \approx \widehat{\rho}(1)$. That is, the Yule–Walker estimators and the conditional least squares estimators are approximately the same. The only difference is the inclusion or exclusion of terms involving the endpoints, $x_1$ and $x_n$. We can also adjust the estimate of $\sigma_w^2$ in (3.112) to be equivalent to the least squares estimator, that is, divide $S_c(\widehat{\mu}, \widehat{\phi})$ by $(n-3)$ instead of $(n-1)$ in (3.112).

For general AR($p$) models, maximum likelihood estimation, unconditional least squares, and conditional least squares follow analogously to the AR(1) example. For general ARMA models, it is difficult to write the likelihood as an explicit function of the parameters. Instead, it is advantageous to write the likelihood in terms of the innovations, or one-step-ahead prediction errors, $x_t - x_t^{t-1}$. This will also be useful in Chapter 6 when we study state-space models.

For a normal ARMA($p, q$) model, let $\boldsymbol{\beta} = (\mu, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)'$ be the $(p+q+1)$-dimensional vector of the model parameters. The likelihood can be written as

$$L(\boldsymbol{\beta}, \sigma_w^2) = \prod_{t=1}^{n} f(x_t \mid x_{t-1}, \dots, x_1).$$

The conditional distribution of $x_t$ given $x_{t-1}, \dots, x_1$ is Gaussian with mean $x_t^{t-1}$ and variance $P_t^{t-1}$. Recall from (3.71) that $P_t^{t-1} = \gamma(0)\prod_{j=1}^{t-1}(1 - \phi_{jj}^2)$. For ARMA models, $\gamma(0) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2$, in which case we may write

$$P_t^{t-1} = \sigma_w^2 \left\{ \left[ \sum_{j=0}^{\infty} \psi_j^2 \right] \left[ \prod_{j=1}^{t-1} (1 - \phi_{jj}^2) \right] \right\} \overset{\text{def}}{=} \sigma_w^2 \, r_t,$$

where $r_t$ is the term in the braces. Note that the $r_t$ terms are functions only of the regression parameters and that they may be computed recursively as $r_{t+1} = (1 - \phi_{tt}^2) r_t$ with initial condition $r_1 = \sum_{j=0}^{\infty} \psi_j^2$. The likelihood of the data can now be written as

$$L(\boldsymbol{\beta}, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2} \left[ r_1(\boldsymbol{\beta}) r_2(\boldsymbol{\beta}) \cdots r_n(\boldsymbol{\beta}) \right]^{-1/2} \exp\left[ -\frac{S(\boldsymbol{\beta})}{2\sigma_w^2} \right], \tag{3.116}$$

where

$$S(\boldsymbol{\beta}) = \sum_{t=1}^{n} \left[ \frac{(x_t - x_t^{t-1}(\boldsymbol{\beta}))^2}{r_t(\boldsymbol{\beta})} \right]. \tag{3.117}$$

Both $x_t^{t-1}$ and $r_t$ are functions of $\boldsymbol{\beta}$ alone, and we make that fact explicit in (3.116)-(3.117). Given values for $\boldsymbol{\beta}$ and $\sigma_w^2$, the likelihood may be evaluated using the techniques of §3.5. Maximum likelihood estimation would now proceed by maximizing (3.116) with respect to $\boldsymbol{\beta}$ and $\sigma_w^2$. As in the AR(1) example, we have

$$\widehat{\sigma}_w^2 = n^{-1} S(\widehat{\boldsymbol{\beta}}), \tag{3.118}$$

where $\widehat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that minimizes the concentrated likelihood

$$l(\boldsymbol{\beta}) = \log\left[ n^{-1} S(\boldsymbol{\beta}) \right] + n^{-1} \sum_{t=1}^{n} \log r_t(\boldsymbol{\beta}). \tag{3.119}$$

For the AR(1) model (3.105) discussed previously, recall that $x_1^0 = \mu$ and $x_t^{t-1} = \mu + \phi(x_{t-1} - \mu)$, for $t = 2, \dots, n$. Also, using the fact that $\phi_{11} = \phi$ and $\phi_{hh} = 0$ for $h > 1$, we have $r_1 = \sum_{j=0}^{\infty} \phi^{2j} = (1 - \phi^2)^{-1}$, $r_2 = (1 - \phi^2)^{-1}(1 - \phi^2) = 1$, and in general, $r_t = 1$ for $t = 2, \dots, n$. Hence, the likelihood presented in (3.106) is identical to the innovations form of the likelihood given by (3.116). Moreover, the generic $S(\boldsymbol{\beta})$ in (3.117) is $S(\mu, \phi)$ given in (3.107) and the generic $l(\boldsymbol{\beta})$ in (3.119) is $l(\mu, \phi)$ in (3.109).

Unconditional least squares would be performed by minimizing (3.117) with respect to $\boldsymbol{\beta}$. Conditional least squares estimation would involve minimizing (3.117) with respect to $\boldsymbol{\beta}$ but where, to ease the computational burden, the predictions and their errors are obtained by conditioning on initial values of the data. In general, numerical optimization routines are used to obtain the actual estimates and their standard errors.

### Example 3.29 The Newton–Raphson and Scoring Algorithms

Two common numerical optimization routines for accomplishing maximum likelihood estimation are Newton–Raphson and scoring. We will give a brief account of the mathematical ideas here. The actual implementation of these algorithms is much more complicated than our discussion might imply. For details, the reader is referred to any of the *Numerical Recipes* books, for example, Press et al. (1993).

