5.2. *The Likelihood Function for a Gaussian AR(1) Process*

Evaluating the Likelihood Function

A Gaussian $AR(1)$ process takes the form

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t, \qquad [5.2.1]$$

with $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. For this case, the vector of population parameters to be estimated consists of $\boldsymbol{\theta} \equiv (c, \phi, \sigma^2)'$.

Consider the probability distribution of Y_1, the first observation in the sample. From equations [3.4.3] and [3.4.4] this is a random variable with mean

$$E(Y_1) = \mu = c/(1 - \phi)$$

and variance

$$E(Y_1 - \mu)^2 = \sigma^2/(1 - \phi^2).$$

Since $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ is Gaussian, Y_1 is also Gaussian. Hence, the density of the first observation takes the form

$$f_{Y_1}(y_1; \boldsymbol{\theta}) = f_{Y_1}(y_1; c, \phi, \sigma^2) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2/(1 - \phi^2)}} \exp\left[\frac{-\{y_1 - [c/(1 - \phi)]\}^2}{2\sigma^2/(1 - \phi^2)}\right]. \qquad [5.2.2]$$

Next consider the distribution of the second observation Y_2 conditional on observing $Y_1 = y_1$. From [5.2.1],

$$Y_2 = c + \phi Y_1 + \varepsilon_2. \qquad [5.2.3]$$

Conditioning on $Y_1 = y_1$ means treating the random variable Y_1 as if it were the deterministic constant y_1. For this case, [5.2.3] gives Y_2 as the constant $(c + \phi y_1)$ plus the $N(0, \sigma^2)$ variable ε_2. Hence,

$$(Y_2 | Y_1 = y_1) \sim N((c + \phi y_1), \sigma^2),$$

meaning

$$f_{Y_2|Y_1}(y_2|y_1; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_2 - c - \phi y_1)^2}{2\sigma^2}\right]. \qquad [5.2.4]$$

The joint density of observations 1 and 2 is then just the product of [5.2.4] and [5.2.2]:

$$f_{Y_2, Y_1}(y_2, y_1; \boldsymbol{\theta}) = f_{Y_2|Y_1}(y_2|y_1; \boldsymbol{\theta}) \cdot f_{Y_1}(y_1; \boldsymbol{\theta}).$$

Similarly, the distribution of the third observation conditional on the first two is

$$f_{Y_3|Y_2, Y_1}(y_3|y_2, y_1; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_3 - c - \phi y_2)^2}{2\sigma^2}\right],$$

from which

$$f_{Y_3, Y_2, Y_1}(y_3, y_2, y_1; \boldsymbol{\theta}) = f_{Y_3|Y_2, Y_1}(y_3|y_2, y_1; \boldsymbol{\theta}) \cdot f_{Y_2, Y_1}(y_2, y_1; \boldsymbol{\theta}).$$

In general, the values of $Y_1, Y_2, \ldots, Y_{t-1}$ matter for Y_t only through the value of Y_{t-1}, and the density of observation t conditional on the preceding $t - 1$ observations is given by

$$f_{Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_1}(y_t|y_{t-1}, y_{t-2}, \ldots, y_1; \boldsymbol{\theta}) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\right]. \qquad [5.2.5]$$

The joint density of the first t observations is then

$$f_{Y_t,Y_{t-1},\ldots,Y_1}(y_t, y_{t-1}, \ldots, y_1; \boldsymbol{\theta}) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}) \cdot f_{Y_{t-1},Y_{t-2},\ldots,Y_1}(y_{t-1}, y_{t-2}, \ldots, y_1; \boldsymbol{\theta}). \quad [5.2.6]$$

The likelihood of the complete sample can thus be calculated as

$$f_{Y_T,Y_{T-1},\ldots,Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta}) = f_{Y_1}(y_1; \boldsymbol{\theta}) \cdot \prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}). \quad [5.2.7]$$

The log likelihood function (denoted $\mathcal{L}(\boldsymbol{\theta})$) can be found by taking logs of [5.2.7]:

$$\mathcal{L}(\boldsymbol{\theta}) = \log f_{Y_1}(y_1; \boldsymbol{\theta}) + \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}). \quad [5.2.8]$$

Clearly, the value of $\boldsymbol{\theta}$ that maximizes [5.2.8] is identical to the value that maximizes [5.2.7]. However, Section 5.8 presents a number of useful results that can be calculated as a by-product of the maximization if one always poses the problem as maximization of the log likelihood function [5.2.8] rather than the likelihood function [5.2.7].

Substituting [5.2.2] and [5.2.5] into [5.2.8], the log likelihood for a sample of size T from a Gaussian $AR(1)$ process is seen to be

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = {} & -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\log[\sigma^2/(1-\phi^2)] \\ & - \frac{\{y_1 - [c/(1-\phi)]\}^2}{2\sigma^2/(1-\phi^2)} - [(T-1)/2]\log(2\pi) \\ & - [(T-1)/2]\log(\sigma^2) - \sum_{t=2}^{T}\left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\right].\end{aligned} \quad [5.2.9]$$

An Alternative Expression for the Likelihood Function

A different description of the likelihood function for a sample of size T from a Gaussian $AR(1)$ process is sometimes useful. Collect the full set of observations in a $(T \times 1)$ vector,

$$\underset{(T\times 1)}{\mathbf{y}} \equiv (y_1, y_2, \ldots, y_T)'.$$

This vector could be viewed as a single realization from a T-dimensional Gaussian distribution. The mean of this $(T \times 1)$ vector is

$$\begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_T) \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix}, \quad [5.2.10]$$

where, as before, $\mu = c/(1-\phi)$. In vector form, [5.2.10] could be written

$$E(\mathbf{Y}) = \boldsymbol{\mu},$$

where $\boldsymbol{\mu}$ denotes the $(T \times 1)$ vector on the right side of [5.2.10]. The variance-covariance matrix of $\mathbf{Y}$ is given by

$$E[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'] = \boldsymbol{\Omega}, \quad [5.2.11]$$

where

$$\boldsymbol{\Omega} = \begin{bmatrix} E(Y_1 - \mu)^2 & E(Y_1 - \mu)(Y_2 - \mu) & \cdots & E(Y_1 - \mu)(Y_T - \mu) \\ E(Y_2 - \mu)(Y_1 - \mu) & E(Y_2 - \mu)^2 & \cdots & E(Y_2 - \mu)(Y_T - \mu) \\ \vdots & \vdots & \cdots & \vdots \\ E(Y_T - \mu)(Y_1 - \mu) & E(Y_T - \mu)(Y_2 - \mu) & \cdots & E(Y_T - \mu)^2 \end{bmatrix}. \quad [5.2.12]$$

The elements of this matrix correspond to autocovariances of Y. Recall that the jth autocovariance for an $AR(1)$ process is given by

$$E(Y_t - \mu)(Y_{t-j} - \mu) = \sigma^2\phi^j/(1 - \phi^2). \quad [5.2.13]$$

Hence, [5.2.12] can be written as

$$\boldsymbol{\Omega} = \sigma^2\mathbf{V}, \quad [5.2.14]$$

where

$$\mathbf{V} = \frac{1}{1 - \phi^2}\begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{T-1} \\ \phi & 1 & \phi & \cdots & \phi^{T-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{T-3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \cdots & 1 \end{bmatrix}. \quad [5.2.15]$$

Viewing the observed sample y as a single draw from a $N(\boldsymbol{\mu}, \boldsymbol{\Omega})$ distribution, the sample likelihood could be written down immediately from the formula for the multivariate Gaussian density:

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2}\,|\boldsymbol{\Omega}^{-1}|^{1/2}\exp[-\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})], \quad [5.2.16]$$

with log likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = (-T/2)\log(2\pi) + \tfrac{1}{2}\log|\boldsymbol{\Omega}^{-1}| - \tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu}). \quad [5.2.17]$$

Evidently, [5.2.17] and [5.2.9] must represent the identical function of $(y_1, y_2, \ldots, y_T)$. To verify that this is indeed the case, define

$$\underset{(T\times T)}{\mathbf{L}} \equiv \begin{bmatrix} \sqrt{1 - \phi^2} & 0 & 0 & \cdots & 0 & 0 \\ -\phi & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\phi & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\phi & 1 \end{bmatrix}. \quad [5.2.18]$$

It is straightforward to show that[1]

$$\mathbf{L}'\mathbf{L} = \mathbf{V}^{-1}, \quad [5.2.19]$$

[1]By direct multiplication, one calculates

$$\mathbf{LV} = \frac{1}{1 - \phi^2}\begin{bmatrix} \sqrt{1 - \phi^2} & \phi\sqrt{1 - \phi^2} & \phi^2\sqrt{1 - \phi^2} & \cdots & \phi^{T-1}\sqrt{1 - \phi^2} \\ 0 & (1 - \phi^2) & \phi(1 - \phi^2) & \cdots & \phi^{T-2}(1 - \phi^2) \\ 0 & 0 & (1 - \phi^2) & \cdots & \phi^{T-3}(1 - \phi^2) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & (1 - \phi^2) \end{bmatrix},$$

and premultiplying this by L' produces the $(T \times T)$ identity matrix. Thus, $\mathbf{L}'\mathbf{LV} = \mathbf{I}_T$, confirming [5.2.19].

implying from [5.2.14] that

$$\mathbf{\Omega}^{-1} = \sigma^{-2}\mathbf{L}'\mathbf{L}. \qquad [5.2.20]$$

Substituting [5.2.20] into [5.2.17] results in

$$\mathcal{L}(\boldsymbol{\theta}) = (-T/2)\log(2\pi) + \tfrac{1}{2}\log|\sigma^{-2}\mathbf{L}'\mathbf{L}| - \tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\sigma^{-2}\mathbf{L}'\mathbf{L}(\mathbf{y} - \boldsymbol{\mu}). \qquad [5.2.21]$$

Define the $(T \times 1)$ vector $\tilde{\mathbf{y}}$ to be

$$\begin{aligned}\tilde{\mathbf{y}} &\equiv \mathbf{L}(\mathbf{y} - \boldsymbol{\mu}) \\ &= \begin{bmatrix} \sqrt{1-\phi^2} & 0 & 0 & \cdots & 0 & 0 \\ -\phi & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\phi & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\phi & 1 \end{bmatrix} \begin{bmatrix} y_1 - \mu \\ y_2 - \mu \\ y_3 - \mu \\ \vdots \\ y_T - \mu \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{1-\phi^2}\,(y_1 - \mu) \\ (y_2 - \mu) - \phi(y_1 - \mu) \\ (y_3 - \mu) - \phi(y_2 - \mu) \\ \vdots \\ (y_T - \mu) - \phi(y_{T-1} - \mu) \end{bmatrix}. \end{aligned} \qquad [5.2.22]$$

Substituting $\mu = c/(1 - \phi)$, this becomes

$$\tilde{\mathbf{y}} = \begin{bmatrix} \sqrt{1-\phi^2}\,[y_1 - c/(1-\phi)] \\ y_2 - c - \phi y_1 \\ y_3 - c - \phi y_2 \\ \vdots \\ y_T - c - \phi y_{T-1} \end{bmatrix}.$$

The last term in [5.2.21] can thus be written

$$\begin{aligned}\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\sigma^{-2}\mathbf{L}'\mathbf{L}(\mathbf{y} - \boldsymbol{\mu}) &= [1/(2\sigma^2)]\tilde{\mathbf{y}}'\tilde{\mathbf{y}} \\ &= [1/(2\sigma^2)](1 - \phi^2)[y_1 - c/(1-\phi)]^2 \\ &\quad + [1/(2\sigma^2)] \sum_{t=2}^{T} (y_t - c - \phi y_{t-1})^2. \end{aligned} \qquad [5.2.23]$$

The middle term in [5.2.21] is similarly

$$\begin{aligned}\tfrac{1}{2}\log|\sigma^{-2}\mathbf{L}'\mathbf{L}| &= \tfrac{1}{2}\log\{\sigma^{-2T} \cdot |\mathbf{L}'\mathbf{L}|\} \\ &= -\tfrac{1}{2}\log\sigma^{2T} + \tfrac{1}{2}\log|\mathbf{L}'\mathbf{L}| \\ &= (-T/2)\log\sigma^2 + \log|\mathbf{L}|, \end{aligned} \qquad [5.2.24]$$

where use has been made of equations [A.4.8], [A.4.9], and [A.4.11] in the Mathematical Review (Appendix A) at the end of the book. Moreover, since $\mathbf{L}$ is lower triangular, its determinant is given by the product of the terms along the principal diagonal: $|\mathbf{L}| = \sqrt{1-\phi^2}$. Thus, [5.2.24] states that

$$\tfrac{1}{2}\log|\sigma^{-2}\mathbf{L}'\mathbf{L}| = (-T/2)\log\sigma^2 + \tfrac{1}{2}\log(1 - \phi^2). \qquad [5.2.25]$$

Substituting [5.2.23] and [5.2.25] into [5.2.21] reproduces [5.2.9]. Thus, equations [5.2.17] and [5.2.9] are just two different expressions for the same magnitude, as claimed. Either expression accurately describes the log likelihood function.

Expression [5.2.17] requires inverting a $(T \times T)$ matrix, whereas [5.2.9] does not. Thus, expression [5.2.9] is clearly to be preferred for computations. It avoids inverting a $(T \times T)$ matrix by writing Y_t as the sum of a forecast $(c + \phi Y_{t-1})$ and a forecast error (ε_t). The forecast error is independent from previous observations by construction, so the log of its density is simply added to the log likelihood of the preceding observations. This approach is known as a *prediction-error decomposition* of the likelihood function.

Exact Maximum Likelihood Estimates for the Gaussian AR(1) Process

The *MLE* $\hat{\boldsymbol{\theta}}$ is the value for which [5.2.9] is maximized. In principle, this requires differentiating [5.2.9] and setting the result equal to zero. In practice, when an attempt is made to carry this out, the result is a system of nonlinear equations in $\boldsymbol{\theta}$ and $(y_1, y_2, \ldots, y_T)$ for which there is no simple solution for $\boldsymbol{\theta}$ in terms of $(y_1, y_2, \ldots, y_T)$. Maximization of [5.2.9] thus requires iterative or numerical procedures described in Section 5.7.

