12

Estimation of VARMA Models

In this chapter, maximum likelihood estimation of the coefficients of a VARMA model is considered. Before we can proceed to the actual estimation, a unique set of parameters must be specified. In this context, the problem of nonuniqueness of a VARMA representation becomes important. This identification problem, that is, the problem of identifying a unique structure among many equivalent ones, is treated in Section 12.1. In Section 12.2, the Gaussian likelihood function of a VARMA model is considered. A numerical algorithm for maximizing it and, thus, for computing the actual estimates is discussed in Section 12.3. The asymptotic properties of the ML estimators are the subject of Section 12.4. Forecasting with estimated processes and impulse response analysis are dealt with in Sections 12.5 and 12.6, respectively.

12.1 The Identification Problem

12.1.1 Nonuniqueness of VARMA Representations

In the previous chapter, we have considered K-dimensional, stationary processes y_t with VARMA(p, q) representations

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t + M_1 u_{t-1} + \cdots + M_q u_{t-q}. \tag{12.1.1}$$

Because the mean term is of no importance for the presently considered problem, we have set it to zero. Therefore, no intercept term appears in (12.1.1). This model can be written in lag operator notation as

$$A(L)y_t = M(L)u_t, \tag{12.1.2}$$

where $A(L) := I_K - A_1 L - \cdots - A_p L^p$ and $M(L) := I_K + M_1 L + \cdots + M_q L^q$. Assuming that the VARMA representation is stable and invertible, the well-defined process described by the model (12.1.1) or (12.1.2) is given by

H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*

$$y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i} = \Phi(L)u_t = A(L)^{-1}M(L)u_t.$$

In practice, it is sometimes useful to consider a slightly more general type of VARMA model by attaching nonidentity coefficient matrices to y_t and u_t, that is, one may want to consider representations of the type

$$A_0 y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + M_0 v_t + M_1 v_{t-1} + \cdots + M_q v_{t-q}, \quad (12.1.3)$$

where v_t is a suitable white noise process. Such a form may be suggested by subject matter theory which may imply instantaneous effects of some variables on other variables. It will also turn out to be useful in finding unique structures for VARMA models. By the specification (12.1.3) we mean the well-defined process

$$y_t = (A_0 - A_1 L - \cdots - A_p L^p)^{-1}(M_0 + M_1 L + \cdots + M_q L^q)v_t.$$

Such a process has a standard VARMA(p, q) representation with identity coefficient matrices attached to the instantaneous y_t and u_t if A_0 and M_0 are nonsingular. To see this, we premultiply (12.1.3) by A_0^{-1} and define $u_t = A_0^{-1}M_0 v_t$ which gives

$$\begin{aligned} y_t &= A_0^{-1}A_1 y_{t-1} + \cdots + A_0^{-1}A_p y_{t-p} + u_t + A_0^{-1}M_1 M_0^{-1}A_0 u_{t-1} + \cdots \\ &\quad + A_0^{-1}M_q M_0^{-1}A_0 u_{t-q}. \end{aligned}$$

Redefining the matrices appropriately, this, of course, is a representation of the type (12.1.1) with identity coefficient matrices at lag zero which describes the same process as (12.1.3). The assumption that both A_0 and M_0 are nonsingular does not entail any loss of generality, as long as none of the components of y_t can be written as a linear combination of the other components. We call a stable and invertible representation as in (12.1.1) a *VARMA representation in standard form* or a *standard VARMA representation* to distinguish it from representations with nonidentity matrices at lag zero as in (12.1.3). This discussion shows that VARMA representations are not unique, that is, a given process y_t can be written in standard form or in nonstandard form by premultiplying by any nonsingular $(K \times K)$ matrix. We have encountered a similar problem in dealing with finite order structural VAR processes in Chapter 9. However, once we consider standard reduced form VAR models only, we have unique representations. This property is in sharp contrast to the presently considered VARMA case, where, in general, a standard form is not a unique representation, as we will see shortly.

It may be useful at this stage to emphasize what we mean by equivalent representations of a process. Generally, two representations of a process y_t are equivalent if they give rise to the same realizations (except on a set of measure zero) and, thus, to the same multivariate distributions of any finite subcollection of variables $y_t, y_{t+1}, \ldots, y_{t+h}$, for arbitrary integers t and h. Of course, this specification just says that equivalent representations really

represent the same process. If y_t is a zero mean process with canonical MA representation

$$\begin{aligned} y_t &= \sum_{i=0}^{\infty} \Phi_i u_{t-i}, \quad \Phi_0 = I_K, \\ &= \Phi(L)u_t, \end{aligned} \tag{12.1.4}$$

where $\Phi(L) := \sum_{i=0}^{\infty} \Phi_i L^i$, then any VARMA model $A(L)y_t = M(L)u_t$ for which

$$A(L)^{-1}M(L) = \Phi(L) \tag{12.1.5}$$

is an equivalent representation of the process y_t. In other words, all VARMA models are equivalent for which $A(L)^{-1}M(L)$ results in the same operator $\Phi(L)$. Thus, in order to ensure uniqueness of a VARMA representation, we must impose restrictions on the VAR and MA operators such that there is precisely one feasible pair of operators $A(L)$ and $M(L)$ satisfying (12.1.5) for a given $\Phi(L)$.

Obviously, given some stable, invertible VARMA representation $A(L)y_t = M(L)u_t$, an equivalent representation results if we premultiply by any nonsingular matrix A_0. Therefore, to remove this source of nonuniqueness, let us for the moment focus on VARMA representations in standard form. As mentioned earlier, even then uniqueness is not ensured. To see this problem more clearly, let us consider a bivariate VARMA$(1,1)$ process in standard form,

$$y_t = A_1 y_{t-1} + u_t + M_1 u_{t-1}. \tag{12.1.6}$$

From Section 11.3.1, we know that this process has the canonical MA representation

$$y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i} = u_t + \sum_{i=1}^{\infty} (A_1^i + A_1^{i-1} M_1) u_{t-i}. \tag{12.1.7}$$

Thus, for example, any VARMA$(1,1)$ representation with $M_1 = -A_1$ will result in the same canonical MA representation. In other words, if it turns out that y_t is such that $M_1 = -A_1$ for some set of coefficients, then any choice of A_1 matrix that gives rise to a stable VAR operator can be matched by an M_1 matrix that leads to an equivalent VARMA$(1,1)$ representation of y_t. Of course, in this case, the MA coefficient matrices in (12.1.7) are in fact all zero and $y_t = u_t$ is really white noise, that is, y_t actually has a VARMA$(0,0)$ structure. This fact is also quite easy to see from the lag operator representation of (12.1.6),

$$(I_2 - A_1 L)y_t = (I_2 + M_1 L)u_t.$$

Of course, if $M_1 = -A_1$, the MA operator cancels against the VAR operator. This type of parameter indeterminacy is also known from univariate ARMA

processes. It is usually ruled out by the assumption that the AR and MA operators have no common factors. Let us make a similar assumption in the presently considered multivariate case by requiring that y_t is not white noise, i.e., $M_1 \neq -A_1$.

Unfortunately, in the multivariate case, the nonuniqueness problem is not solved by this assumption. To see this, suppose that

$$A_1 = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad M_1 = 0,$$

where $\alpha \neq 0$. In this case, the canonical MA representation (12.1.4) has coefficient matrices

$$\Phi_1 = A_1, \quad \Phi_2 = \Phi_3 = \cdots = 0, \tag{12.1.8}$$

because $A_1^i = 0$ for $i > 1$. The same MA representation results if

$$A_1 = 0 \quad \text{and} \quad M_1 = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix}.$$

More generally, a canonical MA representation with coefficient matrices as in (12.1.8) is obtained if

$$A_1 = \begin{bmatrix} 0 & \alpha + m \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad M_1 = \begin{bmatrix} 0 & -m \\ 0 & 0 \end{bmatrix},$$

whatever the value of m. Note also that the VARMA representation will be stable and invertible for any value of m.

To understand where the parameter indeterminacy comes from, consider the VAR operator

$$I_2 - \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} L. \tag{12.1.9}$$

The inverse of this operator is

$$I_2 + \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} L, \tag{12.1.10}$$

which is easily checked by multiplying the two operators together. Thus, the operator (12.1.9) has a finite order inverse. Operators of this type are precisely the ones that cause trouble in setting up a uniquely parameterized VARMA representation of a given process because multiplying by such an operator may cancel part of one operator (VAR or MA) while at the same time the finite order of the other operator is maintained.