Let $l(\boldsymbol{\beta})$ be a criterion function of $k$ parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)$ that we wish to minimize with respect to $\boldsymbol{\beta}$. For example, consider the likelihood function given by (3.109) or by (3.119). Suppose $l(\widehat{\boldsymbol{\beta}})$ is the extremum that we are interested in finding, and $\widehat{\boldsymbol{\beta}}$ is found by solving $\partial l(\boldsymbol{\beta})/\partial \beta_j = 0$, for $j = 1, \dots, k$. Let $l^{(1)}(\boldsymbol{\beta})$ denote the $k \times 1$ vector of partials

$$l^{(1)}(\boldsymbol{\beta}) = \left( \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1}, \dots, \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_k} \right)'.$$

Note, $l^{(1)}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, the $k \times 1$ zero vector. Let $l^{(2)}(\boldsymbol{\beta})$ denote the $k \times k$ matrix of second-order partials

$$l^{(2)}(\boldsymbol{\beta}) = \left\{ -\frac{\partial l^2(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \right\}_{i,j=1}^{k},$$

and assume $l^{(2)}(\boldsymbol{\beta})$ is nonsingular. Let $\boldsymbol{\beta}_{(0)}$ be an initial estimator of $\boldsymbol{\beta}$. Then, using a Taylor expansion, we have the following approximation:

$$\mathbf{0} = l^{(1)}(\widehat{\boldsymbol{\beta}}) \approx l^{(1)}(\boldsymbol{\beta}_{(0)}) - l^{(2)}(\boldsymbol{\beta}_{(0)}) \left[ \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{(0)} \right].$$

Setting the right-hand side equal to zero and solving for $\widehat{\boldsymbol{\beta}}$ [call the solution $\boldsymbol{\beta}_{(1)}$], we get

$$\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)} + \left[ l^{(2)}(\boldsymbol{\beta}_{(0)}) \right]^{-1} l^{(1)}(\boldsymbol{\beta}_{(0)}).$$

The Newton–Raphson algorithm proceeds by iterating this result, replacing $\boldsymbol{\beta}_{(0)}$ by $\boldsymbol{\beta}_{(1)}$ to get $\boldsymbol{\beta}_{(2)}$, and so on, until convergence. Under a set of appropriate conditions, the sequence of estimators, $\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}, \dots$, will converge to $\widehat{\boldsymbol{\beta}}$, the MLE of $\boldsymbol{\beta}$.

For maximum likelihood estimation, the criterion function used is $l(\boldsymbol{\beta})$ given by (3.119); $l^{(1)}(\boldsymbol{\beta})$ is called the score vector, and $l^{(2)}(\boldsymbol{\beta})$ is called the Hessian. In the method of scoring, we replace $l^{(2)}(\boldsymbol{\beta})$ by $E[l^{(2)}(\boldsymbol{\beta})]$, the information matrix. Under appropriate conditions, the inverse of the information matrix is the asymptotic variance–covariance matrix of the estimator $\widehat{\boldsymbol{\beta}}$. This is sometimes approximated by the inverse of the Hessian at $\widehat{\boldsymbol{\beta}}$. If the derivatives are difficult to obtain, it is possible to use quasi-maximum likelihood estimation where numerical techniques are used to approximate the derivatives.

**Example 3.30 MLE for the Recruitment Series**

So far, we have fit an AR(2) model to the Recruitment series using ordinary least squares (Example 3.17) and using Yule–Walker (Example 3.27). The following is an R session used to fit an AR(2) model via maximum likelihood estimation to the Recruitment series; these results can be compared to the results in Examples 3.17 and 3.27.

```
rec.mle = ar.mle(rec, order=2)
rec.mle$x.mean   # 62.26
rec.mle$ar       # 1.35, -.46
sqrt(diag(rec.mle$asy.var.coef))  # .04, .04
rec.mle$var.pred     # 89.34
```

We now discuss least squares for ARMA$(p, q)$ models via Gauss–Newton. For general and complete details of the Gauss–Newton procedure, the reader is referred to Fuller (1996). As before, write $\boldsymbol{\beta} = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)'$, and for the ease of discussion, we will put $\mu = 0$. We write the model in terms of the errors

$$w_t(\boldsymbol{\beta}) = x_t - \sum_{j=1}^{p} \phi_j x_{t-j} - \sum_{k=1}^{q} \theta_k w_{t-k}(\boldsymbol{\beta}), \tag{3.120}$$

emphasizing the dependence of the errors on the parameters.

For conditional least squares, we approximate the residual sum of squares by conditioning on $x_1, \dots, x_p$ (if $p > 0$) and $w_p = w_{p-1} = w_{p-2} = \cdots = w_{1-q} = 0$ (if $q > 0$), in which case, given $\boldsymbol{\beta}$, we may evaluate (3.120) for $t = p+1, p+2, \dots, n$. Using this conditioning argument, the conditional error sum of squares is

$$S_c(\boldsymbol{\beta}) = \sum_{t=p+1}^{n} w_t^2(\boldsymbol{\beta}). \tag{3.121}$$

Minimizing $S_c(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ yields the conditional least squares estimates. If $q = 0$, the problem is linear regression and no iterative technique is needed to minimize $S_c(\phi_1, \dots, \phi_p)$. If $q > 0$, the problem becomes nonlinear regression and we will have to rely on numerical optimization.

When $n$ is large, conditioning on a few initial values will have little influence on the final parameter estimates. In the case of small to moderate sample sizes, one may wish to rely on unconditional least squares. The unconditional least squares problem is to choose $\boldsymbol{\beta}$ to minimize the unconditional sum of squares, which we have generically denoted by $S(\boldsymbol{\beta})$ in this section. The unconditional sum of squares can be written in various ways, and one useful form in the case of ARMA$(p, q)$ models is derived in Box et al. (1994, Appendix A7.3). They showed (see Problem 3.19) the unconditional sum of squares can be written as

$$S(\boldsymbol{\beta}) = \sum_{t=-\infty}^{n} \widehat{w}_t^2(\boldsymbol{\beta}), \tag{3.122}$$

where $\widehat{w}_t(\boldsymbol{\beta}) = E(w_t \,|\, x_1, \dots, x_n)$. When $t \le 0$, the $\widehat{w}_t(\boldsymbol{\beta})$ are obtained by backcasting. As a practical matter, we approximate $S(\boldsymbol{\beta})$ by starting the sum at $t = -M + 1$, where $M$ is chosen large enough to guarantee $\sum_{t=-\infty}^{-M} \widehat{w}_t^2(\boldsymbol{\beta}) \approx 0$. In the case of unconditional least squares estimation, a numerical optimization technique is needed even when $q = 0$.