Conditional Maximum Likelihood Estimates

An alternative to numerical maximization of the exact likelihood function is to regard the value of y_1 as deterministic and maximize the likelihood conditioned on the first observation,

$$f_{Y_T,Y_{T-1},\ldots,Y_2|Y_1}(y_T, y_{T-1}, \ldots, y_2|y_1; \boldsymbol{\theta}) = \prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}), \quad [5.2.26]$$

the objective then being to maximize

$$\begin{aligned} &\log f_{Y_T,Y_{T-1},\ldots,Y_2|Y_1}(y_T, y_{T-1}, \ldots, y_2|y_1; \boldsymbol{\theta}) \\ &\quad = -[(T-1)/2]\log(2\pi) - [(T-1)/2]\log(\sigma^2) \\ &\qquad - \sum_{t=2}^{T}\left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\right]. \end{aligned} \quad [5.2.27]$$

Maximization of [5.2.27] with respect to c and ϕ is equivalent to minimization of

$$\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2, \quad [5.2.28]$$

which is achieved by an ordinary least squares (*OLS*) regression of y_t on a constant and its own lagged value. The conditional maximum likelihood estimates of c and ϕ are therefore given by

$$\begin{bmatrix} \hat{c} \\ \hat{\phi} \end{bmatrix} = \begin{bmatrix} T-1 & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma y_t \\ \Sigma y_{t-1} y_t \end{bmatrix},$$

where Σ denotes summation over $t = 2, 3, \ldots, T$.

The conditional maximum likelihood estimate of the innovation variance is found by differentiating [5.2.27] with respect to σ^2 and setting the result equal to zero:

$$\frac{-(T-1)}{2\sigma^2} + \sum_{t=2}^{T}\left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^4}\right] = 0,$$

or

$$\hat{\sigma}^2 = \sum_{t=2}^{T} \left[\frac{(y_t - \hat{c} - \hat{\phi} y_{t-1})^2}{T - 1} \right].$$

In other words, the conditional *MLE* is the average squared residual from the *OLS* regression [5.2.28].

In contrast to exact maximum likelihood estimates, the conditional maximum likelihood estimates are thus trivial to compute. Moreover, if the sample size T is sufficiently large, the first observation makes a negligible contribution to the total likelihood. The exact *MLE* and conditional *MLE* turn out to have the same large-sample distribution, provided that $|\phi| < 1$. And when $|\phi| > 1$, the conditional *MLE* continues to provide consistent estimates, whereas maximization of [5.2.9] does not. This is because [5.2.9] is derived from [5.2.2], which does not accurately describe the density of Y_1 when $|\phi| > 1$. For these reasons, in most applications the parameters of an autoregression are estimated by *OLS* (conditional maximum likelihood) rather than exact maximum likelihood.

5.3. *The Likelihood Function for a Gaussian* AR(p) *Process*

This section discusses a Gaussian $AR(p)$ process,

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t, \qquad [5.3.1]$$

with $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. In this case, the vector of population parameters to be estimated is $\boldsymbol{\theta} = (c, \phi_1, \phi_2, \ldots, \phi_p, \sigma^2)'$.

Evaluating the Likelihood Function

A combination of the two methods described for the $AR(1)$ case is used to calculate the likelihood function for a sample of size T for an $AR(p)$ process. The first p observations in the sample $(y_1, y_2, \ldots, y_p)$ are collected in a $(p \times 1)$ vector $\mathbf{y}_p$, which is viewed as the realization of a p-dimensional Gaussian variable. The mean of this vector is $\boldsymbol{\mu}_p$, which denotes a $(p \times 1)$ vector each of whose elements is given by

$$\mu = c/(1 - \phi_1 - \phi_2 - \cdots - \phi_p). \qquad [5.3.2]$$

Let $\sigma^2 \mathbf{V}_p$ denote the $(p \times p)$ variance-covariance matrix of $(Y_1, Y_2, \ldots, Y_p)$:

$$\sigma^2 \mathbf{V}_p = \begin{bmatrix} E(Y_1 - \mu)^2 & E(Y_1 - \mu)(Y_2 - \mu) & \cdots & E(Y_1 - \mu)(Y_p - \mu) \\ E(Y_2 - \mu)(Y_1 - \mu) & E(Y_2 - \mu)^2 & \cdots & E(Y_2 - \mu)(Y_p - \mu) \\ \vdots & \vdots & \cdots & \vdots \\ E(Y_p - \mu)(Y_1 - \mu) & E(Y_p - \mu)(Y_2 - \mu) & \cdots & E(Y_p - \mu)^2 \end{bmatrix}. \qquad [5.3.3]$$

For example, for a first-order autoregression ($p = 1$), $\mathbf{V}_p$ is the scalar $1/(1 - \phi^2)$. For a general pth-order autoregression,

$$\sigma^2 \mathbf{V}_p = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{p-3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 \end{bmatrix},$$

where γ_j, the jth autocovariance for an $AR(p)$ process, can be calculated using the methods in Chapter 3. The density of the first p observations is then that of a $N(\boldsymbol{\mu}_p, \sigma^2\mathbf{V}_p)$ variable:

$$\begin{aligned} f_{Y_p,Y_{p-1},\ldots,Y_1}(y_p, y_{p-1}, \ldots, y_1; \boldsymbol{\theta}) \\ &= (2\pi)^{-p/2}|\sigma^{-2}\mathbf{V}_p^{-1}|^{1/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p - \boldsymbol{\mu}_p)\right] \\ &= (2\pi)^{-p/2}(\sigma^{-2})^{p/2}|\mathbf{V}_p^{-1}|^{1/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p - \boldsymbol{\mu}_p)\right], \end{aligned} \tag{5.3.4}$$

where use has been made of result [A.4.8].

For the remaining observations in the sample, $(y_{p+1}, y_{p+2}, \ldots, y_T)$, the prediction-error decomposition can be used. Conditional on the first $t-1$ observations, the tth observation is Gaussian with mean

$$c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p}$$

and variance σ^2. Only the p most recent observations matter for this distribution. Hence, for $t > p$,

$$\begin{aligned} f_{Y_t|Y_{t-1},Y_{t-2},\ldots,Y_1}(y_t|y_{t-1}, y_{t-2}, \ldots, y_1; \boldsymbol{\theta}) \\ &= f_{Y_t|Y_{t-1},Y_{t-2},\ldots,Y_{t-p}}(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p}; \boldsymbol{\theta}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}\right]. \end{aligned}$$

The likelihood function for the complete sample is then

$$\begin{aligned} f_{Y_T,Y_{T-1},\ldots,Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta}) \\ &= f_{Y_p,Y_{p-1},\ldots,Y_1}(y_p, y_{p-1}, \ldots, y_1; \boldsymbol{\theta}) \\ &\quad \times \prod_{t=p+1}^{T} f_{Y_t|Y_{t-1},Y_{t-2},\ldots,Y_{t-p}}(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p}; \boldsymbol{\theta}), \end{aligned} \tag{5.3.5}$$

and the log likelihood is therefore

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \log f_{Y_T,Y_{T-1},\ldots,Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta}) \\ &= -\frac{p}{2}\log(2\pi) - \frac{p}{2}\log(\sigma^2) + \frac{1}{2}\log|\mathbf{V}_p^{-1}| \\ &\quad - \frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p - \boldsymbol{\mu}_p) \\ &\quad - \frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) \\ &\quad - \sum_{t=p+1}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2} \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log|\mathbf{V}_p^{-1}| \\ &\quad - \frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p - \boldsymbol{\mu}_p) \\ &\quad - \sum_{t=p+1}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}. \end{aligned} \tag{5.3.6}$$

Evaluation of [5.3.6] requires inverting the $(p \times p)$ matrix $\mathbf{V}_p$. Denote the row i, column j element of $\mathbf{V}_p^{-1}$ by $v^{ij}(p)$. Galbraith and Galbraith (1974, equation

16, p. 70) showed that

$$v^{ij}(p) = \left[\sum_{k=0}^{i-1} \phi_k\phi_{k+j-i} - \sum_{k=p+1-j}^{p+i-j} \phi_k\phi_{k+j-i}\right] \quad \text{for } 1 \le i \le j \le p, \quad [5.3.7]$$

where $\phi_0 \equiv -1$. Values of $v^{ij}(p)$ for $i > j$ can be inferred from the fact that $\mathbf{V}_p^{-1}$ is symmetric ($v^{ij}(p) = v^{ji}(p)$). For example, for an $AR(1)$ process, $\mathbf{V}_p^{-1}$ is a scalar whose value is found by taking $i = j = p = 1$:

$$\mathbf{V}_1^{-1} = \left[\sum_{k=0}^{0} \phi_k\phi_k - \sum_{k=1}^{1} \phi_k\phi_k\right] = (\phi_0^2 - \phi_1^2) = (1 - \phi^2).$$

Thus $\sigma^2\mathbf{V}_1 = \sigma^2/(1 - \phi^2)$, which indeed reproduces the formula for the variance of an $AR(1)$ process. For $p = 2$, equation [5.3.7] implies

$$\mathbf{V}_2^{-1} = \begin{bmatrix} (1 - \phi_2^2) & -(\phi_1 + \phi_1\phi_2) \\ -(\phi_1 + \phi_1\phi_2) & (1 - \phi_2^2) \end{bmatrix},$$

from which one readily calculates

$$|\mathbf{V}_2^{-1}| = \left|(1 + \phi_2)\begin{bmatrix} (1 - \phi_2) & -\phi_1 \\ -\phi_1 & (1 - \phi_2) \end{bmatrix}\right| = (1 + \phi_2)^2[(1 - \phi_2)^2 - \phi_1^2]$$

and

$$\begin{aligned}(\mathbf{y}_2 - \boldsymbol{\mu}_2)'\mathbf{V}_2^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2) &= [(y_1 - \mu) \quad (y_2 - \mu)](1 + \phi_2)\begin{bmatrix} (1 - \phi_2) & -\phi_1 \\ -\phi_1 & (1 - \phi_2) \end{bmatrix}\begin{bmatrix} (y_1 - \mu) \\ (y_2 - \mu) \end{bmatrix} \\ &= (1 + \phi_2) \times \{(1 - \phi_2)(y_1 - \mu)^2 \\ &\quad - 2\phi_1(y_1 - \mu)(y_2 - \mu) + (1 - \phi_2)(y_2 - \mu)^2\}.\end{aligned}$$

The exact log likelihood for a Gaussian $AR(2)$ process is thus given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = &-\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log\{(1 + \phi_2)^2[(1 - \phi_2)^2 - \phi_1^2]\} \\ &- \left\{\frac{1 + \phi_2}{2\sigma^2}\right\} \times \{(1 - \phi_2)(y_1 - \mu)^2 \\ &- 2\phi_1(y_1 - \mu)(y_2 - \mu) + (1 - \phi_2)(y_2 - \mu)^2\} \\ &- \sum_{t=3}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2})^2}{2\sigma^2},\end{aligned} \quad [5.3.8]$$

where $\mu = c/(1 - \phi_1 - \phi_2)$.

Conditional Maximum Likelihood Estimates

Maximization of the exact log likelihood for an $AR(p)$ process [5.3.6] must be accomplished numerically. In contrast, the log of the likelihood conditional on the first p observations assumes the simple form

$$\begin{aligned}&\log f_{Y_T, Y_{T-1}, \ldots, Y_{p+1}|Y_p, \ldots, Y_1}(y_T, y_{T-1}, \ldots, y_{p+1}|y_p, \ldots, y_1; \boldsymbol{\theta}) \\ &= -\frac{T - p}{2}\log(2\pi) - \frac{T - p}{2}\log(\sigma^2) \\ &\quad - \sum_{t=p+1}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}.\end{aligned} \quad [5.3.9]$$

The values of $c, \phi_1, \phi_2, \ldots, \phi_p$ that maximize [5.3.9] are the same as those that minimize

$$\sum_{t=p+1}^{T} (y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2. \qquad [5.3.10]$$

Thus, the conditional maximum likelihood estimates of these parameters can be obtained from an *OLS* regression of y_t on a constant and p of its own lagged values. The conditional maximum likelihood estimate of σ^2 turns out to be the average squared residual from this regression:

$$\hat{\sigma}^2 = \frac{1}{T-p} \sum_{t=p+1}^{T} (y_t - \hat{c} - \hat{\phi}_1 y_{t-1} - \hat{\phi}_2 y_{t-2} - \cdots - \hat{\phi}_p y_{t-p})^2.$$

The exact maximum likelihood estimates and the conditional maximum likelihood estimates again have the same large-sample distribution.

Maximum Likelihood Estimation for Non-Gaussian Time Series

We noted in Chapter 4 that an *OLS* regression of a variable on a constant and p of its lags would yield a consistent estimate of the coefficients of the linear projection,

$$\hat{E}(Y_t | Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}),$$

provided that the process is ergodic for second moments. This *OLS* regression also maximizes the Gaussian conditional log likelihood [5.3.9]. Thus, even if the process is non-Gaussian, if we mistakenly form a Gaussian log likelihood function and maximize it, the resulting estimates $(\hat{c}, \hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p)$ will provide consistent estimates of the population parameters in [5.3.1].

An estimate that maximizes a misspecified likelihood function (for example, an *MLE* calculated under the assumption of a Gaussian process when the true data are non-Gaussian) is known as a *quasi-maximum likelihood estimate*. Sometimes, as turns out to be the case here, quasi-maximum likelihood estimation provides consistent estimates of the population parameters of interest. However, standard errors for the estimated coefficients that are calculated under the Gaussianity assumption need not be correct if the true data are non-Gaussian.[2]

Alternatively, if the raw data are non-Gaussian, sometimes a simple transformation such as taking logs will produce a Gaussian time series. For a positive random variable Y_t, Box and Cox (1964) proposed the general class of transformations

$$Y_t^{(\lambda)} = \begin{cases} \dfrac{Y_t^{\lambda} - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log Y_t & \text{for } \lambda = 0. \end{cases}$$

One approach is to pick a particular value of λ and maximize the likelihood function for $Y_t^{(\lambda)}$ under the assumption that $Y_t^{(\lambda)}$ is a Gaussian *ARMA* process. The value of λ that is associated with the highest value of the maximized likelihood is taken as the best transformation. However, Nelson and Granger (1979) reported discouraging results from this method in practice.

[2]These points were first raised by White (1982) and are discussed further in Sections 5.8 and 14.4.

Li and McLeod (1988) and Janacek and Swift (1990) described approaches to maximum likelihood estimation for some non-Gaussian *ARMA* models. Martin (1981) discussed robust time series estimation for contaminated data.

5.4. *The Likelihood Function for a Gaussian* MA*(1) Process*

Conditional Likelihood Function

Calculation of the likelihood function for an autoregression turned out to be much simpler if we conditioned on initial values for the Y's. Similarly, calculation of the likelihood function for a moving average process is simpler if we condition on initial values for the ε's.