To get a better sense for this problem, let us look at the following VARMA$(1, 1)$ process:

$$A(L)y_t = M(L)u_t,$$

where

$$A(L) := \begin{bmatrix} 1-\alpha_{11}L & -\alpha_{12}L \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad M(L) := \begin{bmatrix} 1+m_{11}L & m_{12}L \\ 0 & 1 \end{bmatrix}.$$

The two operators do not cancel if $\alpha_{11} \neq -m_{11}$ and $\alpha_{12} \neq -m_{12}$. Still we can factor an operator

$$D(L) := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & \gamma \\ 0 & 0 \end{bmatrix} L = \begin{bmatrix} 1 & \gamma L \\ 0 & 1 \end{bmatrix}$$

from both operators without changing their general structure:

$$A(L) = D(L) \begin{bmatrix} 1-\alpha_{11}L & -(\gamma+\alpha_{12})L \\ 0 & 1 \end{bmatrix},$$

$$M(L) = D(L) \begin{bmatrix} 1+m_{11}L & (m_{12}-\gamma)L \\ 0 & 1 \end{bmatrix}.$$

Cancelling $D(L)$ gives operators

$$\begin{bmatrix} 1-\alpha_{11}L & -(\gamma+\alpha_{12})L \\ 0 & 1 \end{bmatrix} = D(L) \begin{bmatrix} 1+\alpha_{11}L & -(2\gamma+\alpha_{12})L \\ 0 & 1 \end{bmatrix}$$

and

$$\begin{bmatrix} 1+m_{11}L & (m_{12}-\gamma)L \\ 0 & 1 \end{bmatrix} = D(L) \begin{bmatrix} 1+m_{11}L & (m_{12}-2\gamma)L \\ 0 & 1 \end{bmatrix}.$$

Thus, we can again factor and cancel $D(L)$. In fact, we can cancel $D(L)$ as often as we like without changing the general structure of the process. Hence, even if the orders of both operators cannot be reduced simultaneously by cancellation, it may still be possible to factor some operator from both $A(L)$ and $M(L)$ without changing their general structure. Note that the troubling operator $D(L)$ is again one with finite order inverse,

$$D(L)^{-1} = \begin{bmatrix} 1 & -\gamma L \\ 0 & 1 \end{bmatrix}.$$

Finite order operators that have a finite order inverse are characterized by the property that their determinant is a nonzero constant, that is, it does not involve L or powers of L. Operators with this property are called *unimodular*. For instance, the operator (12.1.9) has determinant,

$$\left| I_2 - \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} L \right| = \left| \begin{bmatrix} 1 & -\alpha L \\ 0 & 1 \end{bmatrix} \right| = 1$$

and, hence, it is unimodular. The property of a unimodular operator to have a finite order inverse follows because the inverse of an operator $A(L)$ is its adjoint divided by its determinant,

$$A(L)^{-1} = A(L)^{adj}/|A(L)| = |A(L)|^{-1}A(L)^{adj}.$$

The determinant is a univariate operator. A finite order invertible univariate operator, however, has an infinite order inverse, unless its degree is zero, that is, unless it is a constant.

In order to state uniqueness conditions for a VARMA representation, we will first of all require that a representation is chosen for which further cancellation is not possible in the sense that there are no common factors in the VAR and MA parts, except for unimodular operators. Operators $A(L)$ and $M(L)$ with this property are *left-coprime*. This property may be defined by calling the matrix operator $[A(L) : M(L)]$ left-coprime, if the existence of operators $D(L)$, $\bar{A}(L)$, and $\bar{M}(L)$ satisfying

$$D(L)[\bar{A}(L) : \bar{M}(L)] = [A(L) : M(L)] \tag{12.1.11}$$

implies that $D(L)$ is unimodular, that is, $|D(L)|$ is a nonzero constant. From the foregoing examples, it should be understood that in general factoring unimodular operators from $A(L)$ and $M(L)$ is unavoidable if no further constraints are imposed. Thus, to obtain uniqueness of left-coprime operators we have to impose restrictions ensuring that the only feasible unimodular operator $D(L)$ in (12.1.11) is $D(L) = I_K$. We will now give two sets of conditions that ensure uniqueness of a VARMA representation.

12.1.2 Final Equations Form and Echelon Form

Suppose y_t is a stationary zero mean process that has a stable, invertible VARMA representation,

$$A(L)y_t = M(L)u_t, \tag{12.1.12}$$

where $A(L) := A_0 - A_1L - \cdots - A_pL^p$ and $M(L) := M_0 + M_1L + \cdots + M_qL^q$. Further suppose that $A(L)$ and $M(L)$ are left-coprime and the white noise covariance matrix Σ_u is nonsingular.

Definition 12.1 (*Final Equations Form*)
The VARMA representation (12.1.12) is said to be in *final equations form* if $M_0 = I_K$ and $A(L) = \alpha(L)I_K$, where $\alpha(L) := 1 - \alpha_1L - \cdots - \alpha_pL^p$ is a scalar (one-dimensional) operator with $\alpha_p \neq 0$. ■

For instance, the bivariate VARMA$(3, 1)$ model

$$(1 - \alpha_1L - \alpha_2L^2 - \alpha_3L^3)\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} 1 + m_{11,1}L & m_{12,1}L \\ m_{21,1}L & 1 + m_{22,1}L \end{bmatrix}\begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{12.1.13}$$

with $\alpha_3 \neq 0$, is in final equations form. The label "final equations form" for this type of VARMA representation is in line with the terminology used in Chapter 10, Section 10.2.2.

Uniqueness of the final equations form

$$\alpha(L)y_t = M(L)u_t$$

is seen by noting that $D(L) = I_K$ is the only operator that retains the scalar AR part upon multiplication. For the operator $D(L)\alpha(L)I_K$ to maintain the order p, the operator $D(L)$ must have degree zero, that is, $D(L) = D$. However, the only possible matrix D that guarantees a zero order matrix I_K for the VAR operator is $D = I_K$.

Definition 12.2 (*Echelon Form*)
The VARMA representation (12.1.12) is said to be in *echelon form* or ARMA_E form if the VAR and MA operators $A(L) = [\alpha_{ki}(L)]_{k,i=1,\ldots,K}$ and $M(L) = [m_{ki}(L)]$ are left-coprime and satisfy the following conditions: The operators $\alpha_{ki}(L)$ $(i = 1, \ldots, K)$ and $m_{kj}(L)$ $(j = 1, \ldots, K)$ in the k-th row of $A(L)$ and $M(L)$ have degree p_k and they have the form

$$\alpha_{kk}(L) = 1 - \sum_{j=1}^{p_k} \alpha_{kk,j} L^j, \quad \text{for } k = 1, \ldots, K,$$

$$\alpha_{ki}(L) = - \sum_{j=p_k - p_{ki}+1}^{p_k} \alpha_{ki,j} L^j, \quad \text{for } k \neq i,$$

and

$$m_{ki}(L) = \sum_{j=0}^{p_k} m_{ki,j} L^j, \quad \text{for } k, i = 1, \ldots, K, \quad \text{with } M_0 = A_0.$$

In the VAR operators $\alpha_{ki}(L)$,

$$p_{ki} := \begin{cases} \min(p_k + 1, p_i) & \text{for } k \geq i, \\ \min(p_k, p_i) & \text{for } k < i, \end{cases} \qquad k, i = 1, \ldots, K. \tag{12.1.14}$$

That is, p_{ki} specifies the number of free coefficients in the operator $\alpha_{ki}(L)$ for $i \neq k$. The row degrees $(p_1, \ldots, p_K)$ are called the *Kronecker indices* and their sum $\sum_{k=1}^{K} p_i$ is the *McMillan degree*. Obviously, for the VARMA orders we have, in general, $p = q = \max(p_1, \ldots, p_K)$. ■

We will sometimes denote an echelon form VARMA model with Kronecker indices $(p_1, \ldots, p_K)$ by $\text{ARMA}_E(p_1, \ldots, p_K)$. The following model is an example of a bivariate VARMA process in echelon form or, more precisely, an $\text{ARMA}_E(2, 1)$:

$$\begin{bmatrix} 1 - \alpha_{11,1}L - \alpha_{11,2}L^2 & -\alpha_{12,2}L^2 \\ -\alpha_{21,0} - \alpha_{21,1}L & 1 - \alpha_{22,1}L \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix}$$
$$= \begin{bmatrix} 1 + m_{11,1}L + m_{11,2}L^2 & m_{12,1}L + m_{12,2}L^2 \\ -\alpha_{21,0} + m_{21,1}L & 1 + m_{22,1}L \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{12.1.15}$$

or

$$\begin{aligned}
&\begin{bmatrix} 1 & 0 \\ -\alpha_{21,0} & 1 \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} \\
&= \begin{bmatrix} \alpha_{11,1} & 0 \\ \alpha_{21,1} & \alpha_{22,1} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \alpha_{11,2} & \alpha_{12,2} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} \\
&\quad + \begin{bmatrix} 1 & 0 \\ -\alpha_{21,0} & 1 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} + \begin{bmatrix} m_{11,1} & m_{12,1} \\ m_{21,1} & m_{22,1} \end{bmatrix} \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix} \\
&\quad + \begin{bmatrix} m_{11,2} & m_{12,2} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_{1,t-2} \\ u_{2,t-2} \end{bmatrix}.
\end{aligned}$$

In this model, the Kronecker indices (row degrees) are $p_1 = 2$ and $p_2 = 1$. Thus, the McMillan degree is 3. The p_{ki} numbers are

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}$$

(see (12.1.14)). The off-diagonal elements p_{12} and p_{21} of this matrix indicate the numbers of parameters contained in the operators $\alpha_{12}(L)$ and $\alpha_{21}(L)$, respectively. Because $\alpha_{12}(L)$ belongs to the first row or first equation of the system, it has degree $p_1 = 2$. Hence, because it has just one free coefficient ($p_{12} = 1$), it has the form $\alpha_{12}(L) = -\alpha_{12,2}L^2$. Similarly, $\alpha_{21}(L)$ belongs to the second row of the system and, thus, it has degree $p_2 = 1$. Because it has $p_{21} = 2$ free coefficients, it must be of the form $\alpha_{21}(L) = -\alpha_{21,0} - \alpha_{21,1}L$. Another characteristic feature of the echelon form is that A_0 is lower-triangular and has ones on the main diagonal. Moreover, the zero order MA coefficient matrix is identical to the zero order VAR matrix, $M_0 = A_0$.

Some free coefficients of the echelon form of a VARMA model may be zero and, hence, p or q may be less than $\max(p_1, \ldots, p_K)$. For instance, in the example process (12.1.15), $m_{11,2}$ and $m_{12,2}$ may be zero. In that case, $q = 1 < \max(p_1, p_2) = 2$. In order for a representation to be an echelon form with Kronecker indices $(p_1, \ldots, p_K)$, at least one operator in the k-th row of $[A(L) : M(L)]$ must have degree p_k, with nonzero coefficient at lag p_k.

