Estimation

Throughout this section, we assume we have n observations, $\boxed{x_1, \dots, x_n}$ 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, $\boxed{\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q,}$ $\boxed{\text{and } \sigma_w^2.}$ 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 $\mathrm{E}(x_t) = \mu$, then the method of moments estimator of μ 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,

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

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

Method of moments estimators:

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

$$\Gamma_p \phi = \gamma_p, \quad \sigma_w^2 = \gamma(0) - \phi' \gamma_p, \tag{3.100}$$

where $\Gamma_p = \{\gamma(k-j)\}_{j,k=1}^p$ is a $p \times p$ matrix, $\phi = (\phi_1, \dots, \phi_p)'$ is a $p \times 1$ vector, and $\gamma_p = (\gamma(1), \dots, \gamma(p))'$ is a $p \times 1$ vector. Using the method of moments, we replace $\gamma(h)$ in (3.100) by $\hat{\gamma}(h)$ [see equation (1.36)] and solve

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

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 $\hat{\gamma}(0)$ in (3.101), we can write the Yule–Walker estimates as

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

where $\hat{R}_p = \{\hat{\rho}(k-j)\}_{j,k=1}^p$ is a $p \times p$ matrix and $\hat{\rho}_p = (\hat{\rho}(1), \dots, \hat{\rho}(p))'$ 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 $\hat{\sigma}_w^2$ is close to the true value of σ_w^2. We state these results in Property 3.8; for details, see Section 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:

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

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

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

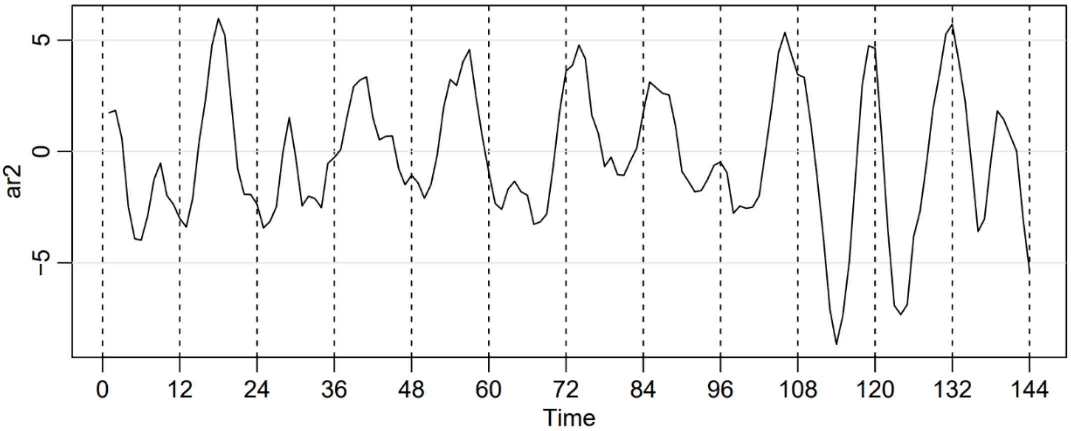

Fig. 3.4. *Simulated AR(2) model, $n = 144$ with $\phi_1 = 1.5$ and $\phi_2 = -.75$.*

Example 3.27 Yule–Walker Estimation for an AR(2) Process

The data shown in Figure 3.4 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, $\hat{\gamma}(0) = 8.903$, $\hat{\rho}(1) = .849$, and $\hat{\rho}(2) = .519$. Thus,

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

and

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

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

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

and it can be used to get confidence regions for, or make inferences about $\hat{\phi}$ and its components. For example, an approximate 95% confidence interval for ϕ_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 $\hat{\phi}_{11} = \hat{\rho}(1) = .849$, $\hat{\phi}_{22} = \hat{\phi}_2 = -.721$, and $\hat{\phi}_{33} = -.085$. According to Property 3.9, the asymptotic standard error of $\hat{\phi}_{33}$ is $1/\sqrt{144} = .083$, and the observed value, $-.085$, is about only one standard deviation from $\phi_{33} = 0$.