Consider the Gaussian $MA(1)$ process

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1} \qquad [5.4.1]$$

with $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. Let $\boldsymbol{\theta} = (\mu, \theta, \sigma^2)'$ denote the population parameters to be estimated. If the value of ε_{t-1} were known with certainty, then

$$Y_t|\varepsilon_{t-1} \sim N((\mu + \theta\varepsilon_{t-1}), \sigma^2)$$

or

$$f_{Y_t|\varepsilon_{t-1}}(y_t|\varepsilon_{t-1}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_t - \mu - \theta\varepsilon_{t-1})^2}{2\sigma^2}\right]. \qquad [5.4.2]$$

Suppose that we knew for certain that $\varepsilon_0 = 0$. Then

$$(Y_1|\varepsilon_0 = 0) \sim N(\mu, \sigma^2).$$

Moreover, given observation of y_1, the value of ε_1 is then known with certainty as well:

$$\varepsilon_1 = y_1 - \mu,$$

allowing application of [5.4.2] again:

$$f_{Y_2|Y_1,\varepsilon_0=0}(y_2|y_1, \varepsilon_0 = 0; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_2 - \mu - \theta\varepsilon_1)^2}{2\sigma^2}\right].$$

Since ε_1 is known with certainty, ε_2 can be calculated from

$$\varepsilon_2 = y_2 - \mu - \theta\varepsilon_1.$$

Proceeding in this fashion, it is clear that given knowledge that $\varepsilon_0 = 0$, the full sequence $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T\}$ can be calculated from $\{y_1, y_2, \ldots, y_T\}$ by iterating on

$$\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1} \qquad [5.4.3]$$

for $t = 1, 2, \ldots, T$, starting from $\varepsilon_0 = 0$. The conditional density of the tth observation can then be calculated from [5.4.2] as

$$\begin{aligned} &f_{Y_t|Y_{t-1},Y_{t-2},\ldots,Y_1,\varepsilon_0=0}(y_t|y_{t-1}, y_{t-2}, \ldots, y_1, \varepsilon_0 = 0; \boldsymbol{\theta}) \\ &\quad = f_{Y_t|\varepsilon_{t-1}}(y_t|\varepsilon_{t-1}; \boldsymbol{\theta}) \qquad [5.4.4] \\ &\quad = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-\varepsilon_t^2}{2\sigma^2}\right]. \end{aligned}$$

The sample likelihood would then be the product of these individual densities:

$$f_{Y_T, Y_{T-1}, \ldots, Y_1 | \varepsilon_0 = 0}(y_T, y_{T-1}, \ldots, y_1 | \varepsilon_0 = 0; \boldsymbol{\theta})$$
$$= f_{Y_1 | \varepsilon_0 = 0}(y_1 | \varepsilon_0 = 0; \boldsymbol{\theta}) \prod_{t=2}^{T} f_{Y_t | Y_{t-1}, Y_{t-2}, \ldots, Y_1, \varepsilon_0 = 0}(y_t | y_{t-1}, y_{t-2}, \ldots, y_1, \varepsilon_0 = 0; \boldsymbol{\theta}).$$

The conditional log likelihood is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \log f_{Y_T, Y_{T-1}, \ldots, Y_1 | \varepsilon_0 = 0}(y_T, y_{T-1}, \ldots, y_1 | \varepsilon_0 = 0; \boldsymbol{\theta}) \\ &= -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) - \sum_{t=1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}. \end{aligned} \qquad [5.4.5]$$

For a particular numerical value of $\boldsymbol{\theta}$, we thus calculate the sequence of ε's implied by the data from [5.4.3]. The conditional log likelihood [5.4.5] is then a function of the sum of squares of these ε's. Although it is simple to program this iteration by computer, the log likelihood is a fairly complicated nonlinear function of μ and θ, so that an analytical expression for the maximum likelihood estimates of μ and θ is not readily calculated. Hence, even the conditional maximum likelihood estimates for an $MA(1)$ process must be found by numerical optimization.

Iteration on [5.4.3] from an arbitrary starting value of ε_0 will result in

$$\begin{aligned} \varepsilon_t = (y_t - \mu) - \theta(y_{t-1} - \mu) + \theta^2(y_{t-2} - \mu) - \cdots \\ + (-1)^{t-1}\theta^{t-1}(y_1 - \mu) + (-1)^t\theta^t\varepsilon_0. \end{aligned}$$

If $|\theta|$ is substantially less than unity, the effect of imposing $\varepsilon_0 = 0$ will quickly die out and the conditional likelihood [5.4.4] will give a good approximation to the unconditional likelihood for a reasonably large sample size. By contrast, if $|\theta| > 1$, the consequences of imposing $\varepsilon_0 = 0$ accumulate over time. The conditional approach is not reasonable in such a case. If numerical optimization of [5.4.5] results in a value of θ that exceeds 1 in absolute value, the results must be discarded. The numerical optimization should be attempted again with the reciprocal of $\hat{\theta}$ used as a starting value for the numerical search procedure.

Exact Likelihood Function

Two convenient algorithms are available for calculating the exact likelihood function for a Gaussian $MA(1)$ process. One approach is to use the Kalman filter discussed in Chapter 13. A second approach uses the triangular factorization of the variance-covariance matrix. The second approach is described here.

As in Section 5.2, the observations on y can be collected in a $(T \times 1)$ vector $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$ with mean $\boldsymbol{\mu} = (\mu, \mu, \ldots, \mu)'$ and $(T \times T)$ variance-covariance matrix

$$\boldsymbol{\Omega} = E(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'.$$

The variance-covariance matrix for T consecutive draws from an $MA(1)$ process is

$$\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} (1 + \theta^2) & \theta & 0 & \cdots & 0 \\ \theta & (1 + \theta^2) & \theta & \cdots & 0 \\ 0 & \theta & (1 + \theta^2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & (1 + \theta^2) \end{bmatrix}.$$

The likelihood function is then

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2} |\boldsymbol{\Omega}|^{-1/2} \exp[-\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \boldsymbol{\mu})]. \qquad [5.4.6]$$

A prediction-error decomposition of the likelihood is provided from the triangular factorization of $\boldsymbol{\Omega}$,

$$\boldsymbol{\Omega} = \mathbf{ADA}', \tag{5.4.7}$$

where A is the lower triangular matrix given in [4.5.18] and D is the diagonal matrix in [4.5.19]. Substituting [5.4.7] into [5.4.6] gives

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = {} & (2\pi)^{-T/2}|\mathbf{ADA}'|^{-1/2} \\ & \times \exp[-\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'[\mathbf{A}']^{-1}\mathbf{D}^{-1}\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu})]. \end{aligned} \tag{5.4.8}$$

But A is a lower triangular matrix with 1s along the principal diagonal. Hence, $|\mathbf{A}| = 1$ and

$$|\mathbf{ADA}'| = |\mathbf{A}|\cdot|\mathbf{D}|\cdot|\mathbf{A}'| = |\mathbf{D}|.$$

Further defining

$$\tilde{\mathbf{y}} \equiv \mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \tag{5.4.9}$$

the likelihood [5.4.8] can be written

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2}\,|\mathbf{D}|^{-1/2} \exp[-\tfrac{1}{2}\tilde{\mathbf{y}}'\mathbf{D}^{-1}\tilde{\mathbf{y}}]. \tag{5.4.10}$$

Notice that [5.4.9] implies

$$\mathbf{A}\tilde{\mathbf{y}} = \mathbf{y} - \boldsymbol{\mu}.$$

The first row of this system states that $\tilde{y}_1 = y_1 - \mu$, while the tth row implies that

$$\tilde{y}_t = y_t - \mu - \frac{\theta[1 + \theta^2 + \theta^4 + \cdots + \theta^{2(t-2)}]}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(t-1)}}\,\tilde{y}_{t-1}. \tag{5.4.11}$$

The vector $\tilde{\mathbf{y}}$ can thus be calculated by iterating on [5.4.11] for $t = 2, 3, \ldots, T$ starting from $\tilde{y}_1 = y_1 - \mu$. The variable $\tilde{y}_t$ has the interpretation as the residual from a linear projection of y_t on a constant and $y_{t-1}, y_{t-2}, \ldots, y_1$, while the tth diagonal element of D gives the *MSE* of this linear projection:

$$d_{tt} = E(\tilde{Y}_t^2) = \sigma^2 \frac{1 + \theta^2 + \theta^4 + \cdots + \theta^{2t}}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(t-1)}}. \tag{5.4.12}$$

Since D is diagonal, its determinant is the product of the terms along the principal diagonal,

$$|\mathbf{D}| = \prod_{t=1}^{T} d_{tt}, \tag{5.4.13}$$

while the inverse of D is obtained by taking reciprocals of the terms along the principal diagonal. Hence,

$$\tilde{\mathbf{y}}'\mathbf{D}^{-1}\tilde{\mathbf{y}} = \sum_{t=1}^{T} \frac{\tilde{y}_t^2}{d_{tt}}. \tag{5.4.14}$$

Substituting [5.4.13] and [5.4.14] into [5.4.10], the likelihood function is

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2}\left[\prod_{t=1}^{T} d_{tt}\right]^{-1/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}\right]. \tag{5.4.15}$$

The exact log likelihood for a Gaussian $MA(1)$ process is therefore

$$\mathcal{L}(\boldsymbol{\theta}) = \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(d_{tt}) - \frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}. \tag{5.4.16}$$

Given numerical values for μ, θ, and σ^2, the sequence $\tilde{y}_t$ is calculated by iterating on [5.4.11] starting with $\tilde{y}_1 = y_1 - \mu$, while d_{tt} is given by [5.4.12].

In contrast to the conditional log likelihood function [5.4.5], expression [5.4.16] will be valid regardless of whether θ is associated with an invertible $MA(1)$ representation. The value of [5.4.16] at $\theta = \tilde{\theta}$, $\sigma^2 = \tilde{\sigma}^2$ will be identical to its value at $\theta = \tilde{\theta}^{-1}$, $\sigma^2 = \tilde{\theta}^2\tilde{\sigma}^2$; see Exercise 5.1.

5.5. *The Likelihood Function for a Gaussian* MA(q) *Process*

Conditional Likelihood Function

For the $MA(q)$ process,

$$Y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}, \qquad [5.5.1]$$

a simple approach is to condition on the assumption that the first q values for ε were all zero:

$$\varepsilon_0 = \varepsilon_{-1} = \cdots = \varepsilon_{-q+1} = 0. \qquad [5.5.2]$$

From these starting values we can iterate on

$$\varepsilon_t = y_t - \mu - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q} \qquad [5.5.3]$$

for $t = 1, 2, \ldots, T$. Let $\boldsymbol{\varepsilon}_0$ denote the $(q \times 1)$ vector $(\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1})'$. The conditional log likelihood is then

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \log f_{Y_T, Y_{T-1}, \ldots, Y_1|\boldsymbol{\varepsilon}_0 = \mathbf{0}}(y_T, y_{T-1}, \ldots, y_1|\boldsymbol{\varepsilon}_0 = \mathbf{0}; \boldsymbol{\theta}) \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}, \end{aligned} \qquad [5.5.4]$$

where $\boldsymbol{\theta} = (\mu, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2)'$. Again, expression [5.5.4] is useful only if all values of z for which

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

lie outside the unit circle.

Exact Likelihood Function

The exact likelihood function is given by

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2}|\boldsymbol{\Omega}|^{-1/2}\exp[-\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})], \qquad [5.5.5]$$

where as before $\mathbf{y} \equiv (y_1, y_2, \ldots, y_T)'$ and $\boldsymbol{\mu} \equiv (\mu, \mu, \ldots, \mu)'$. Here $\boldsymbol{\Omega}$ represents the variance-covariance matrix of T consecutive draws from an $MA(q)$ process:

$$\boldsymbol{\Omega} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_q & 0 & \cdots & 0 \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_q & & \vdots \\ \vdots & \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_q & 0 \\ \gamma_q & \cdots & \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_q \\ 0 & \gamma_q & \cdots & \gamma_1 & \gamma_0 & \gamma_1 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \gamma_1 & \gamma_0 & \gamma_1 \\ 0 & \cdots & 0 & \gamma_q & \cdots & \gamma_1 & \gamma_0 \end{bmatrix}. \qquad [5.5.6]$$

The row i, column j element of $\mathbf{\Omega}$ is given by $\gamma_{|i-j|}$, where γ_k is the kth autocovariance of an $MA(q)$ process:

$$\gamma_k = \begin{cases} \sigma^2(\theta_k + \theta_{k+1}\theta_1 + \theta_{k+2}\theta_2 + \cdots + \theta_q\theta_{q-k}) & \text{for } k = 0, 1, \ldots, q \\ 0 & \text{for } k > q, \end{cases} \tag{5.5.7}$$

where $\theta_0 \equiv 1$. Again, the exact likelihood function [5.5.5] can be evaluated using either the Kalman filter of Chapter 13 or the triangular factorization of $\mathbf{\Omega}$,

$$\mathbf{\Omega} = \mathbf{ADA}', \tag{5.5.8}$$

where $\mathbf{A}$ is the lower triangular matrix given by [4.4.11] and $\mathbf{D}$ is the diagonal matrix given by [4.4.7]. Note that the band structure of $\mathbf{\Omega}$ in [5.5.6] makes $\mathbf{A}$ and $\mathbf{D}$ simple to calculate. After the first $(q + 1)$ rows, all the subsequent entries in the first column of $\mathbf{\Omega}$ are already zero, so no multiple of the first row need be added to make these zero. Hence, $a_{i1} = 0$ for $i > q + 1$. Similarly, beyond the first $(q + 2)$ rows of the second column, no multiple of the second row need be added to make these entries zero, meaning that $a_{i2} = 0$ for $i > q + 2$. Thus $\mathbf{A}$ is a lower triangular band matrix with $a_{ij} = 0$ for $i > q + j$:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 1 & 0 & \cdots & 0 & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ a_{q+1,1} & a_{q+1,2} & a_{q+1,3} & \cdots & 0 & 0 \\ 0 & a_{q+2,2} & a_{q+2,3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{T,T-1} & 1 \end{bmatrix}.$$

A computer can be programmed to calculate these matrices quickly for a given numerical value for $\boldsymbol{\theta}$.