To employ Gauss–Newton, let $\boldsymbol{\beta}_{(0)} = (\phi_1^{(0)}, \dots, \phi_p^{(0)}, \theta_1^{(0)}, \dots, \theta_q^{(0)})'$ be an initial estimate of $\boldsymbol{\beta}$. For example, we could obtain $\boldsymbol{\beta}_{(0)}$ by method of moments. The first-order Taylor expansion of $w_t(\boldsymbol{\beta})$ is

$$w_t(\boldsymbol{\beta}) \approx w_t(\boldsymbol{\beta}_{(0)}) - \left(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}\right)' \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}), \tag{3.123}$$

where

$$\boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}) = \left(-\frac{\partial w_t(\boldsymbol{\beta}_{(0)})}{\partial \beta_1}, \dots, -\frac{\partial w_t(\boldsymbol{\beta}_{(0)})}{\partial \beta_{p+q}}\right)', \quad t = 1, \dots, n.$$

The linear approximation of $S_c(\boldsymbol{\beta})$ is

$$Q(\boldsymbol{\beta}) = \sum_{t=p+1}^{n} \left[w_t(\boldsymbol{\beta}_{(0)}) - \left(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}\right)' \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)})\right]^2 \tag{3.124}$$

and this is the quantity that we will minimize. For approximate unconditional least squares, we would start the sum in (3.124) at $t = -M + 1$, for a large value of $M$, and work with the backcasted values.

Using the results of ordinary least squares (§2.2), we know

$$(\widehat{\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}}) = \left(n^{-1} \sum_{t=p+1}^{n} \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}) \boldsymbol{z}_t'(\boldsymbol{\beta}_{(0)})\right)^{-1} \left(n^{-1} \sum_{t=p+1}^{n} \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}) w_t(\boldsymbol{\beta}_{(0)})\right) \tag{3.125}$$

minimizes $Q(\boldsymbol{\beta})$. From (3.125), we write the one-step Gauss–Newton estimate as

$$\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)} + \Delta(\boldsymbol{\beta}_{(0)}), \tag{3.126}$$

where $\Delta(\boldsymbol{\beta}_{(0)})$ denotes the right-hand side of (3.125). Gauss–Newton estimation is accomplished by replacing $\boldsymbol{\beta}_{(0)}$ by $\boldsymbol{\beta}_{(1)}$ in (3.126). This process is repeated by calculating, at iteration $j = 2, 3, \dots,$

$$\boldsymbol{\beta}_{(j)} = \boldsymbol{\beta}_{(j-1)} + \Delta(\boldsymbol{\beta}_{(j-1)})$$

until convergence.

**Example 3.31 Gauss–Newton for an MA(1)**

Consider an invertible MA(1) process, $x_t = w_t + \theta w_{t-1}$. Write the truncated errors as

$$w_t(\theta) = x_t - \theta w_{t-1}(\theta), \quad t = 1, \dots, n, \tag{3.127}$$

where we condition on $w_0(\theta) = 0$. Taking derivatives,

$$-\frac{\partial w_t(\theta)}{\partial \theta} = w_{t-1}(\theta) + \theta \frac{\partial w_{t-1}(\theta)}{\partial \theta}, \quad t = 1, \dots, n, \tag{3.128}$$

where $\partial w_0(\theta)/\partial \theta = 0$. Using the notation of (3.123), we can also write (3.128) as

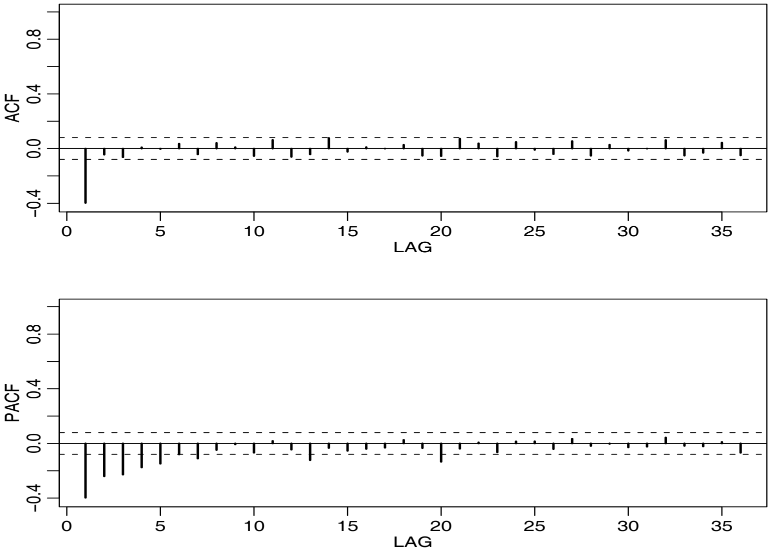

**Fig. 3.7.** ACF and PACF of transformed glacial varves.

$$z_t(\theta) = w_{t-1}(\theta) - \theta z_{t-1}(\theta), \quad t = 1, \dots, n, \tag{3.129}$$

where $z_0(\theta) = 0$.

Let $\theta_{(0)}$ be an initial estimate of $\theta$, for example, the estimate given in Example 3.28. Then, the Gauss–Newton procedure for conditional least squares is given by

$$\theta_{(j+1)} = \theta_{(j)} + \frac{\sum_{t=1}^{n} z_t(\theta_{(j)}) w_t(\theta_{(j)})}{\sum_{t=1}^{n} z_t^2(\theta_{(j)})}, \quad j = 0, 1, 2, \dots, \tag{3.130}$$

where the values in (3.130) are calculated recursively using (3.127) and (3.129). 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 3.32 Fitting the Glacial Varve Series

Consider the series of glacial varve thicknesses from Massachusetts for $n = 634$ years, as analyzed in Example 2.6 and in Problem 2.8, where it was argued that a first-order moving average model might fit the logarithmically transformed and differenced varve series, say,

$$\nabla \log(x_t) = \log(x_t) - \log(x_{t-1}) = \log\left(\frac{x_t}{x_{t-1}}\right),$$

which can be interpreted as being approximately the percentage change in the thickness.