An echelon is a certain positioning of an army in the form of steps. Similarly, the nonzero parameters in an echelon VARMA representation are positioned in a specific way. In particular, the positioning of freely varying parameters in the k-th equation depends only on Kronecker indices $p_i \leq p_k$ and not on Kronecker indices $p_j > p_k$. More precisely, as long as $p_j > p_k$, the positioning of the free parameters in the k-th equation will be the same for any value p_j. For the example process (12.1.15), it is easy to check that the positions of the free parameters in the second equation will remain the same if the row degree of the first equation is increased to $p_1 = 3$. In other words, p_{21} does not change due to an increase in p_1.

It can be shown that the echelon form, just like the final equations form, guarantees uniqueness of the VARMA representation. In other words, if a VARMA representation is in echelon form, then the representation is unique

within the class of all echelon representations. A similar statement applies for the final equations form. Also, for any stable, invertible VARMA(p, q) representation, there exists an equivalent echelon form and an equivalent final equations form.

The reader may wonder why we consider the complicated looking echelon representation although the final equations form serves the same purpose. The reason is that the echelon form is usually preferable in practice because it often involves fewer free parameters than the equivalent final equations form. We will see an example of this phenomenon shortly. Having as few free parameters as possible is important to ease the numerical problems in maximizing the likelihood function and to gain efficiency of the parameter estimators.

There are a number of other unique or *identified* parameterizations of VARMA models. We have chosen to present the final equations form and the echelon form because these two forms will play a role when we discuss the issue of specifying VARMA models in Chapter 13. For proofs of the uniqueness of the echelon form and for other identification conditions we refer to Hannan (1969, 1970, 1976, 1979), Deistler & Hannan (1981), and Hannan & Deistler (1988). We now proceed with illustrations of the final equations form and the echelon form.

12.1.3 Illustrations

Starting from some VARMA(p, q) representation $A(L)y_t = M(L)u_t$, one strategy for finding the corresponding final equations form results from premultiplying with the adjoint $A(L)^{adj}$ of the VAR operator $A(L)$ which gives

$$|A(L)|y_t = A(L)^{adj}M(L)u_t, \tag{12.1.16}$$

where $A(L)^{adj}A(L) = |A(L)|$ has been used. Obviously, (12.1.16) has a scalar VAR operator and, hence, is in final equations form if all superfluous terms are cancelled.

To find the echelon form corresponding to a given VARMA model, we have to cancel as much as possible so as to make the VAR and MA operators left-coprime. Then a unimodular matrix operator has to be determined which, upon premultiplication, transforms the given model into an echelon form. It usually helps to determine the Kronecker indices (row degrees) and the corresponding numbers p_{ki} first. We will now consider examples.

Let us begin with the simple bivariate process

$$\left(I_2 - \begin{bmatrix} 0 & \alpha \\ 0 & 1 \end{bmatrix} L\right) y_t = u_t \tag{12.1.17}$$

with $\alpha \neq 0$. Noting that

$$|A(L)| = \left|\begin{bmatrix} 1 & -\alpha L \\ 0 & 1 \end{bmatrix}\right| = 1 \quad \text{and} \quad A(L)^{adj} = \begin{bmatrix} 1 & \alpha L \\ 0 & 1 \end{bmatrix},$$

the final equations form is seen to be

$$y_t = \left(I_2 + \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} L \right) u_t. \tag{12.1.18}$$

To find the echelon representation, we first determine the Kronecker indices or row degrees and the implied p_{ki} from Definition 12.2. The first row of (12.1.17) has degree $p_1 = 1$ and the second row has degree $p_2 = 0$. Hence,

$$p_{11} = 1, \quad p_{12} = 0, \quad p_{21} = 1, \quad p_{22} = 0,$$

so that

$$\alpha_{11}(L) = 1 - \alpha_{11,1}L,\ \alpha_{12}(L) = 0,\ \alpha_{21}(L) = -\alpha_{21,0},\ \text{and}\ \alpha_{22}(L) = 1.$$

Thus, the echelon form is

$$\begin{bmatrix} 1 - \alpha_{11,1}L & 0 \\ -\alpha_{21,0} & 1 \end{bmatrix} y_t = \begin{bmatrix} 1 + m_{11,1}L & m_{12,1}L \\ -\alpha_{21,0} & 1 \end{bmatrix} u_t. \tag{12.1.19}$$

The unique parameter values in this representation corresponding to the specific process (12.1.17) are easily seen to be

$$\alpha_{11,1} = \alpha_{21,0} = m_{11,1} = 0 \quad \text{and} \quad m_{12,1} = \alpha.$$

Thus, in this particular case, the final equations form and the echelon form coincide.

As another example, we consider a 3-dimensional process with VARMA$(2, 1)$ representation

$$\begin{bmatrix} 1 - \theta_1 L & -\theta_2 L & 0 \\ 0 & 1 - \theta_3 L - \theta_4 L^2 & -\theta_5 L \\ 0 & 0 & 1 \end{bmatrix} y_t$$

$$= \begin{bmatrix} 1 - \eta_1 L & 0 & 0 \\ 0 & 1 - \eta_2 L & 0 \\ 0 & 0 & 1 - \eta_3 L \end{bmatrix} u_t. \tag{12.1.20}$$

Using (12.1.16), its final equations form is seen to be

$$(1 - \theta_1 L)(1 - \theta_3 L - \theta_4 L^2) y_t$$

$$= \begin{bmatrix} 1 - \theta_3 L - \theta_4 L^2 & \theta_2 L & \theta_2 \theta_5 L^2 \\ 0 & 1 - \theta_1 L & \theta_5 L - \theta_1 \theta_5 L^2 \\ 0 & 0 & (1 - \theta_1 L)(1 - \theta_3 L - \theta_4 L^2) \end{bmatrix}$$

$$\times \begin{bmatrix} 1 - \eta_1 L & 0 & 0 \\ 0 & 1 - \eta_2 L & 0 \\ 0 & 0 & 1 - \eta_3 L \end{bmatrix} u_t$$

which is easily recognizable as a VARMA(3, 4) structure with scalar VAR operator.

The Kronecker indices, that is, the row degrees of (12.1.20) are $(p_1, p_2, p_3) = (1, 2, 1)$ and the implied p_{ki}-numbers from (12.1.14) are collected in the following matrix:

$$[p_{ki}]_{k,i=1,2,3} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 2 & 1 \end{bmatrix}.$$

Consequently, the VAR operator of the echelon form becomes

$$\begin{bmatrix} 1-\alpha_{11,1}L & -\alpha_{12,1}L & -\alpha_{13,1}L \\ -\alpha_{21,2}L^2 & 1-\alpha_{22,1}L-\alpha_{22,2}L^2 & -\alpha_{23,2}L^2 \\ -\alpha_{31,1}L & -\alpha_{32,0}-\alpha_{32,1}L & 1-\alpha_{33,1}L \end{bmatrix}$$

or

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\alpha_{32,0} & 1 \end{bmatrix} - \begin{bmatrix} \alpha_{11,1} & \alpha_{12,1} & \alpha_{13,1} \\ 0 & \alpha_{22,1} & 0 \\ \alpha_{31,1} & \alpha_{32,1} & \alpha_{33,1} \end{bmatrix} L - \begin{bmatrix} 0 & 0 & 0 \\ \alpha_{21,2} & \alpha_{22,2} & \alpha_{23,2} \\ 0 & 0 & 0 \end{bmatrix} L^2. \tag{12.1.21}$$

Hence, in the echelon representation,

$$A_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\alpha_{32,0} & 1 \end{bmatrix}$$

is different from I_3, if $\alpha_{32,0} \neq 0$, and, thus, $M_0 = A_0$ is also not the identity matrix. The MA operator is

$$\begin{bmatrix} 1+m_{11,1}L & m_{12,1}L & m_{13,1}L \\ m_{21,1}L+m_{21,2}L^2 & 1+m_{22,1}L+m_{22,2}L^2 & m_{23,1}L+m_{23,2}L^2 \\ m_{31,1}L & -\alpha_{32,0}+m_{32,1}L & 1+m_{33,1}L \end{bmatrix}$$

or

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\alpha_{32,0} & 1 \end{bmatrix} + \begin{bmatrix} m_{11,1} & m_{12,1} & m_{13,1} \\ m_{21,1} & m_{22,1} & m_{23,1} \\ m_{31,1} & m_{32,1} & m_{33,1} \end{bmatrix} L + \begin{bmatrix} 0 & 0 & 0 \\ m_{21,2} & m_{22,2} & m_{23,2} \\ 0 & 0 & 0 \end{bmatrix} L^2. \tag{12.1.22}$$

The reader may be puzzled by the fact that the last element in the second row of (12.1.21) does not involve a term with first power of L while such a term appears in (12.1.20). This model form shows that there is a VARMA representation equivalent to (12.1.20) with the second but not the first power of L in the last operator in the second row of $A(L)$. The fact, that there always

exists an equivalent echelon representation does not mean that there is always an immediately obvious relation between the coefficients of any given VARMA representation and its equivalent echelon form. However, in the present case it is fairly easy to relate the representations (12.1.20) and (12.1.21)/(12.1.22). Premultiplying (12.1.20) by the operator

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \theta_5 L \\ 0 & 0 & 1 \end{bmatrix} \tag{12.1.23}$$

results in a VAR operator

$$\begin{bmatrix} 1-\theta_1 L & -\theta_2 L & 0 \\ 0 & 1-\theta_3 L-\theta_4 L^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and the MA operator changes accordingly. Notice that the operator (12.1.23) has constant determinant and, of course, the resulting VARMA model is equivalent to (12.1.20). The relation between its coefficients and those of the echelon representation (12.1.21)/(12.1.22) is obvious:

$$\begin{aligned} &\alpha_{11,1}=\theta_1, \quad \alpha_{12,1}=\theta_2, \quad \theta_{13,1}=0, \\ &\alpha_{21,2}=0, \quad \alpha_{22,1}=\theta_3, \quad \alpha_{22,2}=\theta_4, \quad \alpha_{23,2}=0, \\ &\alpha_{31,1}=\alpha_{32,0}=\alpha_{32,1}=\alpha_{33,1}=0, \end{aligned}$$

and the relation between (12.1.22) and the coefficients of (12.1.20) is also apparent. Of course, if the zero coefficients are known, then this knowledge may be used to reduce the number of free coefficients in the echelon form.