Substituting [5.5.8] into [5.5.5], the exact likelihood function for a Gaussian $MA(q)$ process can be written as in [5.4.10]:

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2}|\mathbf{D}|^{-1/2} \exp[-\tfrac{1}{2}\tilde{\mathbf{y}}'\mathbf{D}^{-1}\tilde{\mathbf{y}}]$$

where

$$\mathbf{A}\tilde{\mathbf{y}} = \mathbf{y} - \boldsymbol{\mu}. \tag{5.5.9}$$

The elements of $\tilde{\mathbf{y}}$ can be calculated recursively by working down the rows of [5.5.9]:

$$\begin{aligned} \tilde{y}_1 &= y_1 - \mu \\ \tilde{y}_2 &= (y_2 - \mu) - a_{21}\tilde{y}_1 \\ \tilde{y}_3 &= (y_3 - \mu) - a_{32}\tilde{y}_2 - a_{31}\tilde{y}_1 \\ &\vdots \\ \tilde{y}_t &= (y_t - \mu) - a_{t,t-1}\tilde{y}_{t-1} - a_{t,t-2}\tilde{y}_{t-2} - \cdots - a_{t,t-q}\tilde{y}_{t-q}. \end{aligned}$$

The exact log likelihood function can then be calculated as in [5.4.16]:

$$\mathcal{L}(\boldsymbol{\theta}) = \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(d_{tt}) - \frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}. \tag{5.5.10}$$

5.6. *The Likelihood Function for a Gaussian* ARMA(p, q) *Process*

Conditional Likelihood Function

A Gaussian $ARMA(p, q)$ process takes the form

$$\begin{aligned} Y_t = c &+ \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \\ &+ \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \end{aligned} \tag{5.6.1}$$

where $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. The goal is to estimate the vector of population parameters $\boldsymbol{\theta} = (c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2)'$.

The approximation to the likelihood function for an autoregression conditioned on initial values of the y's. The approximation to the likelihood function for a moving average process conditioned on initial values of the ε's. A common approximation to the likelihood function for an $ARMA(p, q)$ process conditions on both y's and ε's.

Taking initial values for $\mathbf{y}_0 \equiv (y_0, y_{-1}, \ldots, y_{-p+1})'$ and $\boldsymbol{\varepsilon}_0 \equiv (\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1})'$ as given, the sequence $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T\}$ can be calculated from $\{y_1, y_2, \ldots, y_T\}$ by iterating on

$$\begin{aligned} \varepsilon_t = y_t &- c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} \\ &- \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \end{aligned} \tag{5.6.2}$$

for $t = 1, 2, \ldots, T$. The conditional log likelihood is then

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \log f_{Y_T, Y_{T-1}, \ldots, Y_1 | \mathbf{y}_0, \boldsymbol{\varepsilon}_0}(y_T, y_{T-1}, \ldots, y_1 | \mathbf{y}_0, \boldsymbol{\varepsilon}_0; \boldsymbol{\theta}) \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}. \end{aligned} \tag{5.6.3}$$

One option is to set initial y's and ε's equal to their expected values. That is, set $y_s = c/(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$ for $s = 0, -1, \ldots, -p + 1$ and set $\varepsilon_s = 0$ for $s = 0, -1, \ldots, -q + 1$, and then proceed with the iteration in [5.6.2] for $t = 1, 2, \ldots, T$. Alternatively, Box and Jenkins (1976, p. 211) recommended setting ε's to zero but y's equal to their actual values. Thus, iteration on [5.6.2] is started at date $t = p + 1$ with $y_1, y_2, \ldots, y_p$ set to the observed values and

$$\varepsilon_p = \varepsilon_{p-1} = \cdots = \varepsilon_{p-q+1} = 0.$$

Then the conditional likelihood calculated is

$$\begin{aligned} \log f(y_T, \ldots, y_{p+1} | y_p, \ldots, y_1, \varepsilon_p = 0, \ldots, \varepsilon_{p-q+1} = 0) \\ = -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}. \end{aligned}$$

As in the case for the moving average processes, these approximations should be used only if all values of z satisfying

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

lie outside the unit circle.

Alternative Algorithms

The simplest approach to calculating the exact likelihood function for a Gaussian $ARMA$ process is to use the Kalman filter described in Chapter 13. For more

details on exact and approximate maximum likelihood estimation of *ARMA* models, see Galbraith and Galbraith (1974), Box and Jenkins (1976, Chapter 6), Hannan and Rissanen (1982), and Koreisha and Pukkila (1989).

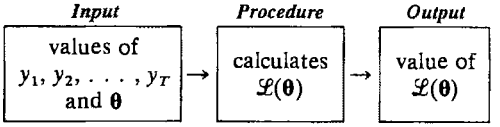
5.7. *Numerical Optimization*

Previous sections of this chapter have shown how to calculate the log likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \log f_{Y_T, Y_{T-1}, \ldots, Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta}) \qquad [5.7.1]$$

for various specifications of the process thought to have generated the observed data $y_1, y_2, \ldots, y_T$. Given the observed data, the formulas given could be used to calculate the value of $\mathcal{L}(\boldsymbol{\theta})$ for any given numerical value of $\boldsymbol{\theta}$.

This section discusses how to find the value of $\hat{\boldsymbol{\theta}}$ that maximizes $\mathcal{L}(\boldsymbol{\theta})$ given no more knowledge than this ability to calculate the value of $\mathcal{L}(\boldsymbol{\theta})$ for any particular value of $\boldsymbol{\theta}$. The general approach is to write a procedure that enables a computer to calculate the numerical value of $\mathcal{L}(\boldsymbol{\theta})$ for any particular numerical values for $\boldsymbol{\theta}$ and the observed data $y_1, y_2, \ldots, y_T$. We can think of this procedure as a "black box" that enables us to guess some value of $\boldsymbol{\theta}$ and see what the resulting value of $\mathcal{L}(\boldsymbol{\theta})$ would be:



The idea will be to make a series of different guesses for $\boldsymbol{\theta}$, compare the value of $\mathcal{L}(\boldsymbol{\theta})$ for each guess, and try to infer from these values for $\mathcal{L}(\boldsymbol{\theta})$ the value $\hat{\boldsymbol{\theta}}$ for which $\mathcal{L}(\boldsymbol{\theta})$ is largest. Such methods are described as *numerical maximization*.

Grid Search

The simplest approach to numerical maximization is known as the *grid search* method. To illustrate this approach, suppose we have data generated by an $AR(1)$ process, for which the log likelihood was seen to be given by [5.2.9]. To keep the example very simple, it is assumed to be known that the mean of the process is zero ($c = 0$) and that the innovations have unit variance ($\sigma^2 = 1$). Thus the only unknown parameter is the autoregressive coefficient ϕ, and [5.2.9] simplifies to

$$\begin{aligned}\mathcal{L}(\phi) = &-\frac{T}{2}\log(2\pi) + \frac{1}{2}\log(1 - \phi^2) \\ &- \frac{1}{2}(1 - \phi^2)y_1^2 - \frac{1}{2}\sum_{t=2}^{T}(y_t - \phi y_{t-1})^2.\end{aligned} \qquad [5.7.2]$$

Suppose that the observed sample consists of the following $T = 5$ observations:

$$y_1 = 0.8 \quad y_2 = 0.2 \quad y_3 = -1.2 \quad y_4 = -0.4 \quad y_5 = 0.0.$$

If we make an arbitrary guess as to the value of ϕ, say, $\phi = 0.0$, and plug this guess into expression [5.7.2], we calculate that $\mathcal{L}(\phi) = -5.73$ at $\phi = .0.0$. Trying another guess ($\phi = 0.1$), we calculate $\mathcal{L}(\phi) = -5.71$ at $\phi = 0.1$—the log likelihood is higher at $\phi = 0.1$ than at $\phi = 0.0$. Continuing in this fashion, we could calculate the value of $\mathcal{L}(\phi)$ for every value of ϕ between -0.9 and $+0.9$ in increments of

0.1. The results are reported in Figure 5.1. It appears from these calculations that the log likelihood function $\mathcal{L}(\phi)$ is nicely behaved with a unique maximum at some value of ϕ between 0.1 and 0.3. We could then focus on this subregion of the parameter space and evaluate $\mathcal{L}(\phi)$ at a finer grid, calculating the value of $\mathcal{L}(\phi)$ for all values of ϕ between 0.1 and 0.3 in increments of 0.02. Proceeding in this fashion, it should be possible to get arbitrarily close to the value of ϕ that maximizes $\mathcal{L}(\phi)$ by making the grid finer and finer.

Note that this procedure does not find the *exact MLE* $\hat{\phi}$, but instead approximates it with any accuracy desired. In general, this will be the case with any numerical maximization algorithm. To use these algorithms we therefore have to specify a *convergence criterion*, or some way of deciding when we are close enough to the true maximum. For example, suppose we want an estimate $\hat{\phi}$ that differs from the true *MLE* by no more than ± 0.0001. Then we would continue refining the grid until the increments are in steps of 0.0001, and the best estimate among the elements of that grid would be the numerical *MLE* of ϕ.

For the simple $AR(1)$ example in Figure 5.1, the log likelihood function is *unimodal*—there is a unique value $\boldsymbol{\theta}$ for which $\partial\mathcal{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$. For a general numerical maximization problem, this need not be the case. For example, suppose that we are interested in estimating a scalar parameter θ for which the log likelihood function is as displayed in Figure 5.2. The value $\theta = -0.6$ is a *local maximum*, meaning that the likelihood function is higher there than for any other θ in a neighborhood around $\theta = -0.6$. However, the *global maximum* occurs around $\theta = 0.2$. The grid search method should work well for a unimodal likelihood as long as $\mathcal{L}(\boldsymbol{\theta})$ is continuous. When there are multiple local maxima, the grid must be sufficiently fine to reveal all of the local "hills" on the likelihood surface.

Steepest Ascent

Grid search can be a very good method when there is a single unknown parameter to estimate. However, it quickly becomes intractable when the number of elements of $\boldsymbol{\theta}$ becomes large. An alternative numerical method that often suc-

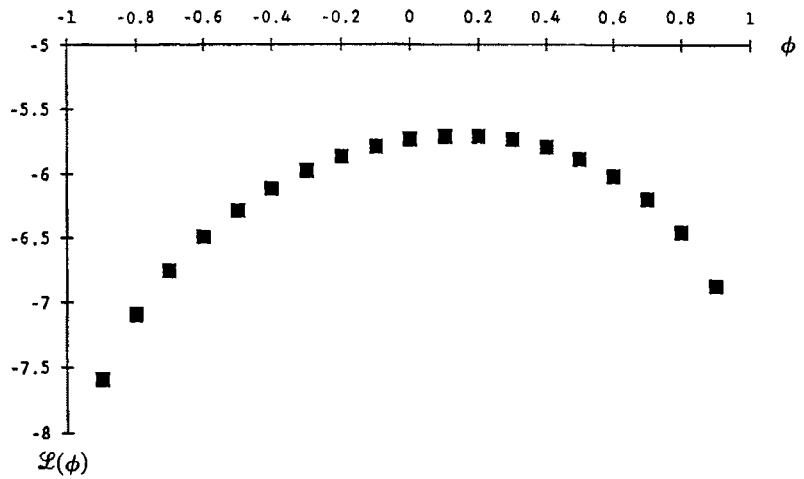


FIGURE 5.1 Log likelihood for an $AR(1)$ process for various guesses of ϕ.

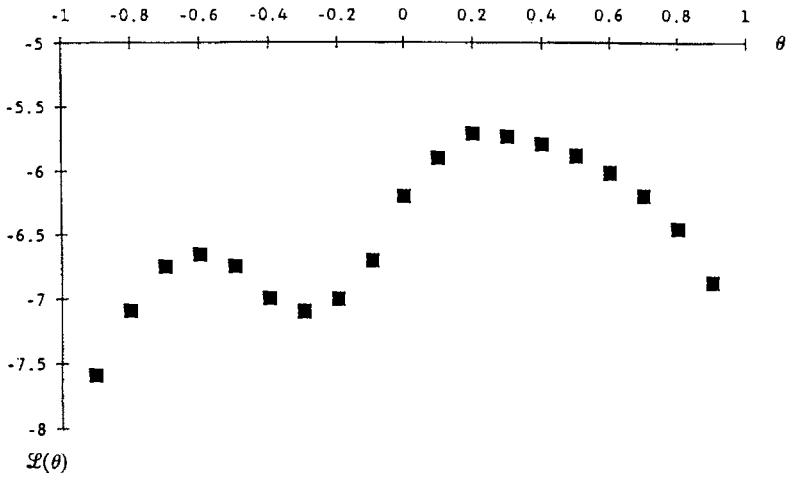


FIGURE 5.2 Bimodal log likelihood function.

ceeds in maximizing a continuously differentiable function of a large number of parameters is known as *steepest ascent*.

To understand this approach, let us temporarily disregard the "black box" nature of the investigation and instead examine how we would proceed analytically with a particular maximization problem. Suppose we have an initial estimate of the parameter vector, denoted $\boldsymbol{\theta}^{(0)}$, and wish to come up with a better estimate $\boldsymbol{\theta}^{(1)}$. Imagine that we are constrained to choose $\boldsymbol{\theta}^{(1)}$ so that the squared distance between $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\theta}^{(1)}$ is some fixed number k:

$$\{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\}'\{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\} = k.$$

The optimal value to choose for $\boldsymbol{\theta}^{(1)}$ would then be the solution to the following constrained maximization problem:

$$\max_{\boldsymbol{\theta}^{(1)}} \mathcal{L}(\boldsymbol{\theta}^{(1)}) \qquad \text{subject to} \quad \{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\}'\{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\} = k.$$

To characterize the solution to this problem,[3] form the Lagrangean,

$$J(\boldsymbol{\theta}^{(1)}) = \mathcal{L}(\boldsymbol{\theta}^{(1)}) + \lambda[k - \{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\}'\{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\}], \qquad [5.7.3]$$

where λ denotes a Lagrange multiplier. Differentiating [5.7.3] with respect to $\boldsymbol{\theta}^{(1)}$ and setting the result equal to zero yields

$$\left.\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(1)}} - (2\lambda)\{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\} = \mathbf{0}. \qquad [5.7.4]$$

Let $\mathbf{g}(\boldsymbol{\theta})$ denote the gradient vector of the log likelihood function:

$$\mathbf{g}(\boldsymbol{\theta}) \equiv \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

If there are a elements of $\boldsymbol{\theta}$, then $\mathbf{g}(\boldsymbol{\theta})$ is an $(a \times 1)$ vector whose ith element represents the derivative of the log likelihood with respect to the ith element of $\boldsymbol{\theta}$.

[3]See Chiang (1974) for an introduction to the use of Lagrange multipliers for solving a constrained optimization problem.