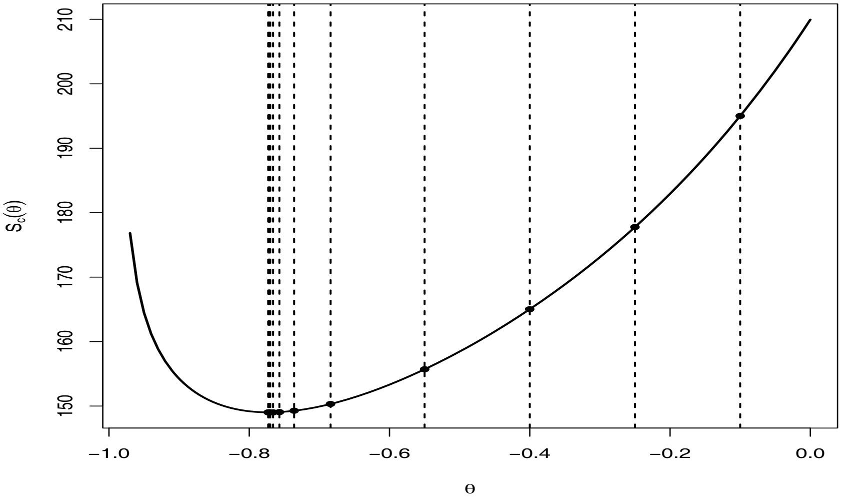

**Fig. 3.8.** Conditional sum of squares versus values of the moving average parameter for the glacial varve example, Example 3.32. Vertical lines indicate the values of the parameter obtained via Gauss–Newton; see Table 3.2 for the actual values.

The sample ACF and PACF, shown in Figure 3.7, confirm the tendency of $\nabla\log(x_t)$ to behave as a first-order moving average process as the ACF has only a significant peak at lag one and the PACF decreases exponentially. Using Table 3.1, this sample behavior fits that of the MA(1) very well.

The results of eleven iterations of the Gauss–Newton procedure, (3.130), starting with $\theta_{(0)} = -.10$ are given in Table 3.2. The final estimate is $\widehat{\theta} = \theta_{(11)} = -.773$; interim values and the corresponding value of the conditional sum of squares, $S_c(\theta)$ given in (3.121), are also displayed in the table. The final estimate of the error variance is $\widehat{\sigma}^2_w = 148.98/632 = .236$ with 632 degrees of freedom (one is lost in differencing). The value of the sum of the squared derivatives at convergence is $\sum_{t=1}^n z_t^2(\theta_{(11)}) = 369.73$, and consequently, the estimated standard error of $\widehat{\theta}$ is $\sqrt{.236/369.73} = .025$;[7] this leads to a $t$-value of $-.773/.025 = -30.92$ with 632 degrees of freedom.

Figure 3.8 displays the conditional sum of squares, $S_c(\theta)$ as a function of $\theta$, 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. When there is only one parameter, as in this case, it would be easy to evaluate $S_c(\theta)$ on a grid of points, and then choose the appropriate value of $\theta$ from the grid search. It would be difficult, however, to perform grid searches when there are many parameters.

[7] To estimate the standard error, we are using the standard regression results from (2.9) as an approximation

**Table 3.2.** Gauss–Newton Results for Example 3.32

| $j$ | $\theta_{(j)}$ | $S_c(\theta_{(j)})$ | $\sum_{t=1}^n z_t^2(\theta_{(j)})$ |
|---|---|---|---|
| 0 | −0.100 | 195.0010 | 183.3464 |
| 1 | −0.250 | 177.7614 | 163.3038 |
| 2 | −0.400 | 165.0027 | 161.6279 |
| 3 | −0.550 | 155.6723 | 182.6432 |
| 4 | −0.684 | 150.2896 | 247.4942 |
| 5 | −0.736 | 149.2283 | 304.3125 |
| 6 | −0.757 | 149.0272 | 337.9200 |
| 7 | −0.766 | 148.9885 | 355.0465 |
| 8 | −0.770 | 148.9812 | 363.2813 |
| 9 | −0.771 | 148.9804 | 365.4045 |
| 10 | −0.772 | 148.9799 | 367.5544 |
| 11 | −0.773 | 148.9799 | 369.7314 |

In the general case of causal and invertible ARMA$(p, q)$ models, maximum likelihood estimation and conditional and unconditional least squares estimation (and Yule–Walker estimation in the case of AR models) all lead to optimal estimators. The proof of this general result can be found in a number of texts on theoretical time series analysis (for example, Brockwell and Davis, 1991, or Hannan, 1970, to mention a few). We will denote the ARMA coefficient parameters by $\boldsymbol{\beta} = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)'$.

**Property 3.10 Large Sample Distribution of the Estimators**

*Under appropriate conditions, for causal and invertible ARMA processes, the maximum likelihood, the unconditional least squares, and the conditional least squares estimators, each initialized by the method of moments estimator, all provide optimal estimators of $\sigma_w^2$ and $\boldsymbol{\beta}$, in the sense that $\widehat{\sigma}_w^2$ is consistent, and the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ is the best asymptotic normal distribution. In particular, as $n \to \infty$,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma_w^2\ \boldsymbol{\Gamma}_{p,q}^{-1}\right). \tag{3.131}$$

*The asymptotic variance–covariance matrix of the estimator $\widehat{\boldsymbol{\beta}}$ is the inverse of the information matrix. In particular, the $(p+q) \times (p+q)$ matrix $\Gamma_{p,q}$, has the form*

$$\Gamma_{p,q} = \begin{pmatrix} \Gamma_{\phi\phi} & \Gamma_{\phi\theta} \\ \Gamma_{\theta\phi} & \Gamma_{\theta\theta} \end{pmatrix}. \tag{3.132}$$