Example 3.28 Yule–Walker Estimation of the Recruitment Series

In Example 3.18 we fit an AR(2) model to the recruitment series using ordinary least squares (OLS). For AR models, the estimators obtained via OLS and Yule-Walker are nearly identical; we will see this when we discuss conditional sum of squares estimation in (3.111)–(3.116).

Below are the results of fitting the same model using Yule-Walker estimation in R, which are nearly identical to the values in Example 3.18.

```
rec.yw = ar.yw(rec, order=2)
rec.yw$x.mean    # = 62.26 (mean estimate)
rec.yw$ar        # = 1.33, -.44  (coefficient 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 (not shown) as in Example 3.25, use the R commands:

```
rec.pr = predict(rec.yw, n.ahead=24)
ts.plot(rec, rec.pr$pred, col=1:2)
lines(rec.pr$pred + rec.pr$se, col=4, lty=2)
lines(rec.pr$pred - rec.pr$se, col=4, lty=2)
```

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 3.29 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 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 $|\hat{\rho}(1)| \le \frac{1}{2}$, 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 $|\hat{\rho}(1)| \ge \frac{1}{2}$ because it is an estimator. For example, the following simulation in R produces a value of $\hat{\rho}(1) = .507$ 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 $|\hat{\rho}(1)| < \frac{1}{2}$, the invertible estimate is

$$\hat{\theta} = \frac{1 - \sqrt{1 - 4\hat{\rho}(1)^2}}{2\hat{\rho}(1)}. \tag{3.105}$$

It can be shown that[3.5]

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

AN is read *asymptotically normal* and is defined in Definition A.5. The maximum likelihood estimator (which we discuss next) of θ, 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 θ 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 θ when $\theta = .5$.

Maximum Likelihood and Least Squares Estimation

To fix ideas, we first focus on the causal $\boxed{\text{AR}(1)}$ case. Let

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t \tag{3.106}$$

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

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

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 $\boxed{x_t \mid x_{t-1} \sim \mathrm{N}\left(\mu + \phi(x_{t-1} - \mu), \, \sigma_w^2\right),}$ 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 σ_w^2. We may then write the likelihood as

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

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 μ and variance $\sigma_w^2/(1-\phi^2)$. Finally, for an AR(1), the likelihood is

$$L(\mu, \phi, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2}(1-\phi^2)^{1/2} \exp\left[-\frac{S(\mu, \phi)}{2\sigma_w^2}\right], \tag{3.107}$$

where

$$S(\mu, \phi) = (1-\phi^2)(x_1-\mu)^2 + \sum_{t=2}^{n} [(x_t-\mu) - \phi(x_{t-1}-\mu)]^2 . \tag{3.108}$$

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

Taking the partial derivative of the log of (3.107) with respect to σ_w^2 and setting the result equal to zero, we get the typical normal result that for any given values of μ and ϕ in the parameter space, $\sigma_w^2 = n^{-1}S(\mu, \phi)$ maximizes the likelihood. Thus, the maximum likelihood estimate of σ_w^2 is

$$\hat{\sigma}_w^2 = n^{-1}S(\hat{\mu}, \hat{\phi}), \tag{3.109}$$

where $\hat{\mu}$ and $\hat{\phi}$ are the MLEs of μ and ϕ, respectively. If we replace n in (3.109) by $n-2$, we would obtain the unconditional least squares estimate of σ_w^2.