Using this notation, expression [5.7.4] can be written as

$$\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)} = [1/(2\lambda)] \cdot \mathbf{g}(\boldsymbol{\theta}^{(1)}). \qquad [5.7.5]$$

Expression [5.7.5] asserts that if we are allowed to change $\boldsymbol{\theta}$ by only a fixed amount, the biggest increase in the log likelihood function will be achieved if the change in $\boldsymbol{\theta}$ (the magnitude $\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}$) is chosen to be a constant $1/(2\lambda)$ times the gradient vector $\mathbf{g}(\boldsymbol{\theta}^{(1)})$. If we are contemplating a very small step (so that k is near zero), the value $\mathbf{g}(\boldsymbol{\theta}^{(1)})$ will approach $\mathbf{g}(\boldsymbol{\theta}^{(0)})$. In other words, the gradient vector $\mathbf{g}(\boldsymbol{\theta}^{(0)})$ gives the direction in which the log likelihood function increases most steeply from $\boldsymbol{\theta}^{(0)}$.

For illustration, suppose that $a = 2$ and let the log likelihood be

$$\mathcal{L}(\boldsymbol{\theta}) = -1.5\theta_1^2 - 2\theta_2^2. \qquad [5.7.6]$$

We can easily see analytically for this example that the *MLE* is given by $\hat{\boldsymbol{\theta}} = (0, 0)'$. Let us nevertheless use this example to illustrate how the method of steepest ascent works. The elements of the gradient vector are

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_1} = -3\theta_1 \qquad \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_2} = -4\theta_2. \qquad [5.7.7]$$

Suppose that the initial guess is $\boldsymbol{\theta}^{(0)} = (-1, 1)'$. Then

$$\left.\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_1}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}} = 3 \qquad \left.\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_2}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}} = -4.$$

An increase in θ_1 would increase the likelihood, while an increase in θ_2 would decrease the likelihood. The gradient vector evaluated at $\boldsymbol{\theta}^{(0)}$ is

$$\mathbf{g}(\boldsymbol{\theta}^{(0)}) = \begin{bmatrix} 3 \\ -4 \end{bmatrix},$$

so that the optimal step $\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}$ should be proportional to $(3, -4)'$. For example, with $k = 1$ we would choose

$$\theta_1^{(1)} - \theta_1^{(0)} = \tfrac{3}{5}$$
$$\theta_2^{(1)} - \theta_2^{(0)} = -\tfrac{4}{5};$$

that is, the new guesses would be $\theta_1^{(1)} = -0.4$ and $\theta_2^{(1)} = 0.2$. To increase the likelihood by the greatest amount, we want to increase θ_1 and decrease θ_2 relative to their values at the initial guess $\boldsymbol{\theta}^{(0)}$. Since a one-unit change in θ_2 has a bigger effect on $\mathcal{L}(\boldsymbol{\theta})$ than would a one-unit change in θ_1, the change in θ_2 is larger in absolute value than the change in θ_1.

Let us now return to the black box perspective, where the only capability we have is to calculate the value of $\mathcal{L}(\boldsymbol{\theta})$ for a specified numerical value of $\boldsymbol{\theta}$. We might start with an arbitrary initial guess for the value of $\boldsymbol{\theta}$, denoted $\boldsymbol{\theta}^{(0)}$. Suppose we then calculate the value of the gradient vector at $\boldsymbol{\theta}^{(0)}$:

$$\mathbf{g}(\boldsymbol{\theta}^{(0)}) = \left.\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}}. \qquad [5.7.8]$$

This gradient could in principle be calculated analytically, by differentiating the general expression for $\mathcal{L}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and writing a computer procedure to calculate each element of $\mathbf{g}(\boldsymbol{\theta})$ given the data and a numerical value for $\boldsymbol{\theta}$. For example, expression [5.7.7] could be used to calculate $\mathbf{g}(\boldsymbol{\theta})$ for any particular value of $\boldsymbol{\theta}$. Alternatively, if it is too hard to differentiate $\mathcal{L}(\boldsymbol{\theta})$ analytically, we can always

get a numerical approximation to the gradient by seeing how $\mathcal{L}(\boldsymbol{\theta})$ changes for a small change in each element of $\boldsymbol{\theta}$. In particular, the ith element of $\mathbf{g}(\boldsymbol{\theta}^{(0)})$ might be approximated by

$$g_i(\boldsymbol{\theta}^{(0)}) \cong \frac{1}{\Delta}[\mathcal{L}(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_{i-1}^{(0)}, \theta_i^{(0)} + \Delta, \theta_{i+1}^{(0)}, \theta_{i+2}^{(0)}, \ldots, \theta_a^{(0)}) - \mathcal{L}(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_{i-1}^{(0)}, \theta_i^{(0)}, \theta_{i+1}^{(0)}, \theta_{i+2}^{(0)}, \ldots, \theta_a^{(0)})\}, \quad [5.7.9]$$

where Δ represents some arbitrarily chosen small scalar such as $\Delta = 10^{-6}$. By numerically calculating the value of $\mathcal{L}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^{(0)}$ and at a different values of $\boldsymbol{\theta}$ corresponding to small changes in each of the individual elements of $\boldsymbol{\theta}^{(0)}$, an estimate of the full vector $\mathbf{g}(\boldsymbol{\theta}^{(0)})$ can be uncovered.

Result [5.7.5] suggests that we should change the value of $\boldsymbol{\theta}$ in the direction of the gradient, choosing

$$\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)} = s \cdot \mathbf{g}(\boldsymbol{\theta}^{(0)})$$

for some positive scalar s. A suitable choice for s could be found by an adaptation of the grid search method. For example, we might calculate the value of $\mathcal{L}\{\boldsymbol{\theta}^{(0)} + s \cdot \mathbf{g}(\boldsymbol{\theta}^{(0)})\}$ for $s = \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8$, and 16 and choose as the new estimate $\boldsymbol{\theta}^{(1)}$ the value of $\boldsymbol{\theta}^{(0)} + s \cdot \mathbf{g}(\boldsymbol{\theta}^{(0)})$ for which $\mathcal{L}(\boldsymbol{\theta})$ is largest. Smaller or larger values of s could also be explored if the maximum appears to be at one of the extremes. If none of the values of s improves the likelihood, then a very small value for s such as the value $\Delta = 10^{-6}$ used to approximate the derivative should be tried.

We can then repeat the process, taking $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + s \cdot \mathbf{g}(\boldsymbol{\theta}^{(0)})$ as the starting point, evaluating the gradient at the new location $\mathbf{g}(\boldsymbol{\theta}^{(1)})$, and generating a new estimate $\boldsymbol{\theta}^{(2)}$ according to

$$\boldsymbol{\theta}^{(2)} = \boldsymbol{\theta}^{(1)} + s \cdot \mathbf{g}(\boldsymbol{\theta}^{(1)})$$

for the best choice of s. The process is iterated, calculating

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + s \cdot \mathbf{g}(\boldsymbol{\theta}^{(m)})$$

for $m = 0, 1, 2, \ldots$ until some convergence criterion is satisfied, such as that the gradient vector $\mathbf{g}(\boldsymbol{\theta}^{(m)})$ is within some specified tolerance of zero, the distance between $\boldsymbol{\theta}^{(m+1)}$ and $\boldsymbol{\theta}^{(m)}$ is less than some specified threshold, or the change between $\mathcal{L}(\boldsymbol{\theta}^{(m+1)})$ and $\mathcal{L}(\boldsymbol{\theta}^{(m)})$ is smaller than some desired amount.

Figure 5.3 illustrates the method of steepest ascent when $\boldsymbol{\theta}$ contains $a = 2$ elements. The figure displays contour lines for the log likelihood $\mathcal{L}(\boldsymbol{\theta})$; along a given contour, the log likelihood $\mathcal{L}(\boldsymbol{\theta})$ is constant. If the iteration is started at the initial guess $\boldsymbol{\theta}^{(0)}$, the gradient $\mathbf{g}(\boldsymbol{\theta}^{(0)})$ describes the direction of steepest ascent. Finding the optimal step in that direction produces the new estimate $\boldsymbol{\theta}^{(1)}$. The gradient at that point $\mathbf{g}(\boldsymbol{\theta}^{(1)})$ then determines a new search direction on which a new estimate $\boldsymbol{\theta}^{(2)}$ is based, until the top of the hill is reached.

Figure 5.3 also illustrates a multivariate generalization of the problem with multiple local maxima seen earlier in Figure 5.2. The procedure should converge to a local maximum, which in this case is different from the global maximum $\boldsymbol{\theta}^*$. In Figure 5.3, it appears that if $\boldsymbol{\theta}^{(0)*}$ were used to begin the iteration in place of $\boldsymbol{\theta}^{(0)}$, the procedure would converge to the true global maximum $\boldsymbol{\theta}^*$. In practice, the only way to ensure that a global maximum is found is to begin the iteration from a number of different starting values for $\boldsymbol{\theta}^{(0)}$ and to continue the sequence from each starting value until the top of the hill associated with that starting value is discovered.

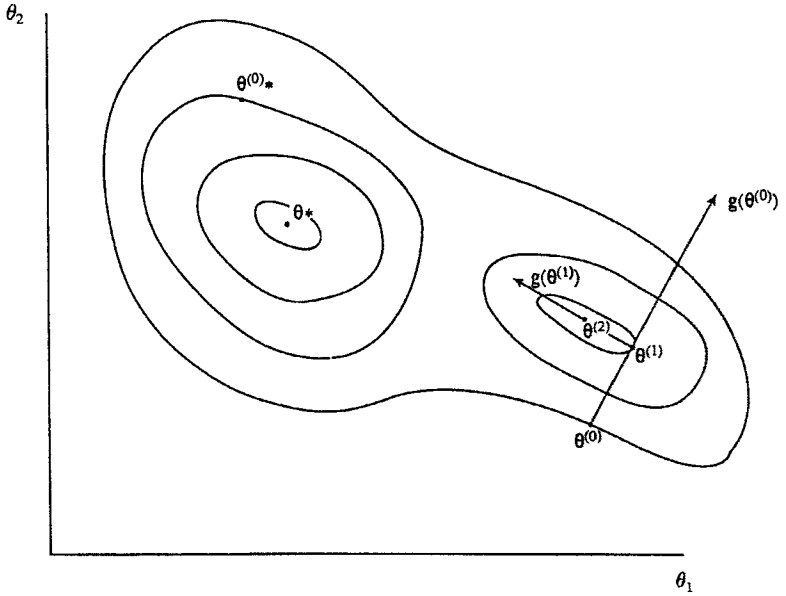


FIGURE 5.3 Likelihood contours and maximization by steepest ascent.

Newton-Raphson

One drawback to the steepest-ascent method is that it may require a very large number of iterations to close in on the local maximum. An alternative method known as *Newton-Raphson* often converges more quickly provided that (1) second derivatives of the log likelihood function $\mathcal{L}(\boldsymbol{\theta})$ exist and (2) the function $\mathcal{L}(\boldsymbol{\theta})$ is concave, meaning that -1 times the matrix of second derivatives is everywhere positive definite.

Suppose that $\boldsymbol{\theta}$ is an $(a \times 1)$ vector of parameters to be estimated. Let $\mathbf{g}(\boldsymbol{\theta}^{(0)})$ denote the gradient vector of the log likelihood function at $\boldsymbol{\theta}^{(0)}$:

$$\underset{(a\times 1)}{\mathbf{g}(\boldsymbol{\theta}^{(0)})} = \left.\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}};$$

and let $\mathbf{H}(\boldsymbol{\theta}^{(0)})$ denote -1 times the matrix of second derivatives of the log likelihood function:

$$\underset{(a\times a)}{\mathbf{H}(\boldsymbol{\theta}^{(0)})} = -\left.\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}}.$$

Consider approximating $\mathcal{L}(\boldsymbol{\theta})$ with a second-order Taylor series around $\boldsymbol{\theta}^{(0)}$:

$$\mathcal{L}(\boldsymbol{\theta}) \cong \mathcal{L}(\boldsymbol{\theta}^{(0)}) + [\mathbf{g}(\boldsymbol{\theta}^{(0)})]'[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}] - \tfrac{1}{2}[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}]'\mathbf{H}(\boldsymbol{\theta}^{(0)})[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}]. \quad [5.7.10]$$

The idea behind the Newton-Raphson method is to choose $\boldsymbol{\theta}$ so as to maximize [5.7.10]. Setting the derivative of [5.7.10] with respect to $\boldsymbol{\theta}$ equal to zero results in

$$\mathbf{g}(\boldsymbol{\theta}^{(0)}) - \mathbf{H}(\boldsymbol{\theta}^{(0)})[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}] = \mathbf{0}. \quad [5.7.11]$$

Let $\boldsymbol{\theta}^{(0)}$ denote an initial guess as to the value of $\boldsymbol{\theta}$. One can calculate the derivative of the log likelihood at that initial guess ($\mathbf{g}(\boldsymbol{\theta}^{(0)})$) either analytically, as in [5.7.7], or numerically, as in [5.7.9]. One can also use analytical or numerical methods to calculate the negative of the matrix of second derivatives at the initial guess ($\mathbf{H}(\boldsymbol{\theta}^{(0)})$). Expression [5.7.11] suggests that an improved estimate of $\boldsymbol{\theta}$ (denoted $\boldsymbol{\theta}^{(1)}$) would satisfy

$$\mathbf{g}(\boldsymbol{\theta}^{(0)}) = \mathbf{H}(\boldsymbol{\theta}^{(0)})[\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}]$$

or

$$\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)} = [\mathbf{H}(\boldsymbol{\theta}^{(0)})]^{-1}\mathbf{g}(\boldsymbol{\theta}^{(0)}). \qquad [5.7.12]$$

One could next calculate the gradient and Hessian at $\boldsymbol{\theta}^{(1)}$ and use these to find a new estimate $\boldsymbol{\theta}^{(2)}$ and continue iterating in this fashion. The mth step in the iteration updates the estimate of $\boldsymbol{\theta}$ by using the formula

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + [\mathbf{H}(\boldsymbol{\theta}^{(m)})]^{-1}\mathbf{g}(\boldsymbol{\theta}^{(m)}). \qquad [5.7.13]$$

If the log likelihood function happens to be a perfect quadratic function, then [5.7.10] holds exactly and [5.7.12] will generate the exact *MLE* in a single step:

$$\boldsymbol{\theta}^{(1)} = \hat{\boldsymbol{\theta}}_{MLE}.$$

If the quadratic approximation is reasonably good, Newton-Raphson should converge to the local maximum more quickly than the steepest-ascent method. However, if the likelihood function is not concave, Newton-Raphson behaves quite poorly. Thus, steepest ascent is often slower to converge but sometimes proves to be more robust compared with Newton-Raphson.