In this example, the unrestricted final equations form has 3 AR coefficients and 36 MA coefficients. Thus, the unrestricted form contains 39 parameters, apart from white noise covariance coefficients. In contrast, the unrestricted echelon form (12.1.21)/(12.1.22) has only 23 free parameters and is therefore preferable in terms of parameter parsimony. Note that, in practice, the true coefficient values are unknown and we pick an identified structure, for example, a final equations form or an echelon form. At that stage, further parameter restrictions may not be available. Hence, if (12.1.20) is the actual data generation process we may pick a VARMA$(3,4)$ model with scalar AR operator if we decide to go with a final equations representation and we may choose the model (12.1.21)/(12.1.22) if we decide to use an echelon form representation. Obviously, the latter choice results in a more parsimonious parameterization. As mentioned earlier, for estimation purposes the more parsimonious representation is advantageous.

Although $A_0 \neq I$ in the previous example, it should be understood that in many echelon representations $A_0 = M_0 = I_K$. In particular, if the row degrees $p_1 = \cdots = p_K = p$, all $p_{ki} = p, i, k = 1, \ldots, K$, and the echelon form is easily seen to be a standard VARMA(p,p) model with $A_0 = M_0 = I_K$. We are now ready to turn to the actual estimation of the parameters of an identified VARMA model and we shall discuss its Gaussian likelihood function next.

12.2 The Gaussian Likelihood Function

For maximum likelihood (ML) estimation the likelihood function is needed. We will now derive useful approximations to the likelihood function of a Gaussian VARMA(p, q) process. Special case MA processes will be considered first.

12.2.1 The Likelihood Function of an MA(1) Process

Because a zero mean MA(1) process is the simplest member of the finite order MA family, we use that as a starting point. Hence, we assume to have a sample $y_1, \ldots, y_T$ which is generated by the Gaussian, K-dimensional, invertible MA(1) process

$$y_t = u_t + M_1 u_{t-1}, \tag{12.2.1}$$

where u_t is a Gaussian white noise process with covariance matrix Σ_u. Thus,

$$\mathbf{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} = \overline{\mathfrak{M}}_1 \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_T \end{bmatrix},$$

where

$$\overline{\mathfrak{M}}_1 := \begin{bmatrix} M_1 & I_K & 0 & \ldots & 0 & 0 \\ 0 & M_1 & I_K & & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & M_1 & I_K \end{bmatrix} \tag{12.2.2}$$

is a $(KT \times K(T+1))$ matrix. Using that u_t is Gaussian white noise and, thus,

$$\begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_T \end{bmatrix} \sim \mathcal{N}(0, I_{T+1} \otimes \Sigma_u),$$

if follows that

$$\mathbf{y} \sim \mathcal{N}(0, \overline{\mathfrak{M}}_1 (I_{T+1} \otimes \Sigma_u) \overline{\mathfrak{M}}_1')$$

and the likelihood function is seen to be

$$\begin{aligned} &l(M_1, \Sigma_u | \mathbf{y}) \\ &\propto |\overline{\mathfrak{M}}_1 (I_{T+1} \otimes \Sigma_u) \overline{\mathfrak{M}}_1'|^{-1/2} \ \exp\{-\tfrac{1}{2} \mathbf{y}' [\overline{\mathfrak{M}}_1 (I_{T+1} \otimes \Sigma_u) \overline{\mathfrak{M}}_1']^{-1} \mathbf{y}\}, \end{aligned} \tag{12.2.3}$$

where $\propto$ stands for "is proportional to". In other words, we have dropped a multiplicative constant from the likelihood function which does not change the maximizing values of M_1 and Σ_u.

It is inconvenient that this function involves the determinant and the inverse of a $(KT \times KT)$ matrix. A simpler form is obtained if u_0 is set to zero, that is, the MA(1) process is assumed to be started up with a nonrandom fixed vector $u_0 = 0$. In that case,

$$\mathbf{y} = \mathfrak{M}_1 \mathbf{u},$$

where

$$\mathfrak{M}_1 := \underset{(KT \times KT)}{\begin{bmatrix} I_K & 0 & \dots & 0 & 0 \\ M_1 & I_K & & 0 & 0 \\ & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & M_1 & I_K \end{bmatrix}} \quad \text{and} \quad \mathbf{u} := \underset{(KT \times 1)}{\begin{bmatrix} u_1 \\ \vdots \\ u_T \end{bmatrix}}. \tag{12.2.4}$$

The likelihood function is then proportional to

$$\begin{aligned} l_0(M_1, \Sigma_u | \mathbf{y}) &= |\mathfrak{M}_1 (I_T \otimes \Sigma_u) \mathfrak{M}_1'|^{-1/2} \exp\{-\tfrac{1}{2} \mathbf{y}' [\mathfrak{M}_1 (I_T \otimes \Sigma_u) \mathfrak{M}_1']^{-1} \mathbf{y}\} \\ &= |\Sigma_u|^{-T/2} \exp\{-\tfrac{1}{2} \mathbf{y}' \mathfrak{M}_1'^{-1} (I_T \otimes \Sigma_u^{-1}) \mathfrak{M}_1^{-1} \mathbf{y}\} \\ &= |\Sigma_u|^{-T/2} \exp\left\{-\frac{1}{2} \sum_{t=1}^{T} u_t' \Sigma_u^{-1} u_t\right\}, \end{aligned} \tag{12.2.5}$$

where it has been used that $|\mathfrak{M}_1| = 1$ and

$$\begin{aligned} \mathfrak{M}_1^{-1} &= \begin{bmatrix} I_K & 0 & \dots & 0 & 0 \\ -M_1 & I_K & & 0 & 0 \\ (-M_1)^2 & -M_1 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ (-M_1)^{T-1} & (-M_1)^{T-2} & \dots & -M_1 & I_K \end{bmatrix} \\ &= \begin{bmatrix} I_K & 0 & \dots & 0 \\ -\Pi_1 & I_K & & 0 \\ \vdots & & \ddots & \vdots \\ -\Pi_{T-1} & -\Pi_{T-2} & \dots & I_K \end{bmatrix}, \end{aligned}$$

where the $\Pi_i = -(-M_1)^i$ are the coefficients of the pure VAR representation of the process. By successive substitution, the MA(1) process in (12.2.1) can be rewritten as

$$y_t + \sum_{i=1}^{t-1} (-M_1)^i y_{t-i} + (-M_1)^t u_0 = u_t. \tag{12.2.6}$$

Thus, if $u_0 = 0$,

$$u_t = y_t + \sum_{i=1}^{t-1} (-M_1)^i y_{t-i},$$

from which the last expression in (12.2.5) is obtained.

The equation (12.2.6) also shows that, for large t, the assumption regarding u_0 becomes inconsequential because, for an invertible process, M_1^t approaches zero as $t \to \infty$. The impact of u_0 disappears more rapidly for processes for which M_1^t goes to zero more rapidly as t gets large. In other words, if all eigenvalues of M_1 are close to zero or, equivalently, all roots of $\det(I_K + M_1 z)$ are far outside the unit circle, then the impact of u_0 is lower than for processes with roots close to the unit circle. In summary, the likelihood approximation in (12.2.5) will improve as the sample size gets large and will become exact as $T \to \infty$. In small samples, it is better for processes with roots of $\det(I_K + M_1 z)$ far away from the unit circle than for those with roots close to the noninvertibility region. Because we will be concerned predominantly with large sample properties in the following, we will often work with likelihood approximations such as l_0 in (12.2.5).

12.2.2 The MA(q) Case

A similar reasoning as for MA(1) processes can also be employed for higher order MA processes. Suppose the generation process of y_t has a zero mean MA(q) representation

$$y_t = u_t + M_1 u_{t-1} + \cdots + M_q u_{t-q}. \tag{12.2.7}$$

Then

$$\mathbf{y} = \overline{\mathfrak{M}}_q \begin{bmatrix} u_{-q+1} \\ \vdots \\ u_0 \\ u_1 \\ \vdots \\ u_T \end{bmatrix},$$

where

$$\overline{\mathfrak{M}}_q := \begin{bmatrix} M_q & M_{q-1} & \dots & M_1 & I_K & 0 & \dots & \dots & 0 \\ 0 & M_q & \dots & M_2 & M_1 & I_K & & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & & M_q & \dots & M_2 & M_1 & I_K \end{bmatrix} \tag{12.2.8}$$

is a $(KT \times K(T+q))$ matrix and the exact likelihood for a sample of size T is seen to be

$$l(M_1, \ldots, M_q, \Sigma_u | \mathbf{y}) \propto |\overline{\mathfrak{M}}_q (I_{T+q} \otimes \Sigma_u) \overline{\mathfrak{M}}_q'|^{-1/2} \times \exp\{-\tfrac{1}{2} \mathbf{y}' [\overline{\mathfrak{M}}_q (I_{T+q} \otimes \Sigma_u) \overline{\mathfrak{M}}_q']^{-1} \mathbf{y}\}. \quad (12.2.9)$$

Again a convenient approximation to the likelihood function is obtained by setting $u_{-q+1} = \cdots = u_0 = 0$. In that case, the likelihood is, apart from a multiplicative constant,

$$l_0(M_1, \ldots, M_q, \Sigma_u | \mathbf{y}) = |\Sigma_u|^{-T/2} \exp\{-\tfrac{1}{2} \mathbf{y}' [\mathfrak{M}_q'^{-1} (I_T \otimes \Sigma_u^{-1}) \mathfrak{M}_q^{-1}] \mathbf{y}\}, \quad (12.2.10)$$

where

$$\mathfrak{M}_q := \begin{bmatrix} I_K & 0 & \ldots & \ldots & \ldots & \ldots & 0 & 0 \\ M_1 & I_K & & & & & 0 & 0 \\ M_2 & M_1 & \ddots & & & & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & & & & \vdots \\ M_q & M_{q-1} & & \ddots & \ddots & & & \vdots \\ 0 & M_q & & & \ddots & \ddots & & \\ \vdots & & \ddots & & & \ddots & \ddots & \\ 0 & 0 & \ldots & M_q & \ldots & \ldots & M_1 & I_K \end{bmatrix} \quad (12.2.11)$$

and, hence,

$$\mathfrak{M}_q^{-1} = \begin{bmatrix} I_K & 0 & \ldots & 0 \\ -\Pi_1 & I_K & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\Pi_{T-1} & -\Pi_{T-2} & \ldots & I_K \end{bmatrix}.$$

Here the Π_i are the coefficient matrices of the pure VAR representation of the process y_t. Thus, the Π_i can be computed recursively as in Section 11.2 of Chapter 11.