*The $p \times p$ matrix $\Gamma_{\phi\phi}$ is given by (3.100), that is, the $ij$-th element of $\Gamma_{\phi\phi}$, for $i, j = 1, \dots, p$, is $\gamma_x(i-j)$ from an AR(p) process, $\phi(B)x_t = w_t$. Similarly, $\Gamma_{\theta\theta}$ is a $q \times q$ matrix with the $ij$-th element, for $i, j = 1, \dots, q$, equal to $\gamma_y(i-j)$ from an AR(q) process, $\theta(B)y_t = w_t$. The $p \times q$ matrix $\Gamma_{\phi\theta} = \{\gamma_{xy}(i-j)\}$, for $i = 1, \dots, p$; $j = 1, \dots, q$; that is, the $ij$-th element is the cross-covariance* *between the two AR processes given by $\phi(B)x_t = w_t$ and $\theta(B)y_t = w_t$. Finally, $\Gamma_{\theta\phi} = \Gamma'_{\phi\theta}$ is $q \times p$.*

Further discussion of Property 3.10, including a proof for the case of least squares estimators for AR($p$) processes, can be found in Appendix B, §B.3.

**Example 3.33 Some Specific Asymptotic Distributions**

The following are some specific cases of Property 3.10.

**AR(1):** $\gamma_x(0) = \sigma_w^2/(1-\phi^2)$, so $\sigma_w^2\Gamma_{1,0}^{-1} = (1-\phi^2)$. Thus,

$$\widehat{\phi} \sim \text{AN}\left[\phi, n^{-1}(1-\phi^2)\right]. \tag{3.133}$$

**AR(2):** The reader can verify that

$$\gamma_x(0) = \left(\frac{1-\phi_2}{1+\phi_2}\right)\frac{\sigma_w^2}{(1-\phi_2)^2-\phi_1^2}$$

and $\gamma_x(1) = \phi_1\gamma_x(0) + \phi_2\gamma_x(1)$. From these facts, we can compute $\Gamma_{2,0}^{-1}$. In particular, we have

$$\begin{pmatrix}\widehat{\phi}_1\\ \widehat{\phi}_2\end{pmatrix} \sim \text{AN}\left[\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix},\ n^{-1}\begin{pmatrix}1-\phi_2^2 & -\phi_1(1+\phi_2)\\ \text{sym} & 1-\phi_2^2\end{pmatrix}\right]. \tag{3.134}$$

**MA(1):** In this case, write $\theta(B)y_t = w_t$, or $y_t + \theta y_{t-1} = w_t$. Then, analogous to the AR(1) case, $\gamma_y(0) = \sigma_w^2/(1-\theta^2)$, so $\sigma_w^2\Gamma_{0,1}^{-1} = (1-\theta^2)$. Thus,

$$\widehat{\theta} \sim \text{AN}\left[\theta, n^{-1}(1-\theta^2)\right]. \tag{3.135}$$

**MA(2):** Write $y_t + \theta_1 y_{t-1} + \theta_2 y_{t-2} = w_t$, so , analogous to the AR(2) case, we have

$$\begin{pmatrix}\widehat{\theta}_1\\ \widehat{\theta}_2\end{pmatrix} \sim \text{AN}\left[\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix},\ n^{-1}\begin{pmatrix}1-\theta_2^2 & \theta_1(1+\theta_2)\\ \text{sym} & 1-\theta_2^2\end{pmatrix}\right]. \tag{3.136}$$

**ARMA(1,1):** To calculate $\Gamma_{\phi\theta}$, we must find $\gamma_{xy}(0)$, where $x_t - \phi x_{t-1} = w_t$ and $y_t + \theta y_{t-1} = w_t$. We have

$$\begin{aligned}\gamma_{xy}(0) &= \text{cov}(x_t, y_t) = \text{cov}(\phi x_{t-1} + w_t, -\theta y_{t-1} + w_t)\\ &= -\phi\theta\gamma_{xy}(0) + \sigma_w^2.\end{aligned}$$

Solving, we find, $\gamma_{xy}(0) = \sigma_w^2/(1+\phi\theta)$. Thus,

$$\begin{pmatrix}\widehat{\phi}\\ \widehat{\theta}\end{pmatrix} \sim \text{AN}\left[\begin{pmatrix}\phi\\ \theta\end{pmatrix},\ n^{-1}\begin{bmatrix}(1-\phi^2)^{-1} & (1+\phi\theta)^{-1}\\ \text{sym} & (1-\theta^2)^{-1}\end{bmatrix}^{-1}\right]. \tag{3.137}$$

**Example 3.34 Overfitting Caveat**

The asymptotic 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 $n$, $\text{var}(\widehat{\phi}_1) \approx n^{-1}(1-\phi_1^2)$. But, if we fit an AR(2) to the AR(1) process, for large $n$, $\text{var}(\widehat{\phi}_1) \approx n^{-1}(1-\phi_2^2) = n^{-1}$ because $\phi_2 = 0$. Thus, the variance of $\phi_1$ 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(2) model to the data and are satisfied with that model, then adding one more parameter and fitting an AR(3) should lead to approximately the same model as in the AR(2) fit. We will discuss model diagnostics in more detail in §3.8.

The reader might wonder, for example, why the asymptotic distributions of $\widehat{\phi}$ from an AR(1) and $\widehat{\theta}$ from an MA(1) are of the same form; compare (3.133) to (3.135). It is possible to explain this unexpected result heuristically using the intuition of linear regression. That is, for the normal regression model presented in §2.2 with no intercept term, $x_t = \beta z_t + w_t$, we know $\widehat{\beta}$ is normally distributed with mean $\beta$, and from (2.9),

$$\text{var}\left\{\sqrt{n}\left(\widehat{\beta}-\beta\right)\right\} = n\sigma_w^2\left(\sum_{t=1}^{n} z_t^2\right)^{-1} = \sigma_w^2\left(n^{-1}\sum_{t=1}^{n} z_t^2\right)^{-1}.$$

For the causal AR(1) model given by $x_t = \phi x_{t-1} + w_t$, the intuition of regression tells us to expect that, for $n$ large,

$$\sqrt{n}\left(\widehat{\phi}-\phi\right)$$

is approximately normal with mean zero and with variance given by

$$\sigma_w^2\left(n^{-1}\sum_{t=2}^{n} x_{t-1}^2\right)^{-1}.$$

Now, $n^{-1}\sum_{t=2}^{n} x_{t-1}^2$ is the sample variance (recall that the mean of $x_t$ is zero) of the $x_t$, so as $n$ becomes large we would expect it to approach $\text{var}(x_t) = \gamma(0) = \sigma_w^2/(1-\phi^2)$. Thus, the large sample variance of $\sqrt{n}\left(\widehat{\phi}-\phi\right)$ is

$$\sigma_w^2\gamma_x(0)^{-1} = \sigma_w^2\left(\frac{\sigma_w^2}{1-\phi^2}\right)^{-1} = (1-\phi^2);$$

that is, (3.133) holds.