Since [5.7.10] is usually only an approximation to the true log likelihood function, the iteration on [5.7.13] is often modified as follows. Expression [5.7.13] is taken to suggest the search direction. The value of the log likelihood function at several points in that direction is then calculated, and the best value determines the length of the step. This strategy calls for replacing [5.7.13] by

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + s[\mathbf{H}(\boldsymbol{\theta}^{(m)})]^{-1}\mathbf{g}(\boldsymbol{\theta}^{(m)}), \qquad [5.7.14]$$

where s is a scalar controlling the step length. One calculates $\boldsymbol{\theta}^{(m+1)}$ and the associated value for the log likelihood $\mathcal{L}(\boldsymbol{\theta}^{(m+1)})$ for various values of s in [5.7.14] and chooses as the estimate $\boldsymbol{\theta}^{(m+1)}$ the value that produces the biggest value for the log likelihood.

Davidon-Fletcher-Powell

If $\boldsymbol{\theta}$ contains a unknown parameters, then the symmetric matrix $\mathbf{H}(\boldsymbol{\theta})$ has $a(a+1)/2$ separate elements. Calculating all these elements can be extremely time-consuming if a is large. An alternative approach reasons as follows. The matrix of second derivatives ($-\mathbf{H}(\boldsymbol{\theta})$) corresponds to the first derivatives of the gradient vector ($\mathbf{g}(\boldsymbol{\theta})$), which tell us how $\mathbf{g}(\boldsymbol{\theta})$ changes as $\boldsymbol{\theta}$ changes. We get some independent information about this by comparing $\mathbf{g}(\boldsymbol{\theta}^{(1)}) - \mathbf{g}(\boldsymbol{\theta}^{(0)})$ with $\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}$. This is not enough information by itself to estimate $\mathbf{H}(\boldsymbol{\theta})$, but it is information that could be used to update an initial guess about the value of $\mathbf{H}(\boldsymbol{\theta})$. Thus, rather than evaluate $\mathbf{H}(\boldsymbol{\theta})$ directly at each iteration, the idea will be to start with an initial guess about $\mathbf{H}(\boldsymbol{\theta})$ and update the guess solely on the basis of how much $\mathbf{g}(\boldsymbol{\theta})$ changes between iterations, given the magnitude of the change in $\boldsymbol{\theta}$. Such methods are sometimes described as *modified Newton-Raphson*.

One of the most popular modified Newton-Raphson methods was proposed by Davidon (1959) and Fletcher and Powell (1963). Since it is $\mathbf{H}^{-1}$ rather than $\mathbf{H}$

itself that appears in the updating formula [5.7.14], the Davidon-Fletcher-Powell algorithm updates an estimate of $\mathbf{H}^{-1}$ at each step on the basis of the size of the change in $\mathbf{g}(\boldsymbol{\theta})$ relative to the change in $\boldsymbol{\theta}$. Specifically, let $\boldsymbol{\theta}^{(m)}$ denote an estimate of $\boldsymbol{\theta}$ that has been calculated at the mth iteration, and let $\mathbf{A}^{(m)}$ denote an estimate of $[\mathbf{H}(\boldsymbol{\theta}^{(m)})]^{-1}$. The new estimate $\boldsymbol{\theta}^{(m+1)}$ is given by

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}) \qquad [5.7.15]$$

for s the positive scalar that maximizes $\mathcal{L}\{\boldsymbol{\theta}^{(m)} + s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)})\}$. Once $\boldsymbol{\theta}^{(m+1)}$ and the gradient at $\boldsymbol{\theta}^{(m+1)}$ have been calculated, a new estimate $\mathbf{A}^{(m+1)}$ is found from

$$\mathbf{A}^{(m+1)} = \mathbf{A}^{(m)} - \frac{\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}}{(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})} - \frac{(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})} \qquad [5.7.16]$$

where

$$\Delta\boldsymbol{\theta}^{(m+1)} \equiv \boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}$$
$$\Delta\mathbf{g}^{(m+1)} \equiv \mathbf{g}(\boldsymbol{\theta}^{(m+1)}) - \mathbf{g}(\boldsymbol{\theta}^{(m)}).$$

In what sense should $\mathbf{A}^{(m+1)}$ as calculated from [5.7.16] be regarded as an estimate of the inverse of $\mathbf{H}(\boldsymbol{\theta}^{(m+1)})$? Consider first the case when $\boldsymbol{\theta}$ is a scalar ($a = 1$). Then [5.7.16] simplifies to

$$\begin{aligned} A^{(m+1)} &= A^{(m)} - \frac{(A^{(m)})^2(\Delta g^{(m+1)})^2}{(\Delta g^{(m+1)})^2(A^{(m)})} - \frac{(\Delta\theta^{(m+1)})^2}{(\Delta g^{(m+1)})(\Delta\theta^{(m+1)})} \\ &= A^{(m)} - A^{(m)} - \frac{\Delta\theta^{(m+1)}}{\Delta g^{(m+1)}} \\ &= -\frac{\Delta\theta^{(m+1)}}{\Delta g^{(m+1)}}. \end{aligned}$$

In this case,

$$[A^{(m+1)}]^{-1} = -\frac{\Delta g^{(m+1)}}{\Delta\theta^{(m+1)}},$$

which is the natural discrete approximation to

$$H(\theta^{(m+1)}) = -\frac{\partial^2\mathcal{L}}{\partial\theta^2}\bigg|_{\theta=\theta^{(m+1)}} = -\frac{\partial g}{\partial\theta}\bigg|_{\theta=\theta^{(m+1)}}.$$

More generally (for $a > 1$), an estimate of the derivative of $\mathbf{g}(\cdot)$ should be related to the observed change in $\mathbf{g}(\cdot)$ according to

$$\mathbf{g}(\boldsymbol{\theta}^{(m+1)}) \cong \mathbf{g}(\boldsymbol{\theta}^{(m)}) + \frac{\partial\mathbf{g}}{\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m+1)}}[\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}].$$

That is,

$$\mathbf{g}(\boldsymbol{\theta}^{(m+1)}) \cong \mathbf{g}(\boldsymbol{\theta}^{(m)}) - \mathbf{H}(\boldsymbol{\theta}^{(m+1)})[\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}]$$

or

$$\Delta\boldsymbol{\theta}^{(m+1)} \cong -[\mathbf{H}(\boldsymbol{\theta}^{(m+1)})]^{-1}\,\Delta\mathbf{g}^{(m+1)}.$$

Hence an estimate $\mathbf{A}^{(m+1)}$ of $[\mathbf{H}(\boldsymbol{\theta}^{(m+1)})]^{-1}$ should satisfy

$$\mathbf{A}^{(m+1)}\,\Delta\mathbf{g}^{(m+1)} = -\Delta\boldsymbol{\theta}^{(m+1)}. \qquad [5.7.17]$$

Postmultiplication of [5.7.16] by $\Delta\mathbf{g}^{(m+1)}$ confirms that [5.7.17] is indeed satisfied by the Davidon-Fletcher-Powell estimate $\mathbf{A}^{(m+1)}$:

$$\begin{aligned}\mathbf{A}^{(m+1)}\,\Delta\mathbf{g}^{(m+1)} &= \mathbf{A}^{(m)}\,\Delta\mathbf{g}^{(m+1)} \\ &\quad - \frac{\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})}{(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})} \\ &\quad - \frac{(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'(\Delta\mathbf{g}^{(m+1)})}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})} \\ &= \mathbf{A}^{(m)}\,\Delta\mathbf{g}^{(m+1)} - \mathbf{A}^{(m)}\,\Delta\mathbf{g}^{(m+1)} - \Delta\boldsymbol{\theta}^{(m+1)} \\ &= -\Delta\boldsymbol{\theta}^{(m+1)}.\end{aligned}$$

Thus, calculation of [5.7.16] produces an estimate of $[\mathbf{H}(\boldsymbol{\theta}^{(m+1)})]^{-1}$ that is consistent with the magnitude of the observed change between $\mathbf{g}(\boldsymbol{\theta}^{(m+1)})$ and $\mathbf{g}(\boldsymbol{\theta}^{(m)})$ given the size of the change between $\boldsymbol{\theta}^{(m+1)}$ and $\boldsymbol{\theta}^{(m)}$.

The following proposition (proved in Appendix 5.A at the end of the chapter) establishes some further useful properties of the updating formula [5.7.16].

Proposition 5.1: *(Fletcher and Powell (1963)). Consider $\mathcal{L}(\boldsymbol{\theta})$, where $\mathcal{L}$: $\mathbb{R}^a \to \mathbb{R}^1$ has continuous first derivatives denoted*

$$\underset{(a\times 1)}{\mathbf{g}(\boldsymbol{\theta}^{(m)})} = \left.\frac{\partial\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}}.$$

Suppose that some element of $\mathbf{g}(\boldsymbol{\theta}^{(m)})$ is nonzero, and let $\mathbf{A}^{(m)}$ be a positive definite symmetric $(a \times a)$ matrix. Then the following hold.

(a) There exists a scalar $s > 0$ such that $\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) > \mathcal{L}(\boldsymbol{\theta}^{(m)})$ for

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}). \qquad [5.7.18]$$

(b) If s in [5.7.18] is chosen so as to maximize $\mathcal{L}(\boldsymbol{\theta}^{(m+1)})$, then the first-order conditions for an interior maximum imply that

$$[\mathbf{g}(\boldsymbol{\theta}^{(m+1)})]'[\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}] = 0. \qquad [5.7.19]$$

(c) Provided that [5.7.19] holds and that some element of $\mathbf{g}(\boldsymbol{\theta}^{(m+1)}) - \mathbf{g}(\boldsymbol{\theta}^{(m)})$ is nonzero, then $\mathbf{A}^{(m+1)}$ described by [5.7.16] is a positive definite symmetric matrix.

Result (a) establishes that as long as we are not already at an optimum ($\mathbf{g}(\boldsymbol{\theta}^{(m)}) \neq \mathbf{0}$), there exists a step in the direction suggested by the algorithm that will increase the likelihood further, provided that $\mathbf{A}^{(m)}$ is a positive definite matrix. Result (c) establishes that provided that the iteration is begun with $\mathbf{A}^{(0)}$ a positive definite matrix, then the sequence of matrices $\{\mathbf{A}^{(m)}\}_{m=1}^{N}$ should all be positive definite, meaning that each step of the iteration should increase the likelihood function. A standard procedure is to start the iteration with $\mathbf{A}^{(0)} = \mathbf{I}_a$, the $(a \times a)$ identity matrix.

If the function $\mathcal{L}(\boldsymbol{\theta})$ is exactly quadratic, so that

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}^{(0)}) + \mathbf{g}'[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}] - \tfrac{1}{2}[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}]'\mathbf{H}[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}],$$

with $\mathbf{H}$ positive definite, then Fletcher and Powell (1963) showed that iteration on [5.7.15] and [5.7.16] will converge to the true global maximum in a steps:

$$\boldsymbol{\theta}^{(a)} = \hat{\boldsymbol{\theta}}_{MLE} = \boldsymbol{\theta}^{(0)} + \mathbf{H}^{-1}\mathbf{g};$$

and the weighting matrix will converge to the inverse of -1 times the matrix of second derivatives:

$$\mathbf{A}^{(a)} = \mathbf{H}^{-1}.$$

More generally, if $\mathcal{L}(\boldsymbol{\theta})$ is well approximated by a quadratic function, then the Davidon-Fletcher-Powell search procedure should approach the global maximum more quickly than the steepest-ascent method,

$$\boldsymbol{\theta}^{(N)} \cong \hat{\boldsymbol{\theta}}_{MLE}$$

for large N, while $\mathbf{A}^{(m)}$ should converge to the negative of the matrix of second derivatives of the log likelihood function:

$$\mathbf{A}^{(N)} \cong -\left[\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MLE}}\right]^{-1}. \qquad [5.7.20]$$

In practice, however, the approximation in [5.7.20] can be somewhat poor, and it is better to evaluate the matrix of second derivatives numerically for purposes of calculating standard errors, as discussed in Section 5.8.

If the function $\mathcal{L}(\boldsymbol{\theta})$ is not globally concave or if the starting value $\boldsymbol{\theta}^{(0)}$ is far from the true maximum, the Davidon-Fletcher-Powell procedure can do very badly. If problems are encountered, it often helps to try a different starting value $\boldsymbol{\theta}^{(0)}$, to rescale the data or parameters so that the elements of $\boldsymbol{\theta}$ are in comparable units, or to rescale the initial matrix $\mathbf{A}^{(0)}$—for example, by setting

$$\mathbf{A}^{(0)} = (1 \times 10^{-4})\mathbf{I}_a.$$

Other Numerical Optimization Methods

A variety of other modified Newton-Raphson methods are available, which use alternative techniques for updating an estimate of $\mathbf{H}(\boldsymbol{\theta}^{(m)})$ or its inverse. Two of the more popular methods are those of Broyden (1965, 1967) and Berndt, Hall, Hall, and Hausman (1974). Surveys of these and a variety of other approaches are provided by Judge, Griffiths, Hill, and Lee (1980, pp. 719–72) and Quandt (1983).

Obviously, these same methods can be used to minimize a function $Q(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. We simply multiply the objective function by -1 and then maximize the function $-Q(\boldsymbol{\theta})$.

5.8. *Statistical Inference with Maximum Likelihood Estimation*

The previous section discussed ways to find the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ given only the numerical ability to evaluate the log likelihood function $\mathcal{L}(\boldsymbol{\theta})$. This section summarizes general approaches that can be used to test a hypothesis about $\boldsymbol{\theta}$. The section merely summarizes a number of useful results without providing any proofs. We will return to these issues in more depth in Chapter 14, where the statistical foundation behind many of these claims will be developed.

Before detailing these results, however, it is worth calling attention to two of the key assumptions behind the formulas presented in this section. First, it is assumed that the observed data are strictly stationary. Second, it is assumed that neither the estimate $\hat{\boldsymbol{\theta}}$ nor the true value $\boldsymbol{\theta}_0$ falls on a boundary of the allowable parameter space. For example, suppose that the first element of $\boldsymbol{\theta}$ is a parameter corresponding to the probability of a particular event, which must be between 0 and 1. If the event did not occur in the sample, the maximum likelihood estimate of the probability might be zero. This is an example where the estimate $\hat{\boldsymbol{\theta}}$ falls on the boundary of the allowable parameter space, in which case the formulas presented in this section will not be valid.