An alternative expression for the approximate likelihood is easily seen to be

$$l_0(M_1, \ldots, M_q, \Sigma_u | \mathbf{y}) = |\Sigma_u|^{-T/2} \exp\left\{-\frac{1}{2} \sum_{t=1}^{T} u_t' \Sigma_u^{-1} u_t\right\}, \quad (12.2.12)$$

where

$$u_t = y_t - \sum_{i=1}^{t-1} \Pi_i y_{t-i}.$$

Again, the likelihood approximation will be quite precise if T is reasonably large and the roots of $\det(I_K + M_1 z + \cdots + M_q z^q)$ are not close to the unit circle.

Although we will work with likelihood approximations in the following, it is perhaps worth noting that an expression for the exact likelihood of an MA(q) process can be derived that is more manageable than the one in (12.2.9) (see, e.g., Hillmer & Tiao (1979), Kohn (1981)).

12.2.3 The VARMA(1, 1) Case

Before we tackle general mixed VARMA models, we shall consider the simplest candidate, namely a Gaussian zero mean, stationary, stable, and invertible VARMA(1, 1) process,

$$y_t = A_1 y_{t-1} + u_t + M_1 u_{t-1}. \tag{12.2.13}$$

Assuming that we have a sample $y_1, \ldots, y_T$, generated by this process and defining

$$\mathfrak{A}_p := \begin{bmatrix} I_K & 0 & & \cdots & & 0 & 0 \\ -A_1 & I_K & \ddots & & & 0 & 0 \\ -A_2 & -A_1 & \ddots & \ddots & & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ -A_p & -A_{p-1} & & \ddots & \ddots & 0 & 0 \\ 0 & -A_p & \ddots & & \ddots & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots & \vdots \\ 0 & 0 & & \ddots & & I_K & 0 \\ 0 & 0 & \cdots & -A_p & \cdots & -A_1 & I_K \end{bmatrix}, \tag{12.2.14}$$

we get

$$\mathfrak{A}_1 \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} + \begin{bmatrix} -A_1 y_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \overline{\mathfrak{M}}_1 \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_T \end{bmatrix}.$$

Hence, for given, fixed presample values y_0,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} \sim \mathcal{N}(\mathfrak{A}_1^{-1}\mathbf{y}_0, \mathfrak{A}_1^{-1}\overline{\mathfrak{M}}_1 (I_{T+1} \otimes \Sigma_u)\overline{\mathfrak{M}}_1' \mathfrak{A}_1'^{-1}), \tag{12.2.15}$$

where

$$\mathbf{y}_0 := \begin{bmatrix} A_1 y_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The corresponding likelihood function, conditional on y_0, is

$$\begin{aligned} & l(A_1, M_1, \Sigma_u | \mathbf{y}, y_0) \\ & \quad \propto \ |\mathfrak{A}_1^{-1} \overline{\mathfrak{M}}_1 (I_{T+1} \otimes \Sigma_u) \overline{\mathfrak{M}}_1' \mathfrak{A}_1'^{-1}|^{-1/2} \\ & \qquad \times \exp\{-\tfrac{1}{2}(\mathbf{y} - \mathfrak{A}_1^{-1}\mathbf{y}_0)' \mathfrak{A}_1' [\overline{\mathfrak{M}}_1 (I_{T+1} \otimes \Sigma_u) \overline{\mathfrak{M}}_1']^{-1} \mathfrak{A}_1 (\mathbf{y} - \mathfrak{A}_1^{-1}\mathbf{y}_0)\} \\ & \quad = \ |\overline{\mathfrak{M}}_1 (I_{T+1} \otimes \Sigma_u) \overline{\mathfrak{M}}_1'|^{-1/2} \\ & \qquad \times \exp\{-\tfrac{1}{2}(\mathfrak{A}_1 \mathbf{y} - \mathbf{y}_0)' [\overline{\mathfrak{M}}_1 (I_{T+1} \otimes \Sigma_u) \overline{\mathfrak{M}}_1']^{-1} (\mathfrak{A}_1 \mathbf{y} - \mathbf{y}_0)\}, \quad (12.2.16) \end{aligned}$$

where $|\mathfrak{A}_1| = 1$ has been used.

With the same arguments as in the pure MA case, a simple approximation is obtained by setting $u_0 = y_0 = 0$. Then we get

$$\begin{aligned} l_0(A_1, M_1, \Sigma_u) &= |\Sigma_u|^{-T/2} \exp\{-\tfrac{1}{2}(\mathfrak{M}_1^{-1}\mathfrak{A}_1 \mathbf{y})' (I_T \otimes \Sigma_u^{-1}) \mathfrak{M}_1^{-1} \mathfrak{A}_1 \mathbf{y}\} \\ &= |\Sigma_u|^{-T/2} \exp\left\{-\frac{1}{2} \sum_{t=1}^{T} u_t' \Sigma_u^{-1} u_t\right\}, \quad (12.2.17) \end{aligned}$$

where

$$u_t = y_t - \sum_{i=1}^{t-1} \Pi_i y_{t-i} \quad (12.2.18)$$

and the Π_i are the coefficient matrices of the pure VAR representation, that is, for the present case $\Pi_i = (-1)^{i-1}(M_1^i + M_1^{i-1} A_1)$, $i = 1, 2, \ldots$ (see Section 11.3.1). Note that in writing the likelihood approximation l_0 we have dropped the conditions $\mathbf{y}$ and y_0 for notational simplicity.

The effect of starting up the process with $y_0 = u_0 = 0$ is quite easily seen in (12.2.18), namely, for observation y_t, the infinite order pure VAR representation is truncated at lag $t - 1$. Such a truncation has little effect if the sample size is large and the roots of the MA operator are not close to the unit circle.

12.2.4 The General VARMA(p, q) Case

Now suppose a sample $y_1, \ldots, y_T$ is generated by the Gaussian K-dimensional, stable, invertible VARMA(p, q) process

$$\begin{aligned} A_0(y_t - \mu) &= A_1(y_{t-1} - \mu) + \cdots + A_p(y_{t-p} - \mu) \\ &\quad + A_0 u_t + M_1 u_{t-1} + \cdots + M_q u_{t-q} \quad (12.2.19) \end{aligned}$$

with mean vector μ and nonsingular white noise covariance matrix Σ_u. Notice that A_0 appears as the coefficient matrix of y_t and of u_t as in the echelon form. Thus, the echelon form is covered by our treatment of the general VARMA(p, q) case. We have chosen the mean-adjusted form of the process because this form has certain advantages in ML estimation, as we will see later.

Usually some elements of the coefficient matrices will be zero or obey some other type of restrictions. Therefore, to be realistic, we define

$$\alpha_0 := \text{vec}(A_0) \text{ and } \boldsymbol{\beta} := \text{vec}[A_1, \ldots, A_p, M_1, \ldots, M_q] \tag{12.2.20}$$

and assume that these coefficients are linearly related to an $(N \times 1)$ parameter vector $\boldsymbol{\gamma}$, that is,

$$\begin{bmatrix} \alpha_0 \\ \boldsymbol{\beta} \end{bmatrix} = R\boldsymbol{\gamma} + r \tag{12.2.21}$$

for a suitable, known $(K^2(p+q+1) \times N)$ matrix R and a known $K^2(p+q+1)$-vector r. For example, for a bivariate ARMA$_E(1, 0)$ process with Kronecker indices $p_1 = 1$ and $p_2 = 0$,

$$\begin{bmatrix} 1 - \alpha_{11,1}L & 0 \\ -\alpha_{21,0} & 1 \end{bmatrix}(y_t - \mu) = \begin{bmatrix} 1 + m_{11,1}L & m_{12,1}L \\ -\alpha_{21,0} & 1 \end{bmatrix} u_t$$

or

$$\begin{aligned} \begin{bmatrix} 1 & 0 \\ -\alpha_{21,0} & 1 \end{bmatrix}(y_t - \mu) &= \begin{bmatrix} \alpha_{11,1} & 0 \\ 0 & 0 \end{bmatrix}(y_{t-1} - \mu) + \begin{bmatrix} 1 & 0 \\ -\alpha_{21,0} & 1 \end{bmatrix} u_t \\ &\quad + \begin{bmatrix} m_{11,1} & m_{12,1} \\ 0 & 0 \end{bmatrix} u_{t-1}, \end{aligned}$$

we have

$$\alpha_0 = \begin{bmatrix} 1 \\ -\alpha_{21,0} \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \alpha_{11,1} \\ 0 \\ 0 \\ 0 \\ m_{11,1} \\ 0 \\ m_{12,1} \\ 0 \end{bmatrix}, \quad R = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \alpha_{21,0} \\ \alpha_{11,1} \\ m_{11,1} \\ m_{12,1} \end{bmatrix}, \quad \text{and} \quad r = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Similarly, for the final equations form