In the case of an MA(1), we may use the discussion of Example 3.31 to write an approximate regression model for the MA(1). That is, consider the approximation (3.129) as the regression model

$$z_t(\widehat{\theta}) = -\theta z_{t-1}(\widehat{\theta}) + w_{t-1},$$

where now, $z_{t-1}(\widehat{\theta})$ as defined in Example 3.31, plays the role of the regressor. Continuing with the analogy, we would expect the asymptotic distribution of $\sqrt{n}\left(\widehat{\theta} - \theta\right)$ to be normal, with mean zero, and approximate variance

$$\sigma_w^2 \left( n^{-1} \sum_{t=2}^{n} z_{t-1}^2(\widehat{\theta}) \right)^{-1} .$$

As in the AR(1) case, $n^{-1}\sum_{t=2}^{n} z_{t-1}^2(\widehat{\theta})$ is the sample variance of the $z_t(\widehat{\theta})$ so, for large $n$, this should be $\text{var}\{z_t(\theta)\} = \gamma_z(0)$, say. But note, as seen from (3.129), $z_t(\theta)$ is approximately an AR(1) process with parameter $-\theta$. Thus,

$$\sigma_w^2 \gamma_z(0)^{-1} = \sigma_w^2 \left( \frac{\sigma_w^2}{1-(-\theta)^2} \right)^{-1} = (1-\theta^2),$$

which agrees with (3.135). Finally, the asymptotic distributions of the AR parameter estimates and the MA parameter estimates are of the same form because in the MA case, the "regressors" are the differential processes $z_t(\theta)$ that have AR structure, and it is this structure that determines the asymptotic variance of the estimators. For a rigorous account of this approach for the general case, see Fuller (1996, Theorem 5.5.4).

In Example 3.32, the estimated standard error of $\widehat{\theta}$ was .025. In that example, we used regression results to estimate the standard error as the square root of

$$n^{-1}\widehat{\sigma}_w^2 \left( n^{-1} \sum_{t=1}^{n} z_t^2(\widehat{\theta}) \right)^{-1} = \frac{\widehat{\sigma}_w^2}{\sum_{t=1}^{n} z_t^2(\widehat{\theta})},$$

where $n = 632$, $\widehat{\sigma}_w^2 = .236$, $\sum_{t=1}^{n} z_t^2(\widehat{\theta}) = 369.73$ and $\widehat{\theta} = -.773$. Using (3.135), we could have also calculated this value using the asymptotic approximation, the square root of $(1-(-.773)^2)/632$, which is also .025.

If $n$ is small, or if the parameters are close to the boundaries, the asymptotic approximations can be quite poor. The bootstrap can be helpful in this case; for a broad treatment of the bootstrap, see Efron and Tibshirani (1994). We discuss the case of an AR(1) here and leave the general discussion for Chapter 6. For now, we give a simple example of the bootstrap for an AR(1) process.

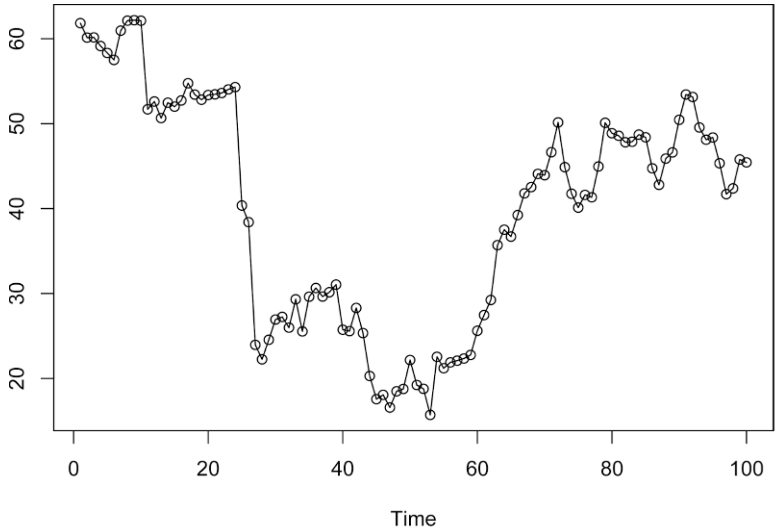

**Fig. 3.9.** One hundred observations generated from the model in Example 3.35.

### Example 3.35 Bootstrapping an AR(1)

We consider an AR(1) model with a regression coefficient near the boundary of causality and an error process that is symmetric but not normal. Specifically, consider the causal model

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t, \tag{3.138}$$

where $\mu = 50$, $\phi = .95$, and $w_t$ are iid double exponential with location zero, and scale parameter $\beta = 2$. The density of $w_t$ is given by

$$f(w) = \frac{1}{2\beta} \exp\{-|w|/\beta\} \qquad -\infty < w < \infty.$$

In this example, $E(w_t) = 0$ and $\text{var}(w_t) = 2\beta^2 = 8$. Figure 3.9 shows $n = 100$ simulated observations from this process. This particular realization is interesting; the data look like they were generated from a nonstationary process with three different mean levels. In fact, the data were generated from a well-behaved, albeit non-normal, stationary and causal model. To show the advantages of the bootstrap, we will act as if we do not know the actual error distribution and we will proceed as if it were normal; of course, this means, for example, that the normal based MLE of $\phi$ will not be the actual MLE because the data are not normal.