Asymptotic Standard Errors for Maximum Likelihood Estimates

If the sample size T is sufficiently large, it often turns out that the distribution of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ can be well approximated by the following distribution:

$$\hat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}_0, T^{-1}\mathcal{I}^{-1}), \qquad [5.8.1]$$

where $\boldsymbol{\theta}_0$ denotes the true parameter vector. The matrix $\mathcal{I}$ is known as the *information matrix* and can be estimated in either of two ways.

The *second-derivative estimate* of the information matrix is

$$\hat{\mathcal{I}}_{2D} = -T^{-1}\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \qquad [5.8.2]$$

Here $\mathcal{L}(\boldsymbol{\theta})$ denotes the log likelihood:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log f_{Y_t|\mathcal{Y}_{t-1}}(y_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta});$$

and $\mathcal{Y}_t$ denotes the history of observations on y obtained through date t. The matrix of second derivatives of the log likelihood is often calculated numerically. Substituting [5.8.2] into [5.8.1], the terms involving the sample size T cancel out so that the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ can be approximated by

$$E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \cong \left[-\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]^{-1}. \qquad [5.8.3]$$

A second estimate of the information matrix $\mathcal{I}$ in [5.8.1] is called the *outer-product estimate*:

$$\hat{\mathcal{I}}_{OP} = T^{-1}\sum_{t=1}^{T} [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)] \cdot [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)]'. \qquad [5.8.4]$$

Here $\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)$ denotes the $(a \times 1)$ vector of derivatives of the log of the conditional density of the tth observation with respect to the a elements of the parameter vector $\boldsymbol{\theta}$, with this derivative evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$:

$$\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t) = \frac{\partial \log f(y_t|y_{t-1}, y_{t-2}, \ldots; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

In this case, the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is approximated by

$$E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \cong \left[\sum_{t=1}^{T} [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)] \cdot [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)]'\right]^{-1}.$$

As an illustration of how such approximations can be used, suppose that the log likelihood is given by expression [5.7.6]. For this case, one can see analytically that

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} = \begin{bmatrix} -3 & 0 \\ 0 & -4 \end{bmatrix},$$

and so result [5.8.3] suggests that the variance of the maximum likelihood estimate $\hat{\theta}_2$ can be approximated by $\frac{1}{4}$. The *MLE* for this example was $\hat{\theta}_2 = 0$. Thus an

approximate 95% confidence interval for θ_2 is given by

$$0 \pm 2\sqrt{\tfrac{1}{4}} = \pm 1.$$

Note that unless the off-diagonal elements of $\hat{\mathcal{I}}$ are zero, in general one needs to calculate all the elements of the matrix $\hat{\mathcal{I}}$ and invert this full matrix in order to obtain a standard error for any given parameter.

Which estimate of the information matrix, $\hat{\mathcal{I}}_{2D}$ or $\hat{\mathcal{I}}_{OP}$, is it better to use in practice? Expression [5.8.1] is only an approximation to the true distribution of $\hat{\boldsymbol{\theta}}$, and $\hat{\mathcal{I}}_{2D}$ and $\hat{\mathcal{I}}_{OP}$ are in turn only approximations to the true value of $\mathcal{I}$. The theory that justifies these approximations does not give any clear guidance to which is better to use, and typically, researchers rely on whichever estimate of the information matrix is easiest to calculate. If the two estimates differ a great deal, this may mean that the model is misspecified. White (1982) developed a general test of model specification based on this idea. One option for constructing standard errors when the two estimates differ significantly is to use the "quasi-maximum likelihood" standard errors discussed at the end of this section.

Likelihood Ratio Test

Another popular approach to testing hypotheses about parameters that are estimated by maximum likelihood is the *likelihood ratio test*. Suppose a null hypothesis implies a set of m different restrictions on the value of the $(a \times 1)$ parameter vector $\boldsymbol{\theta}$. First, we maximize the likelihood function ignoring these restrictions to obtain the unrestricted maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. Next, we find an estimate $\tilde{\boldsymbol{\theta}}$ that makes the likelihood as large as possible while still satisfying all the restrictions. In practice, this is usually achieved by defining a new $[(a - m) \times 1]$ vector $\boldsymbol{\lambda}$ in terms of which all of the elements of $\boldsymbol{\theta}$ can be expressed when the restrictions are satisfied. For example, if the restriction is that the last m elements of $\boldsymbol{\theta}$ are zero, then $\boldsymbol{\lambda}$ consists of the first $a - m$ elements of $\boldsymbol{\theta}$. Let $\mathcal{L}(\hat{\boldsymbol{\theta}})$ denote the value of the log likelihood function at the unrestricted estimate, and let $\mathcal{L}(\tilde{\boldsymbol{\theta}})$ denote the value of the log likelihood function at the restricted estimate. Clearly $\mathcal{L}(\hat{\boldsymbol{\theta}}) > \mathcal{L}(\tilde{\boldsymbol{\theta}})$, and it often proves to be the case that

$$2[\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}})] \approx \chi^2(m). \qquad [5.8.5]$$

For example, suppose that $a = 2$ and we are interested in testing the hypothesis that $\theta_2 = \theta_1 + 1$. Under this null hypothesis the vector $(\theta_1, \theta_2)'$ can be written as $(\lambda, \lambda + 1)'$, where $\lambda = \theta_1$. Suppose that the log likelihood is given by expression [5.7.6]. One can find the restricted *MLE* by replacing θ_2 by $\theta_1 + 1$ and maximizing the resulting expression with respect to θ_1:

$$\tilde{\mathcal{L}}(\theta_1) = -1.5\theta_1^2 - 2(\theta_1 + 1)^2.$$

The first-order condition for maximization of $\tilde{\mathcal{L}}(\theta_1)$ is

$$-3\theta_1 - 4(\theta_1 + 1) = 0,$$

or $\theta_1 = -\frac{4}{7}$. The restricted *MLE* is thus $\tilde{\boldsymbol{\theta}} = (-\frac{4}{7}, \frac{3}{7})'$, and the maximum value attained for the log likelihood while satisfying the restriction is

$$\begin{aligned}\mathcal{L}(\tilde{\boldsymbol{\theta}}) &= (-\tfrac{3}{2})(-\tfrac{4}{7})^2 - (\tfrac{4}{2})(\tfrac{3}{7})^2 \\ &= -\{(3 \cdot 4)/(2 \cdot 7 \cdot 7)\}\{4 + 3\} \\ &= -\tfrac{6}{7}.\end{aligned}$$

The unrestricted *MLE* is $\hat{\boldsymbol{\theta}} = \mathbf{0}$, at which $\mathcal{L}(\hat{\boldsymbol{\theta}}) = 0$. Hence, [5.8.5] would be

$$2[\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}})] = \tfrac{12}{7} = 1.71.$$

The test here involves a single restriction, so $m = 1$. From Table B.2 in Appendix B, the probability that a $\chi^2(1)$ variable exceeds 3.84 is 0.05. Since $1.71 < 3.84$, we accept the null hypothesis that $\theta_2 = \theta_1 + 1$ at the 5% significance level.

Lagrange Multiplier Test

In order to use the standard errors from [5.8.2] or [5.8.4] to test a hypothesis about $\boldsymbol{\theta}$, we need only to find the unrestricted *MLE* $\hat{\boldsymbol{\theta}}$. In order to use the likelihood ratio test [5.8.5], it is necessary to find both the unrestricted *MLE* $\hat{\boldsymbol{\theta}}$ and the restricted *MLE* $\tilde{\boldsymbol{\theta}}$. The *Lagrange multiplier test* provides a third principle with which to test a null hypothesis that requires only the restricted *MLE* $\tilde{\boldsymbol{\theta}}$. This test is useful when it is easier to calculate the restricted estimate $\tilde{\boldsymbol{\theta}}$ than the unrestricted estimate $\hat{\boldsymbol{\theta}}$.

Let $\boldsymbol{\theta}$ be an $(a \times 1)$ vector of parameters, and let $\tilde{\boldsymbol{\theta}}$ be an estimate of $\boldsymbol{\theta}$ that maximizes the log likelihood subject to a set of m restrictions on $\boldsymbol{\theta}$. Let $f(y_t|y_{t-1}, y_{t-2}, \ldots; \boldsymbol{\theta})$ be the conditional density of the tth observation, and let $\mathbf{h}(\tilde{\boldsymbol{\theta}}, \mathcal{Y}_t)$ denote the $(a \times 1)$ vector of derivatives of the log of this conditional density evaluated at the restricted estimate $\tilde{\boldsymbol{\theta}}$:

$$\mathbf{h}(\tilde{\boldsymbol{\theta}}, \mathcal{Y}_t) = \left.\frac{\partial \log f(y_t|y_{t-1}, y_{t-2}, \ldots; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}.$$

The Lagrange multiplier test of the null hypothesis that the restrictions are true is given by the following statistic:

$$T^{-1}\left[\sum_{t=1}^{T} \mathbf{h}(\tilde{\boldsymbol{\theta}}, \mathcal{Y}_t)\right]' \mathcal{I}^{-1}\left[\sum_{t=1}^{T} \mathbf{h}(\tilde{\boldsymbol{\theta}}, \mathcal{Y}_t)\right]. \qquad [5.8.6]$$

If the null hypothesis is true, then for large T this should approximately have a $\chi^2(m)$ distribution. The information matrix $\mathcal{I}$ can again be estimated as in [5.8.2] or [5.8.4] with $\hat{\boldsymbol{\theta}}$ replaced by $\tilde{\boldsymbol{\theta}}$.

Quasi-Maximum Likelihood Standard Errors

It was mentioned earlier in this section that if the data were really generated from the assumed density and the sample size is sufficiently large, the second-derivative estimate $\hat{\mathcal{I}}_{2D}$ and the outer-product estimate $\hat{\mathcal{I}}_{OP}$ of the information matrix should be reasonably close to each other. However, maximum likelihood estimation may still be a reasonable way to estimate parameters even if the data were not generated by the assumed density. For example, we noted in Section 5.2 that the conditional *MLE* for a Gaussian $AR(1)$ process is obtained from an *OLS* regression of y_t on y_{t-1}. This *OLS* regression is often a very sensible way to estimate parameters of an $AR(1)$ process even if the true innovations ε_t are not i.i.d. Gaussian. Although maximum likelihood may be yielding a reasonable estimate of $\boldsymbol{\theta}$, when the innovations are not i.i.d. Gaussian, the standard errors proposed in [5.8.2] or [5.8.4] may no longer be valid. An approximate variance-covariance matrix for $\hat{\boldsymbol{\theta}}$ that is sometimes valid even if the probability density is misspecified is given by

$$E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \cong T^{-1}\{\mathcal{I}_{2D}\mathcal{I}_{OP}^{-1}\mathcal{I}_{2D}\}^{-1}. \qquad [5.8.7]$$

This variance-covariance matrix was proposed by White (1982), who described this approach as *quasi-maximum likelihood estimation*.

5.9. *Inequality Constraints*

A Common Pitfall with Numerical Maximization

Suppose we were to apply one of the methods discussed in Section 5.7 such as steepest ascent to the $AR(1)$ likelihood [5.7.2]. We start with an arbitrary initial guess, say, $\phi = 0.1$. We calculate the gradient at this point, and find that it is positive. The computer is then programmed to try to improve this estimate by evaluating the log likelihood at points described by $\phi^{(1)} = \phi^{(0)} + s \cdot g(\phi^{(0)})$ for various values of s, seeing what works best. But if the computer were to try a value for s such that $\phi^{(1)} = \phi^{(0)} + s \cdot g(\phi^{(0)}) = 1.1$, calculation of [5.7.2] would involve finding the log of $(1 - 1.1^2) = -0.21$. Attempting to calculate the log of a negative number would typically be a fatal execution error, causing the search procedure to crash.

Often such problems can be avoided by using modified Newton-Raphson procedures, provided that the initial estimate $\boldsymbol{\theta}^{(0)}$ is chosen wisely and provided that the initial search area is kept fairly small. The latter might be accomplished by setting the initial weighting matrix $\mathbf{A}^{(0)}$ in [5.7.15] and [5.7.16] equal to a small multiple of the identity matrix, such as $\mathbf{A}^{(0)} = (1 \times 10^{-4}) \cdot \mathbf{I}_a$. In later iterations, the algorithm should use the shape of the likelihood function in the vicinity of the maximum to keep the search conservative. However, if the true *MLE* is close to one of the boundaries (for example, if $\hat{\phi}_{MLE} = 0.998$ in the $AR(1)$ example), it will be virtually impossible to keep a numerical algorithm from exploring what happens when ϕ is greater than unity, which would induce a fatal crash.

Solving the Problem by Reparameterizing the Likelihood Function

One simple way to ensure that a numerical search always stays within certain specified boundaries is to reparameterize the likelihood function in terms of an $(a \times 1)$ vector $\boldsymbol{\lambda}$ for which $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\lambda})$, where the function $\mathbf{g}$: $\mathbb{R}^a \rightarrow \mathbb{R}^a$ incorporates the desired restrictions. The scheme is then as follows:

Input		*Procedure*		*Output*
values of $y_1, y_2, \ldots, y_T$ and $\boldsymbol{\lambda}$	$\rightarrow$	set $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\lambda})$; calculate $\mathcal{L}(\boldsymbol{\theta})$	$\rightarrow$	value of $\mathcal{L}(\mathbf{g}(\boldsymbol{\lambda}))$

For example, to ensure that ϕ is always between ± 1, we could take

$$\phi = g(\lambda) = \frac{\lambda}{1 + |\lambda|}. \qquad [5.9.1]$$

The goal is to find the value of λ that produces the biggest value for the log likelihood. We start with an initial guess such as $\lambda = 3$. The procedure to evaluate the log likelihood function first calculates

$$\phi = 3/(1 + 3) = 0.75$$

and then finds the value for the log likelihood associated with this value of ϕ from [5.7.2]. No matter what value for λ the computer guesses, the value of ϕ in [5.9.1] will always be less than 1 in absolute value and the likelihood function will be well

defined. Once we have found the value of $\hat{\lambda}$ that maximizes the likelihood function, the maximum likelihood estimate of ϕ is then given by

$$\hat{\phi} = \frac{\hat{\lambda}}{1 + |\hat{\lambda}|}.$$

This technique of reparameterizing the likelihood function so that estimates always satisfy any necessary constraints is often very easy to implement. However, one note of caution should be mentioned. If a standard error is calculated from the matrix of second derivatives of the log likelihood as in [5.8.3], this represents the standard error of $\hat{\lambda}$, not the standard error of $\hat{\phi}$. To obtain a standard error for $\hat{\phi}$, the best approach is first to parameterize the likelihood function in terms of λ to find the *MLE*, and then to reparameterize in terms of ϕ to calculate the matrix of second derivatives evaluated at $\hat{\phi}$ to get the final standard error for $\hat{\phi}$. Alternatively, one can calculate an approximation to the standard error for $\hat{\phi}$ from the standard error for $\hat{\lambda}$, based on the formula for a Wald test of a nonlinear hypothesis described in Chapter 14.