$$(1 - \alpha_1 L)(y_t - \mu) = \begin{bmatrix} 1 + m_{11}L & m_{12}L \\ m_{21}L & 1 + m_{22}L \end{bmatrix} u_t$$

or

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}(y_t - \mu) = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_1 \end{bmatrix}(y_{t-1} - \mu) + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} u_t + \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} u_{t-1},$$

we get

$$\alpha_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \alpha_1 \\ 0 \\ 0 \\ \alpha_1 \\ m_{11} \\ m_{21} \\ m_{12} \\ m_{22} \end{bmatrix}, \quad R = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \alpha_1 \\ m_{11} \\ m_{21} \\ m_{12} \\ m_{22} \end{bmatrix}, \quad \text{and} \quad r = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The likelihood function is a function of $\mu, \boldsymbol{\gamma}$, and Σ_u. Its exact form, given fixed initial values $y_{-p+1}, \ldots, y_0$, can be derived analogously to the previously considered special cases (see Problem 12.4 and Hillmer & Tiao (1979)). Here we will just give the likelihood approximation obtained by assuming

$$y_{-p+1} - \mu = \cdots = y_0 - \mu = u_{-q+1} = \cdots = u_0 = 0.$$

Apart from a multiplicative constant, we get

$$l_0(\mu, \boldsymbol{\gamma}, \Sigma_u) = |\Sigma_u|^{-T/2} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} u_t(\mu, \boldsymbol{\gamma})' \Sigma_u^{-1} u_t(\mu, \boldsymbol{\gamma})\right\}, \tag{12.2.22}$$

where

$$u_t(\mu, \boldsymbol{\gamma}) = (y_t - \mu) - \sum_{i=1}^{t-1} \Pi_i(\boldsymbol{\gamma})(y_{t-i} - \mu), \tag{12.2.23}$$

with the $\Pi_i(\boldsymbol{\gamma})$'s being again the coefficient matrices of the pure VAR representation of y_t. We have indicated that these matrices are determined by the parameter vector $\boldsymbol{\gamma}$. Formally the likelihood approximation has the same appearance as in the special cases. Of course, the u_t's are now potentially more complicated functions of the parameters.

It is perhaps worth noting that the uniqueness or identification problem discussed in Section 12.1 is reflected in the likelihood function. If the model is parameterized in a unique way, for instance, in final equations form or echelon form, the likelihood function has a locally unique maximum. This property is of obvious importance to guarantee unique ML estimators. Note, however, that the likelihood function in general has more than one local maximum. A more detailed discussion of the properties of the likelihood function can be found in Deistler & Pötscher (1984).

The next section focuses on the maximization of the approximate likelihood function (12.2.22) or, equivalently, the maximization of its logarithm,

$$\ln l_0(\mu, \boldsymbol{\gamma}, \Sigma_u) = -\frac{T}{2} \ln |\Sigma_u| - \frac{1}{2}\sum_{t=1}^{T} u_t(\mu, \boldsymbol{\gamma})' \Sigma_u^{-1} u_t(\mu, \boldsymbol{\gamma}). \tag{12.2.24}$$

12.3 Computation of the ML Estimates

In the pure finite order VAR case considered in Chapters 3 and 5, we have obtained the ML estimates by solving the normal equations. In the presently considered VARMA(p, q) case, we may use the same principle. In other words, we determine the first order partial derivatives of the log-likelihood function or rather its approximation given in (12.2.24) and equate them to zero. We will obtain the normal equations in Section 12.3.1. It turns out that they are nonlinear in the parameters and we discuss algorithms for solving the ML optimization problem in Section 12.3.2. The optimization procedures are iterative algorithms that require starting-up values or preliminary estimates for the parameters. A possible choice of initial estimates is proposed in Section 12.3.4. One of the optimization algorithms involves the information matrix which is given in Section 12.3.3. An example is discussed in Section 12.3.5.

12.3.1 The Normal Equations

In order to set up the normal equations corresponding to the approximate log-likelihood given in (12.2.24), we derive the first order partial derivatives with respect to all the parameters $\mu, \boldsymbol{\gamma}$, and Σ_u.

$$\frac{\partial \ln l_0}{\partial \mu'} = -\sum_{t=1}^{T} u_t' \Sigma_u^{-1} \frac{\partial u_t}{\partial \mu'} = \sum_{t=1}^{T} u_t' \Sigma_u^{-1} \left[I_K - \sum_{i=1}^{t-1} \Pi_i(\boldsymbol{\gamma}) \right], \tag{12.3.1}$$

$$\frac{\partial \ln l_0}{\partial \boldsymbol{\gamma}'} = -\sum_{t=1}^{T} u_t' \Sigma_u^{-1} \frac{\partial u_t}{\partial \boldsymbol{\gamma}'}. \tag{12.3.2}$$

A recursive formula for computing the $\partial u_t / \partial \boldsymbol{\gamma}'$ is given in the following lemma.

Lemma 12.1
Suppose $\mu = 0$ and let

$$u_t = y_t - A_0^{-1}[A_1 y_{t-1} + \cdots + A_p y_{t-p} + M_1 u_{t-1} + \cdots + M_q u_{t-q}], \tag{12.3.3}$$

$$\alpha_0 := \operatorname{vec}(A_0),$$

$$\boldsymbol{\beta} := \operatorname{vec}[A_1, \ldots, A_p, M_1, \ldots, M_q],$$

and suppose

$$\begin{bmatrix} \alpha_0 \\ \boldsymbol{\beta} \end{bmatrix} = R\boldsymbol{\gamma} + r, \tag{12.3.4}$$

where R is a known $(K^2(p+q+1) \times N)$ matrix, r is a known $K^2(p+q+1)$-dimensional vector, and $\boldsymbol{\gamma}$ is an $(N \times 1)$ vector of unknown parameters. Then, defining $\partial u_0 / \boldsymbol{\gamma}' = \partial u_{-1} / \partial \boldsymbol{\gamma}' = \cdots = \partial u_{-q+1} / \partial \boldsymbol{\gamma}' = 0$ and $y_0 = \cdots = y_{-p+1} = u_0 = \cdots = u_{-q+1} = 0$,

$$\begin{aligned} \frac{\partial u_t}{\partial \boldsymbol{\gamma}'} \;=\; & \{(A_0^{-1}[A_1 y_{t-1} + \cdots + A_p y_{t-p} \\ & + M_1 u_{t-1} + \cdots + M_q u_{t-q}])' \otimes A_0^{-1}\}[I_{K^2} : 0 : \cdots : 0]R \\ & -[(y_{t-1}', \ldots, y_{t-p}', u_{t-1}', \ldots, u_{t-q}') \otimes A_0^{-1}][0 : I_{K^2(p+q)}]R \\ & -A_0^{-1}\left[M_1 \frac{\partial u_{t-1}}{\partial \boldsymbol{\gamma}'} + \cdots + M_q \frac{\partial u_{t-q}}{\partial \boldsymbol{\gamma}'} \right], \end{aligned} \tag{12.3.5}$$

for $t = 1, \ldots, T$. ■

Replacing y_t with $y_t - \mu$ in this lemma, the expression in (12.3.5) can be used for recursively computing the $\partial u_t / \partial \boldsymbol{\gamma}'$ required in (12.3.2).

Proof:

$$\frac{\partial u_t}{\partial \boldsymbol{\gamma}'} = -[(A_1 y_{t-1} + \cdots + A_p y_{t-p} + M_1 u_{t-1} + \cdots + M_q u_{t-q})' \otimes I_K]$$

$$\times \frac{\partial \operatorname{vec}(A_0^{-1})}{\partial \boldsymbol{\gamma}'}$$
$$-[(y'_{t-1},\ldots,y'_{t-p},u'_{t-1},\ldots,u'_{t-q})\otimes A_0^{-1}]$$
$$\times \frac{\partial \operatorname{vec}[A_1,\ldots,A_p,M_1,\ldots,M_q]}{\partial \boldsymbol{\gamma}'}$$
$$-A_0^{-1}[A_1,\ldots,A_p,M_1,\ldots,M_q]\left[\partial\begin{bmatrix} y_{t-1}\\ \vdots\\ y_{t-p}\\ u_{t-1}\\ \vdots\\ u_{t-q}\end{bmatrix}/\partial\boldsymbol{\gamma}'\right]. \tag{12.3.6}$$

The lemma follows by noting that

$$\frac{\partial \operatorname{vec}(A_0^{-1})}{\partial \boldsymbol{\gamma}'} = \frac{\partial \operatorname{vec}(A_0^{-1})}{\partial \alpha_0'}\frac{\partial \alpha_0}{\partial \boldsymbol{\gamma}'} = -[(A_0^{-1})'\otimes A_0^{-1}][I_{K^2}:0:\cdots:0]R \tag{12.3.7}$$

(see Rule (9) of Appendix A.13). ■

The partial derivatives of the approximate log-likelihood with respect to the elements of Σ_u are

$$\frac{\partial \ln l_0}{\partial \Sigma_u} = -\frac{T}{2}\Sigma_u^{-1} + \frac{1}{2}\Sigma_u^{-1}\left(\sum_{t=1}^{T} u_t u_t'\right)\Sigma_u^{-1} \tag{12.3.8}$$

(see Problem 12.5). Setting this expression to zero and solving for Σ_u gives

$$\widetilde{\Sigma}_u(\mu,\boldsymbol{\gamma}) = \frac{1}{T}\sum_{t=1}^{T} u_t(\mu,\boldsymbol{\gamma})u_t(\mu,\boldsymbol{\gamma})'. \tag{12.3.9}$$