If, in (3.107), we take logs, replace σ_w^2 by $\hat{\sigma}_w^2$, and ignore constants, $\hat{\mu}$ and $\hat{\phi}$ are the values that minimize the criterion function

$$l(\mu, \phi) = \log\left[n^{-1}S(\mu, \phi)\right] - n^{-1}\log(1-\phi^2); \tag{3.110}$$

that is, $l(\mu, \phi) \propto -2\log L(\mu, \phi, \hat{\sigma}_w^2)$.[3.6] Because (3.108) and (3.110) 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

$$\begin{aligned} L(\mu, \phi, \sigma_w^2 \mid x_1) &= \prod_{t=2}^{n} f_w\left[(x_t-\mu) - \phi(x_{t-1}-\mu)\right] \\ &= (2\pi\sigma_w^2)^{-(n-1)/2} \exp\left[-\frac{S_c(\mu, \phi)}{2\sigma_w^2}\right], \end{aligned} \tag{3.111}$$

where the *conditional sum of squares* is

$$S_c(\mu, \phi) = \sum_{t=2}^{n} [(x_t - \mu) - \phi(x_{t-1} - \mu)]^2 . \tag{3.112}$$

The conditional MLE of σ_w^2 is

$$\hat{\sigma}_w^2 = S_c(\hat{\mu}, \hat{\phi})/(n-1), \tag{3.113}$$

and $\hat{\mu}$ and $\hat{\phi}$ 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

$$S_c(\mu, \phi) = \sum_{t=2}^{n} [x_t - (\alpha + \phi x_{t-1})]^2 . \tag{3.114}$$

The problem is now the linear regression problem stated in Section 2.1. Following the results from least squares estimation, we have $\hat{\alpha} = \bar{x}_{(2)} - \hat{\phi}\bar{x}_{(1)}$, 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

$$\hat{\mu} = \frac{\bar{x}_{(2)} - \hat{\phi}\bar{x}_{(1)}}{1 - \hat{\phi}} \tag{3.115}$$

$$\hat{\phi} = \frac{\sum_{t=2}^{n} (x_t - \bar{x}_{(2)})(x_{t-1} - \bar{x}_{(1)})}{\sum_{t=2}^{n} (x_{t-1} - \bar{x}_{(1)})^2}. \tag{3.116}$$

From (3.115) and (3.116), we see that $\hat{\mu} \approx \bar{x}$ and $\hat{\phi} \approx \hat{\rho}(1)$. 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 σ_w^2 in (3.113) to be equivalent to the least squares estimator, that is, divide $S_c(\hat{\mu}, \hat{\phi})$ by $(n-3)$ instead of $(n-1)$ in (3.113).

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

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

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

$$L(\beta, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2} \left[r_1(\beta) r_2(\beta) \cdots r_n(\beta) \right]^{-1/2} \exp\left[-\frac{S(\beta)}{2\sigma_w^2} \right], \tag{3.117}$$

where

$$S(\beta) = \sum_{t=1}^{n} \left[\frac{(x_t - x_t^{t-1}(\beta))^2}{r_t(\beta)} \right]. \tag{3.118}$$

Both x_t^{t-1} and r_t are functions of β alone, and we make that fact explicit in (3.117)-(3.118). Given values for β and σ_w^2, the likelihood may be evaluated using the techniques of Section 3.4. Maximum likelihood estimation would now proceed by maximizing (3.117) with respect to β and σ_w^2. As in the AR(1) example, we have

$$\hat{\sigma}_w^2 = n^{-1} S(\hat{\beta}), \tag{3.119}$$

where $\hat{\beta}$ is the value of β that minimizes the concentrated likelihood

$$l(\beta) = \log \left[n^{-1} S(\beta) \right] + n^{-1} \sum_{t=1}^{n} \log r_t(\beta). \tag{3.120}$$

For the AR(1) model (3.106) 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 (3.107) is identical to the innovations form of the likelihood given by (3.117). Moreover, the generic $S(\beta)$ in (3.118) is $S(\mu, \phi)$ given in (3.108) and the generic $l(\beta)$ in (3.120) is $l(\mu, \phi)$ in (3.110).

Unconditional least squares would be performed by minimizing (3.118) with respect to β. Conditional least squares estimation would involve minimizing (3.118) with respect to β 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.