Using the data shown in Figure 3.9, we obtained the Yule–Walker estimates $\widehat{\mu} = 40.05$, $\widehat{\phi} = .96$, and $s_w^2 = 15.30$, where $s_w^2$ is the estimate of $\text{var}(w_t)$. Based on Property 3.10, we would say that $\widehat{\phi}$ is approximately normal with mean $\phi$ (which we supposedly do not know) and variance $(1 - \phi^2)/100$, which we would approximate by $(1 - .96^2)/100 = .03^2$.

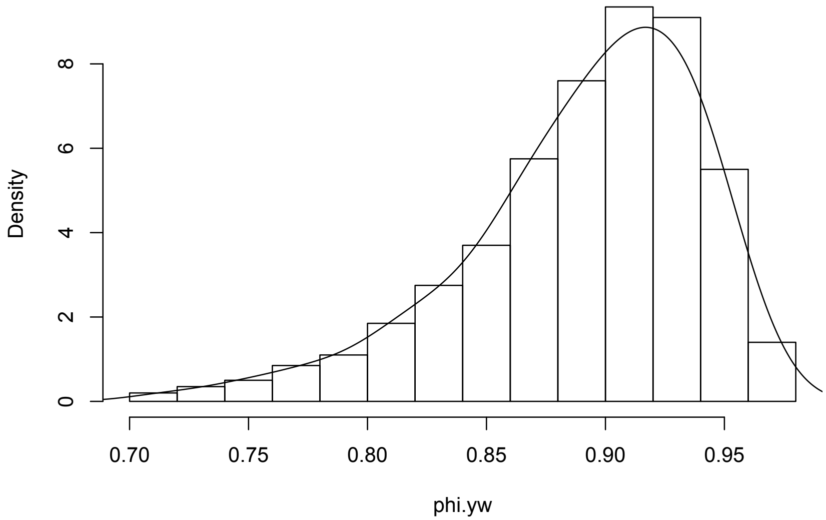

**Fig. 3.10.** Finite sample density of the Yule–Walker estimate of $\phi$ in Example 3.35.

To assess the finite sample distribution of $\widehat{\phi}$ when $n = 100$, we simulated 1000 realizations of this AR(1) process and estimated the parameters via Yule–Walker. The finite sampling density of the Yule–Walker estimate of $\phi$, based on the 1000 repeated simulations, is shown in Figure 3.10. Clearly the sampling distribution is not close to normality for this sample size. The mean of the distribution shown in Figure 3.10 is .89, and the variance of the distribution is $.05^2$; these values are considerably different than the asymptotic values. Some of the quantiles of the finite sample distribution are .79 (5%), .86 (25%), .90 (50%), .93 (75%), and .95 (95%). The R code to perform the simulation and plot the histogram is as follows:

```
set.seed(111)
phi.yw = rep(NA, 1000)
for (i in 1:1000){
  e = rexp(150, rate=.5); u = runif(150,-1,1); de = e*sign(u)
  x = 50 + arima.sim(n=100,list(ar=.95), innov=de, n.start=50)
  phi.yw[i] = ar.yw(x, order=1)$ar }
hist(phi.yw, prob=TRUE, main="")
lines(density(phi.yw, bw=.015))
```

Before discussing the bootstrap, we first investigate the sample innovation process, $x_t - x_t^{t-1}$, with corresponding variances $P_t^{t-1}$. For the AR(1) model in this example,

$$x_t^{t-1} = \mu + \phi(x_{t-1} - \mu), \qquad t = 2, \dots, 100.$$

From this, it follows that

$$P_t^{t-1} = E(x_t - x_t^{t-1})^2 = \sigma_w^2, \qquad t = 2, \dots, 100.$$

When $t = 1$, we have

$$x_1^0 = \mu \quad \text{and} \quad P_1^0 = \sigma_w^2/(1 - \phi^2).$$

Thus, the innovations have zero mean but different variances; in order that all of the innovations have the same variance, $\sigma_w^2$, we will write them as

$$\begin{aligned} \epsilon_1 &= (x_1 - \mu)\sqrt{(1 - \phi^2)} \\ \epsilon_t &= (x_t - \mu) - \phi(x_{t-1} - \mu), \quad \text{for} \quad t = 2, \dots, 100. \end{aligned} \tag{3.139}$$

From these equations, we can write the model in terms of the $\epsilon_t$ as

$$\begin{aligned} x_1 &= \mu + \epsilon_1/\sqrt{(1 - \phi^2)} \\ x_t &= \mu + \phi(x_{t-1} - \mu) + \epsilon_t \quad \text{for} \quad t = 2, \dots, 100. \end{aligned} \tag{3.140}$$

Next, replace the parameters with their estimates in (3.139), that is, $\widehat{\mu} = 40.048$ and $\widehat{\phi} = .957$, and denote the resulting sample innovations as $\{\widehat{\epsilon}_1, \dots, \widehat{\epsilon}_{100}\}$. To obtain one bootstrap sample, first randomly sample, with replacement, $n = 100$ values from the set of sample innovations; call the sampled values $\{\epsilon_1^*, \dots, \epsilon_{100}^*\}$. Now, generate a bootstrapped data set sequentially by setting

$$\begin{aligned} x_1^* &= 40.048 + \epsilon_1^*/\sqrt{(1 - .957^2)} \\ x_t^* &= 40.048 + .957(x_{t-1}^* - 40.048) + \epsilon_t^*, \quad t = 2, \dots, n. \end{aligned} \tag{3.141}$$

Next, estimate the parameters as if the data were $x_t^*$. Call these estimates $\widehat{\mu}(1)$, $\widehat{\phi}(1)$, and $s_w^2(1)$. Repeat this process a large number, $B$, of times, generating a collection of bootstrapped parameter estimates, $\{\widehat{\mu}(b), \widehat{\phi}(b), s_w^2(b), b = 1, \dots, B\}$. We can then approximate the finite sample distribution of an estimator from the bootstrapped parameter values. For example, we can approximate the distribution of $\widehat{\phi} - \phi$ by the empirical distribution of $\widehat{\phi}(b) - \widehat{\phi}$, for $b = 1, \dots, B$.