Substituting for Σ_u in (12.3.1) and (12.3.2) and setting to zero results in a generally nonlinear set of normal equations which may be solved by numerical methods. Before we discuss a possible algorithm, it may be worth pointing out that by substituting $\widetilde{\Sigma}_u(\mu,\boldsymbol{\gamma})$ for Σ_u in $\ln l_0$, we get

$$\begin{aligned}\ln l_0(\mu,\boldsymbol{\gamma}) &= -\frac{T}{2}\ln|\widetilde{\Sigma}_u(\mu,\boldsymbol{\gamma})| - \frac{1}{2}\operatorname{tr}\left(\widetilde{\Sigma}_u(\mu,\boldsymbol{\gamma})^{-1}\sum_{t=1}^{T} u_t(\mu,\boldsymbol{\gamma})u_t(\mu,\boldsymbol{\gamma})'\right)\\ &= -\frac{T}{2}\ln|\widetilde{\Sigma}_u(\mu,\boldsymbol{\gamma})| - \frac{TK}{2}.\end{aligned} \tag{12.3.10}$$

Thus, instead of maximizing $\ln l_0$ we may equivalently minimize

$$\ln|\widetilde{\Sigma}_u(\mu,\boldsymbol{\gamma})| \quad \text{or} \quad |\widetilde{\Sigma}_u(\mu,\boldsymbol{\gamma})|. \tag{12.3.11}$$

12.3.2 Optimization Algorithms

The problem of optimizing (minimizing or maximizing) a function arises not only in ML estimation but also in various other contexts. Therefore, general algorithms have been developed. Following Judge et al. (1985, Section B.2), we will give a brief introduction to so-called gradient algorithms and then address the specific problem at hand. With the objective in mind that we want to find the coefficient values that minimize $-\ln l_0$ or $\ln|\widetilde{\Sigma}_u(\mu,\boldsymbol{\gamma})|$, we assume that the problem is to minimize a twice continuously differentiable, scalar valued function $h(\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is some $(N \times 1)$ vector.

Given a vector $\boldsymbol{\gamma}_i$ in the parameter space, we are looking for a direction (vector) $\mathbf{d}$ in which the objective function declines. Then we can perform a step of length s, say, in that direction which will take us downhill. In other words, we seek an appropriate *step direction* $\mathbf{d}$ and a step length s such that

$$h(\boldsymbol{\gamma}_i + s\mathbf{d}) < h(\boldsymbol{\gamma}_i). \tag{12.3.12}$$

If $\mathbf{d}$ is a downhill direction, a small step in that direction will always decrease the objective function. Thus, we are seeking a $\mathbf{d}$ such that $h(\boldsymbol{\gamma}_i + s\mathbf{d})$ is a decreasing function of s, for s sufficiently close to zero. In other words, $\mathbf{d}$ must be such that

$$0 > \frac{dh(\boldsymbol{\gamma}_i + s\mathbf{d})}{ds}\bigg|_{s=0} = \left[\frac{\partial h(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'}\bigg|_{\boldsymbol{\gamma}_i}\right]\left[\frac{\partial(\boldsymbol{\gamma}_i + s\mathbf{d})}{\partial s}\bigg|_{s=0}\right] = \left[\frac{\partial h(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'}\bigg|_{\boldsymbol{\gamma}_i}\right]\mathbf{d}.$$

Using the abbreviation

$$\mathbf{h}_i := \frac{\partial h(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}_i}$$

for the gradient of $h(\boldsymbol{\gamma})$ at $\boldsymbol{\gamma}_i$, a possible choice of $\mathbf{d}$ is

$$\mathbf{d} = -D_i\mathbf{h}_i,$$

where D_i is any positive definite matrix. With this choice of $\mathbf{d}$,

$$\mathbf{h}_i'\mathbf{d} = -\mathbf{h}_i'D_i\mathbf{h}_i < 0$$

if $\mathbf{h}_i \neq 0$. Because the gradient is zero at a local minimum of the function, we hope to have reached the minimum once $\mathbf{h}_i = 0$ and, hence, $\mathbf{d} = 0$. The general form of an iteration of a *gradient algorithm* is therefore

$$\boldsymbol{\gamma}_{i+1} = \boldsymbol{\gamma}_i - s_iD_i\mathbf{h}_i, \tag{12.3.13}$$

where s_i denotes the step length in the i-th iteration and D_i is a positive definite *direction matrix*. The name "gradient algorithm" stems from the fact that the gradient $\mathbf{h}_i$ is involved in the choice of the step direction. Many such

algorithms have been proposed in the literature (see, for example, Judge et al. (1985, Section B.2)). They differ in their choice of the direction matrix D_i and the step length s_i.

To motivate the choice of the D_i matrix that will be considered in the ML algorithm presented below, we expand the objective function $h(\boldsymbol{\gamma})$ in a Taylor series about $\boldsymbol{\gamma}_i$ (see Appendix A.13, Proposition A.3),

$$h(\boldsymbol{\gamma}) \approx h(\boldsymbol{\gamma}_i) + \mathbf{h}_i'(\boldsymbol{\gamma} - \boldsymbol{\gamma}_i) + \tfrac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_i)' H_i (\boldsymbol{\gamma} - \boldsymbol{\gamma}_i), \tag{12.3.14}$$

where

$$H_i := \left.\frac{\partial^2 h}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'}\right|_{\boldsymbol{\gamma}_i}$$

is the Hessian matrix of second order partial derivatives of $h(\boldsymbol{\gamma})$, evaluated at $\boldsymbol{\gamma}_i$. If $h(\boldsymbol{\gamma})$ were a quadratic function, the right-hand side of (12.3.14) were exactly equal to $h(\boldsymbol{\gamma})$ and the first order conditions for a minimum would result by taking first order partial derivatives of the right-hand side and setting to zero:

$$\mathbf{h}_i' + H_i(\boldsymbol{\gamma} - \boldsymbol{\gamma}_i)' = 0$$

or

$$\boldsymbol{\gamma} = \boldsymbol{\gamma}_i - H_i^{-1}\mathbf{h}_i.$$

Thus, if $h(\boldsymbol{\gamma})$ were a quadratic function, starting from any vector $\boldsymbol{\gamma}_i$, we would reach the minimum in one step of length $s_i = 1$ by choosing the inverse Hessian as the direction matrix. In general, if $h(\boldsymbol{\gamma})$ is not a quadratic function, then the choice $D_i = H_i^{-1}$ is still reasonable once we are close to the minimum. Recall that a positive definite Hessian is the second order condition for a local minimum. Therefore, the inverse Hessian qualifies as a direction matrix. A gradient algorithm with the inverse Hessian as the direction matrix is called a *Newton* or *Newton-Raphson algorithm*.

From the previous subsection, we know that the first order partial derivatives of our objective function $-\ln l_0$ are quite complicated and, thus, finding the Hessian matrix of second order partial derivatives is even more complicated. Therefore we approximate the Hessian by an estimate of the information matrix,

$$\mathcal{I}(\boldsymbol{\gamma}) := E\left[\frac{\partial^2(-\ln l_0)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'}\right], \tag{12.3.15}$$

which is the expected value of the Hessian matrix. The estimate of $\mathcal{I}(\boldsymbol{\gamma})$ will be denoted by $\widehat{\mathcal{I}}(\boldsymbol{\gamma})$. A computable expression will be given in the next subsection. Because the true parameter vector $\boldsymbol{\gamma}$ is unknown, $\widehat{\mathcal{I}}(\boldsymbol{\gamma}_i)$ is used as an estimate of $\mathcal{I}(\boldsymbol{\gamma})$ in the i-th iteration step. Hence, for given mean vector μ

and white noise covariance matrix Σ_u, we get a minimization algorithm with i-th iteration step

$$\boldsymbol{\gamma}_{i+1} = \boldsymbol{\gamma}_i - s_i \widehat{\mathcal{I}}(\boldsymbol{\gamma}_i)^{-1} \left[\frac{\partial(-\ln l_0)}{\partial \boldsymbol{\gamma}} \bigg|_{\boldsymbol{\gamma}_i} \right]. \tag{12.3.16}$$

This algorithm is called the *scoring algorithm*.

As it stands, we still need some more information before we can execute this algorithm. First, we need a starting-up vector $\boldsymbol{\gamma}_1$ for the first iteration. This vector should be close to the minimizing vector to ensure that $\widehat{\mathcal{I}}(\boldsymbol{\gamma}_1)$ is positive definite and we make good progress towards the minimum even in the first iteration. We will consider one possible choice in Section 12.3.4.

Second, we have to choose the step length s_i. There are various possible alternatives (see, e.g., Judge et al. (1985, Section B.2)). Because we are just interested in the main principles of the algorithm, we will ignore the problem here and choose $s_i = 1$.

Third, the algorithm provides an ML estimate of $\boldsymbol{\gamma}$, conditional on some given Σ_u matrix and mean vector μ, because both the information matrix and the gradient vector involve these quantities. They are usually also unknown. As in the pure finite order VAR case, it can be shown that the sample mean

$$\overline{y} = \frac{1}{T} \sum_{t=1}^{T} y_t$$

is an estimator for μ which has the same asymptotic properties as the ML estimator. Therefore, ML estimation of $\boldsymbol{\gamma}$ and Σ_u is often done conditionally on $\mu = \overline{y}$. In other words, the sample mean is subtracted from the data before the VARMA coefficients are estimated.

There are different ways to handle the unknown Σ_u matrix. From (12.3.9), we know that

$$\widetilde{\Sigma}_u(\mu, \boldsymbol{\gamma}) = \frac{1}{T} \sum_{t=1}^{T} u_t(\mu, \boldsymbol{\gamma}) u_t(\mu, \boldsymbol{\gamma})'.$$

Therefore, one possibility is to use $\Sigma_i := \widetilde{\Sigma}_u(\overline{y}, \boldsymbol{\gamma}_i)$ in the i-th iteration. Equivalently, the minimization algorithm can be applied to $\ln |\widetilde{\Sigma}_u(\overline{y}, \boldsymbol{\gamma})|$.