Parameterizations for a Variance-Covariance Matrix

Another common restriction one needs to impose is that a variance parameter σ^2 be positive. An obvious way to achieve this is to parameterize the likelihood in terms of λ which represents ± 1 times the standard deviation. The procedure to evaluate the log likelihood then begins by squaring this parameter λ:

$$\sigma^2 = \lambda^2;$$

and if the standard deviation σ is itself called, it is calculated as

$$\sigma = \sqrt{\lambda^2}.$$

More generally, let $\mathbf{\Omega}$ denote an $(n \times n)$ variance-covariance matrix:

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}.$$

Here one needs to impose the condition that $\mathbf{\Omega}$ is positive definite and symmetric. The best approach is to parameterize $\mathbf{\Omega}$ in terms of the $n(n + 1)/2$ distinct elements of the Cholesky decomposition of $\mathbf{\Omega}$:

$$\mathbf{\Omega} = \mathbf{PP}', \qquad [5.9.2]$$

where

$$\mathbf{P} = \begin{bmatrix} \lambda_{11} & 0 & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \lambda_{n3} & \cdots & \lambda_{nn} \end{bmatrix}.$$

No matter what values the computer guesses for $\lambda_{11}, \lambda_{21}, \ldots, \lambda_{nn}$, the matrix $\mathbf{\Omega}$ calculated from [5.9.2] will be symmetric and positive semidefinite.

Parameterizations for Probabilities

Sometimes some of the unknown parameters are probabilities $p_1, p_2, \ldots, p_K$ which must satisfy the restrictions

$$0 \leq p_i \leq 1 \qquad \text{for } i = 1, 2, \ldots, K$$
$$p_1 + p_2 + \cdots + p_K = 1.$$

In this case, one approach is to parameterize the probabilities in terms of $\lambda_1, \lambda_2, \ldots, \lambda_{K-1}$, where

$$p_i = \lambda_i^2/(1 + \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{K-1}^2) \qquad \text{for } i = 1, 2, \ldots, K-1$$
$$p_K = 1/(1 + \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{K-1}^2).$$

More General Inequality Constraints

For more complicated inequality constraints that do not admit a simple reparameterization, an approach that sometimes works is to put a branching statement in the procedure to evaluate the log likelihood function. The procedure first checks whether the constraint is satisfied. If it is, then the likelihood function is evaluated in the usual way. If it is not, then the procedure returns a large negative number in place of the value of the log likelihood function. Sometimes such an approach will allow an *MLE* satisfying the specified conditions to be found with simple numerical search procedures.

If these measures prove inadequate, more complicated algorithms are available. Judge, Griffiths, Hill, and Lee (1980, pp. 747–49) described some of the possible approaches.

APPENDIX 5.A. *Proofs of Chapter 5 Propositions*

■ **Proof of Proposition 5.1.**

(a) By Taylor's theorem,

$$\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) = \mathcal{L}(\boldsymbol{\theta}^{(m)}) + [\mathbf{g}(\boldsymbol{\theta}^{(m)})]'[\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}] + R_1(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m+1)}). \qquad [5.A.1]$$

Substituting [5.7.18] into [5.A.1],

$$\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) = [\mathbf{g}(\boldsymbol{\theta}^{(m)})]'s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}) + R_1(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m+1)}). \qquad [5.A.2]$$

Since $\mathbf{A}^{(m)}$ is positive definite and since $\mathbf{g}(\boldsymbol{\theta}^{(m)}) \neq \mathbf{0}$, expression [5.A.2] establishes that

$$\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) = s\kappa(\boldsymbol{\theta}^{(m)}) + R_1(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m+1)}),$$

where $\kappa(\boldsymbol{\theta}^{(m)}) > 0$. Moreover, $s^{-1} \cdot R_1(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m+1)}) \to 0$ as $s \to 0$. Hence, there exists an s such that $\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) > 0$, as claimed.

(b) Direct differentiation reveals

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\theta}^{(m+1)})}{\partial s} &= \frac{\partial \mathcal{L}}{\partial \theta_1}\frac{\partial \theta_1}{\partial s} + \frac{\partial \mathcal{L}}{\partial \theta_2}\frac{\partial \theta_2}{\partial s} + \cdots + \frac{\partial \mathcal{L}}{\partial \theta_a}\frac{\partial \theta_a}{\partial s} \\ &= [\mathbf{g}(\boldsymbol{\theta}^{(m+1)})]'\,\frac{\partial \boldsymbol{\theta}^{(m+1)}}{\partial s} \qquad [5.A.3] \\ &= [\mathbf{g}(\boldsymbol{\theta}^{(m+1)})]'\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}),\end{aligned}$$

with the last line following from [5.7.18]. The first-order conditions set [5.A.3] equal to zero, which implies

$$0 = [\mathbf{g}(\boldsymbol{\theta}^{(m+1)})]'s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}) = [\mathbf{g}(\boldsymbol{\theta}^{(m+1)})]'[\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}],$$

with the last line again following from [5.7.18]. This establishes the claim in [5.7.19].

(c) Let $\mathbf{y}$ be any $(a \times 1)$ nonzero vector. The task is to show that $\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} > 0$. Observe from [5.7.16] that

$$\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} = \mathbf{y}'\mathbf{A}^{(m)}\mathbf{y} - \frac{\mathbf{y}'\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})} - \frac{\mathbf{y}'(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})}. \qquad [5.A.4]$$

Since $\mathbf{A}^{(m)}$ is positive definite, there exists a nonsingular matrix $\mathbf{P}$ such that

$$\mathbf{A}^{(m)} = \mathbf{P}\mathbf{P}'.$$

Define

$$\mathbf{y}^* = \mathbf{P}'\mathbf{y}$$
$$\mathbf{x}^* = \mathbf{P}'\Delta\mathbf{g}^{(m+1)}.$$

Then [5.A.4] can be written as

$$\begin{aligned}\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} &= \mathbf{y}'\mathbf{P}\mathbf{P}'\mathbf{y} - \frac{\mathbf{y}'\mathbf{P}\mathbf{P}'(\Delta\mathbf{g}^{(m+1)})(\Delta\mathbf{g}^{(m+1)})'\mathbf{P}\mathbf{P}'\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'\mathbf{P}\mathbf{P}'(\Delta\mathbf{g}^{(m+1)})} \\ &\quad - \frac{\mathbf{y}'(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})} \qquad [5.A.5] \\ &= \mathbf{y}^{*\prime}\mathbf{y}^* - \frac{(\mathbf{y}^{*\prime}\mathbf{x}^*)(\mathbf{x}^{*\prime}\mathbf{y}^*)}{\mathbf{x}^{*\prime}\mathbf{x}^*} - \frac{\mathbf{y}'(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})}.\end{aligned}$$

Recalling equation [4.A.6], the first two terms in the last line of [5.A.5] represent the sum of squared residuals from an *OLS* regression of $\mathbf{y}^*$ on $\mathbf{x}^*$. This cannot be negative,

$$\mathbf{y}^{*\prime}\mathbf{y}^* - \frac{(\mathbf{y}^{*\prime}\mathbf{x}^*)(\mathbf{x}^{*\prime}\mathbf{y}^*)}{\mathbf{x}^{*\prime}\mathbf{x}^*} \geq 0; \qquad [5.A.6]$$

it would equal zero only if the *OLS* regression has a perfect fit, or if $\mathbf{y}^* = \beta\mathbf{x}^*$ or $\mathbf{P}'\mathbf{y} = \beta\mathbf{P}'\Delta\mathbf{g}^{(m+1)}$ for some β. Since $\mathbf{P}$ is nonsingular, expression [5.A.6] would only be zero if $\mathbf{y} = \beta\Delta\mathbf{g}^{(m+1)}$ for some β. Consider two cases.

Case 1. There is no β such that $\mathbf{y} = \beta\Delta\mathbf{g}^{(m+1)}$. In this case, the inequality [5.A.6] is strict and [5.A.5] implies

$$\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} > -\frac{[\mathbf{y}'\Delta\boldsymbol{\theta}^{(m+1)}]^2}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})}.$$

Since $[\mathbf{y}'\Delta\boldsymbol{\theta}^{(m+1)}]^2 \geq 0$, it follows that $\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} > 0$, provided that

$$(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)}) < 0. \qquad [5.A.7]$$

But, from [5.7.19],

$$\begin{aligned}(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)}) &= [\mathbf{g}(\boldsymbol{\theta}^{(m+1)}) - \mathbf{g}(\boldsymbol{\theta}^{(m)})]'(\Delta\boldsymbol{\theta}^{(m+1)}) \\ &= -\mathbf{g}(\boldsymbol{\theta}^{(m)})'(\Delta\boldsymbol{\theta}^{(m+1)}) \qquad [5.A.8] \\ &= -\mathbf{g}(\boldsymbol{\theta}^{(m)})'s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}),\end{aligned}$$

with the last line following from [5.7.18]. But the final term in [5.A.8] must be negative, by virtue of the facts that $\mathbf{A}^{(m)}$ is positive definite, $s > 0$, and $\mathbf{g}(\boldsymbol{\theta}^{(m)}) \neq \mathbf{0}$. Hence, [5.A.7] holds, meaning that $\mathbf{A}^{(m+1)}$ is positive definite for this case.

Case 2. There exists a β such that $\mathbf{y} = \beta\Delta\mathbf{g}^{(m+1)}$. In this case, [5.A.6] is zero, so that [5.A.5] becomes

$$\begin{aligned}\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} &= -\frac{\mathbf{y}'(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})} \\ &= -\frac{\beta(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'\beta(\Delta\mathbf{g}^{(m+1)})}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})} \\ &= -\beta^2(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)}) = \beta^2\mathbf{g}(\boldsymbol{\theta}^{(m)})'s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}) > 0,\end{aligned}$$

as in [5.A.8]. ■

Chapter 5 Exercises

5.1. Show that the value of [5.4.16] at $\theta = \tilde{\theta}$, $\sigma^2 = \tilde{\sigma}^2$ is identical to its value at $\theta = \tilde{\theta}^{-1}$, $\sigma^2 = \tilde{\theta}^2\tilde{\sigma}^2$.

5.2. Verify that expression [5.7.12] calculates the maximum of [5.7.6] in a single step from the initial estimate $\theta^{(0)} = (-1, 1)'$.

5.3. Let $(y_1, y_2, \ldots, y_T)$ be a sample of size T drawn from an i.i.d. $N(\mu, \sigma^2)$ distribution.

(a) Show that the maximum likelihood estimates are given by

$$\hat{\mu} = T^{-1}\sum_{t=1}^{T} y_t$$

$$\hat{\sigma}^2 = T^{-1}\sum_{t=1}^{T} (y_t - \hat{\mu})^2.$$

(b) Show that the matrix $\hat{\mathcal{I}}_{2D}$ in [5.8.2] is

$$\hat{\mathcal{I}}_{2D} = \begin{bmatrix} 1/\hat{\sigma}^2 & 0 \\ 0 & 1/(2\hat{\sigma}^4) \end{bmatrix}.$$

(c) Show that for this example result [5.8.1] suggests

$$\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} \approx N\left(\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \hat{\sigma}^2/T & 0 \\ 0 & 2\hat{\sigma}^4/T \end{bmatrix}\right).$$

Chapter 5 References

Anderson, Brian D. O., and John B. Moore. 1979. *Optimal Filtering*. Englewood Cliffs, N.J.: Prentice-Hall.

Berndt, E. K., B. H. Hall, R. E. Hall, and J. A. Hausman. 1974. "Estimation and Inference in Nonlinear Structural Models." *Annals of Economic and Social Measurement* 3:653–65.

Box, George E. P., and D. R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society* Series B, 26:211–52.

——— and Gwilym M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*, rev. ed. San Francisco: Holden-Day.

Broyden, C. G. 1965. "A Class of Methods for Solving Nonlinear Simultaneous Equations." *Mathematics of Computation* 19:577–93.

———. 1967. "Quasi-Newton Methods and Their Application to Function Minimization." *Mathematics of Computation* 21:368–81.

Chiang, Alpha C. 1974. *Fundamental Methods of Mathematical Economics*, 2d ed. New York: McGraw-Hill.

Davidon, W. C. 1959. "Variable Metric Method of Minimization." A.E.C. Research and Development Report ANL-5990 (rev.).

Fletcher, R., and M. J. D. Powell. 1963. "A Rapidly Convergent Descent Method for Minimization." *Computer Journal* 6:163–68.

Galbraith, R. F., and J. I. Galbraith. 1974. "On the Inverses of Some Patterned Matrices Arising in the Theory of Stationary Time Series." *Journal of Applied Probability* 11:63–71.

Hannan, E., and J. Rissanen. 1982. "Recursive Estimation of Mixed Autoregressive–Moving Average Order." *Biometrika* 69:81–94.

Janacek, G. J., and A. L. Swift. 1990. "A Class of Models for Non-Normal Time Series." *Journal of Time Series Analysis* 11:19–31.

Judge, George G., William E. Griffiths, R. Carter Hill, and Tsoung-Chao Lee. 1980. *The Theory and Practice of Econometrics*. New York: Wiley.

Koreisha, Sergio, and Tarmo Pukkila. 1989. "Fast Linear Estimation Methods for Vector Autoregressive Moving-Average Models." *Journal of Time Series Analysis* 10:325–39.

Li, W. K., and A. I. McLeod. 1988. "ARMA Modelling with Non-Gaussian Innovations." *Journal of Time Series Analysis* 9:155–68.

Martin, R. D. 1981. "Robust Methods for Time Series," in D. F. Findley, ed., *Applied Time Series*, Vol. II. New York: Academic Press.

Nelson, Harold L., and C. W. J. Granger. 1979. "Experience with Using the Box-Cox Transformation When Forecasting Economic Time Series." *Journal of Econometrics* 10:57–69.

Quandt, Richard E. 1983. "Computational Problems and Methods," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, Vol. 1. Amsterdam: North-Holland.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50:1–25.