Figure 3.11 shows the bootstrap histogram of 200 bootstrapped estimates of $\phi$ using the data shown in Figure 3.9. In addition, Figure 3.11 shows a density estimate based on the bootstrap histogram, as well as the asymptotic normal density that would have been used based on Proposition 3.10. Clearly, the bootstrap distribution of $\widehat{\phi}$ is closer to the distribution of $\widehat{\phi}$ shown in Figure 3.10 than to the asymptotic normal approximation. In particular, the mean of the distribution of $\widehat{\phi}(b)$ is .92 with a variance of $.05^2$. Some quantiles of this distribution are .83 (5%), .90 (25%), .93 (50%), .95 (75%), and .98 (95%).

To perform a similar bootstrap exercise in R, use the following commands. We note that the R estimation procedure is conditional on the first observation, so the first residual is not returned. To get around this problem,

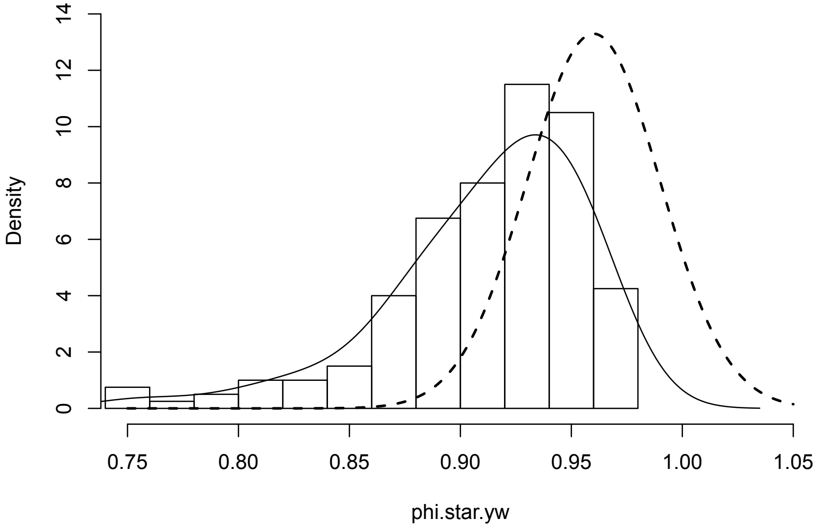

**Fig. 3.11.** Bootstrap histogram of $\widehat{\phi}$ based on 200 bootstraps; a density estimate based on the histogram (solid line) and the corresponding asymptotic normal density (dashed line).

we simply fix the first observation and bootstrap the remaining data. The simulated data are available in the file `ar1boot`, but you can simulate your own data as was done in the code that produced Figure 3.10.

```
x = ar1boot
m = mean(x)   # estimate of mu
fit = ar.yw(x, order=1)
phi = fit$ar  # estimate of phi
nboot = 200   # number of bootstrap replicates
resids = fit$resid[-1]  # the first resid is NA
x.star = x    # initialize x*
phi.star.yw = rep(NA, nboot)
for (i in 1:nboot) {
  resid.star = sample(resids, replace=TRUE)
  for (t in 1:99){ x.star[t+1] = m + phi*(x.star[t]-m) +
     resid.star[t] }
  phi.star.yw[i] = ar.yw(x.star, order=1)$ar }
hist(phi.star.yw, 10, main="", prob=TRUE, ylim=c(0,14),
     xlim=c(.75,1.05))
lines(density(phi.star.yw, bw=.02))
u = seq(.75, 1.05, by=.001)
lines(u, dnorm(u, mean=.96, sd=.03), lty="dashed", lwd=2)
```

## 3.7 Integrated Models for Nonstationary Data

In Chapters 1 and 2, we saw that if $x_t$ is a random walk, $x_t = x_{t-1} + w_t$, then by differencing $x_t$, we find that $\nabla x_t = w_t$ is stationary. In many situations, time series can be thought of as being composed of two components, a nonstationary trend component and a zero-mean stationary component. For example, in §2.2 we considered the model

$$x_t = \mu_t + y_t, \tag{3.142}$$

where $\mu_t = \beta_0 + \beta_1 t$ and $y_t$ is stationary. Differencing such a process will lead to a stationary process:

$$\nabla x_t = x_t - x_{t-1} = \beta_1 + y_t - y_{t-1} = \beta_1 + \nabla y_t.$$

Another model that leads to first differencing is the case in which $\mu_t$ in (3.142) is stochastic and slowly varying according to a random walk. That is,

$$\mu_t = \mu_{t-1} + v_t$$

where $v_t$ is stationary. In this case,

$$\nabla x_t = v_t + \nabla y_t,$$

is stationary. If $\mu_t$ in (3.142) is a $k$-th order polynomial, $\mu_t = \sum_{j=0}^{k} \beta_j t^j$, then (Problem 3.27) the differenced series $\nabla^k y_t$ is stationary. Stochastic trend models can also lead to higher order differencing. For example, suppose

$$\mu_t = \mu_{t-1} + v_t \quad \text{and} \quad v_t = v_{t-1} + e_t,$$

where $e_t$ is stationary. Then, $\nabla x_t = v_t + \nabla y_t$ is not stationary, but

$$\nabla^2 x_t = e_t + \nabla^2 y_t$$

is stationary.

The integrated ARMA, or ARIMA, model is a broadening of the class of ARMA models to include differencing.

**Definition 3.11** *A process $x_t$ is said to be* **ARIMA**$(p, d, q)$ *if*

$$\nabla^d x_t = (1 - B)^d x_t$$

*is ARMA$(p, q)$. In general, we will write the model as*

$$\phi(B)(1 - B)^d x_t = \theta(B) w_t. \tag{3.143}$$

*If $E(\nabla^d x_t) = \mu$, we write the model as*

$$\phi(B)(1 - B)^d x_t = \delta + \theta(B) w_t,$$

*where $\delta = \mu(1 - \phi_1 - \cdots - \phi_p)$.*