A number of computer program packages contain exact or approximate ML algorithms which may be used in practice. The foregoing algorithm is just meant to demonstrate some basic principles. Modifications in actual applications may result in improved convergence properties. Slow convergence or no convergence at all may be the consequence of working with VARMA orders or Kronecker indices which are larger than the true ones and, hence, with an overparameterized model.

12.3.3 The Information Matrix

In the scoring algorithm described previously, an estimate of the information matrix is needed. To see how that can be obtained, we consider the second order partial derivatives of $-\ln l_0$,

$$\begin{aligned}\frac{\partial^2(-\ln l_0)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} &= \partial\left[\sum_{t=1}^{T}\frac{\partial u_t'}{\partial\boldsymbol{\gamma}}\Sigma_u^{-1}u_t\right]\Big/\,\partial\boldsymbol{\gamma}' \qquad \text{(see (12.3.2))}\\ &= \sum_{t=1}^{T}\frac{\partial u_t'}{\partial\boldsymbol{\gamma}}\Sigma_u^{-1}\frac{\partial u_t}{\partial\boldsymbol{\gamma}'} + (u_t'\Sigma_u^{-1}\otimes I)\frac{\partial\,\mathrm{vec}[\partial u_t'/\partial\boldsymbol{\gamma}]}{\partial\boldsymbol{\gamma}'}.\end{aligned}$$

Taking the expectation of this expression, the last term vanishes because $E(u_t)=0$ and $u_t'\Sigma_u^{-1}\otimes I$ is independent of

$$\frac{\partial\,\mathrm{vec}[\partial u_t'/\partial\boldsymbol{\gamma}]}{\partial\boldsymbol{\gamma}'}$$

as this term does not contain current y_t or u_t variables (see Lemma 12.1). Hence,

$$E\left[\frac{\partial^2(-\ln l_0)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right] = \sum_{t=1}^{T}E\left[\frac{\partial u_t'}{\partial\boldsymbol{\gamma}}\Sigma_u^{-1}\frac{\partial u_t}{\partial\boldsymbol{\gamma}'}\right].$$

Estimating the expected value in the usual way by the sample average gives an estimator

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial u_t'}{\partial\boldsymbol{\gamma}}\Sigma_u^{-1}\frac{\partial u_t}{\partial\boldsymbol{\gamma}'}$$

for

$$E\left[\frac{\partial u_t'}{\partial\boldsymbol{\gamma}}\Sigma_u^{-1}\frac{\partial u_t}{\partial\boldsymbol{\gamma}'}\right].$$

These considerations suggest the estimator

$$\widehat{\mathcal{I}}(\boldsymbol{\gamma}) = \sum_{t=1}^{T}\frac{\partial u_t(\overline{y},\boldsymbol{\gamma})'}{\partial\boldsymbol{\gamma}}\Sigma_u^{-1}\frac{\partial u_t(\overline{y},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}'} \tag{12.3.17}$$

for the information matrix $\mathcal{I}(\boldsymbol{\gamma})$. In the i-th iteration of the scoring algorithm, we evaluate this estimator for $\boldsymbol{\gamma}=\boldsymbol{\gamma}_i$. The quantities $\partial u_t/\partial\boldsymbol{\gamma}'$ may be obtained recursively as in Lemma 12.1 to make this estimator operational.

If $\boldsymbol{\gamma}$ is the true parameter value, the asymptotic information matrix equals plim $\widehat{\mathcal{I}}(\boldsymbol{\gamma})/T$. Thus, if we have a consistent estimator $\widetilde{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$, $\widehat{\mathcal{I}}(\widetilde{\boldsymbol{\gamma}})/T$ is a consistent estimator of the asymptotic information matrix, that is,

$$\mathcal{I}_a(\boldsymbol{\gamma}) = \text{plim}\,\widehat{\mathcal{I}}(\widetilde{\boldsymbol{\gamma}})/T. \tag{12.3.18}$$

In Section 12.4, we will see that the inverse of this matrix, if it exists, is the asymptotic covariance matrix of the ML estimator for $\boldsymbol{\gamma}$. If a nonidentified structure is used, this problem is reflected in the asymptotic information matrix being singular. Hence, it is important at this stage to have an identified version of a VARMA model.

12.3.4 Preliminary Estimation

The coefficients of a VARMA(p, q) model in standard form,

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t + M_1 u_{t-1} + \cdots + M_q u_{t-q},$$

could be estimated by multivariate LS, if the lagged u_t were given. We assume that the sample mean $\overline{y}$ has been subtracted previously. It is therefore neglected here. In deriving preliminary estimators for the other parameters, the idea is to fit a long pure autoregression first and then use estimated residuals in place of the true residuals. Hence, we fit a VAR(n) model

$$y_t = \sum_{i=1}^{n} \Pi_i(n) y_{t-i} + u_t(n),$$

where n is larger than p and q. From that estimation, we compute estimated residuals

$$\widehat{u}_t(n) := y_t - \sum_{i=1}^{n} \widehat{\Pi}_i(n) y_{t-i}, \tag{12.3.19}$$

where $\widehat{\Pi}_i(n)$ are the multivariate LS estimators. Then we set up a multivariate regression model

$$Y = [A : M] X_n + U^0, \tag{12.3.20}$$

where $Y := [y_1, \ldots, y_T]$, $A := [A_1, \ldots, A_p]$, $M := [M_1, \ldots, M_q]$,

$$X_n := [Y_{0,n}, \ldots, Y_{T-1,n}] \quad \text{with } Y_{t,n} := \begin{bmatrix} y_t \\ \vdots \\ y_{t-p+1} \\ \widehat{u}_t(n) \\ \vdots \\ \widehat{u}_{t-q+1}(n) \end{bmatrix} \quad (K(p+q) \times 1)$$

and U^0 is a $(K \times T)$ matrix of residuals. Usually restrictions will be imposed on the parameters A and M of the model, for instance, if the model is given in final equations form. Additional restrictions may also be available. Suppose the restrictions are such that there exists a matrix R and a vector $\boldsymbol{\gamma}$ satisfying

$$\text{vec}[A : M] = R\boldsymbol{\gamma}. \tag{12.3.21}$$

Applying the vec operator to (12.3.20) and substituting $R\boldsymbol{\gamma}$ for $\text{vec}[A : M]$ gives

$$\text{vec}(Y) = (X_n' \otimes I_K)R\boldsymbol{\gamma} + \text{vec}(U^0) \tag{12.3.22}$$

and the LS estimator of $\boldsymbol{\gamma}$ is known to be

$$\widehat{\boldsymbol{\gamma}}(n) = [R'(X_n X_n' \otimes I_K)R]^{-1}R'(X_n \otimes I_K)\ \text{vec}(Y) \tag{12.3.23}$$

(see Chapter 5, Section 5.2). This estimator may be used as an initial vector $\boldsymbol{\gamma}_1$ in the ML algorithm described in the previous subsections.

Using this estimator, a new set of residuals may be obtained as

$$\text{vec}(\widehat{U}^0) = \text{vec}(Y) - (X_n' \otimes I_K)R\widehat{\boldsymbol{\gamma}}(n)$$

which may be used to obtain a white noise covariance estimator

$$\widetilde{\Sigma}_u(n) = \widehat{U}^0\widehat{U}^{0\prime}/T. \tag{12.3.24}$$

This estimator may be used in place of Σ_u in the initial round of the iterative optimization algorithm described earlier.

Alternatively, instead of the LS estimator (12.3.23), we may use an EGLS estimator,

$$\widehat{\widehat{\boldsymbol{\gamma}}}(n) = [R'(X_n X_n' \otimes \widetilde{\Sigma}_u)R]^{-1}R'(X_n \otimes \widetilde{\Sigma}_u)\,\text{vec}(Y),$$

with $\widetilde{\Sigma}_u(n)$ in place of $\widetilde{\Sigma}_u$ or a white noise covariance matrix estimator based on the residuals $\widehat{u}_t(n)$.

The echelon form of a VARMA(p, q) process may be of the more general type

$$A_0 y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + A_0 u_t + M_1 u_{t-1} + \cdots + M_q u_{t-q}, \tag{12.3.25}$$

where A_0 is a lower triangular matrix with unit diagonal. To handle this case, we proceed in a similar manner as in the standard case and substitute the residuals $\widehat{u}_t(n)$ for the lagged u_t and for current residuals from other equations. In other words, in the k-th equation we substitute estimation residuals for $u_{it}, i < k$. Because A_0 is the coefficient matrix for both y_t and u_t, we define

$$X_n^c := [Y_{0,n}^c, \ldots, Y_{T-1,n}^c], \quad \text{where} \quad Y_{t,n}^c := \begin{bmatrix} y_{t+1} - \widehat{u}_{t+1}(n) \\ Y_{t,n} \end{bmatrix}$$

and we pick a restriction matrix R_c and a vector $\boldsymbol{\gamma}_c$ such that

$$R_c\boldsymbol{\gamma}_c = \text{vec}[I_K - A_0, A, M].$$

Hence,

$$\mathrm{vec}(Y) = (X_n^{c\prime} \otimes I_K) R_c \boldsymbol{\gamma}_c + \mathrm{vec}(U^0)$$

and the LS estimator of $\boldsymbol{\gamma}_c$ becomes

$$\widehat{\boldsymbol{\gamma}}_c(n) = [R_c'(X_n^c X_n^{c\prime} \otimes I_K) R_c]^{-1} R_c'(X_n^c \otimes I_K)\mathrm{vec}(Y).$$

The starting-up estimator of Σ_u is then obtained from the residuals of this regression. It is possible that the VARMA process corresponding to these coefficients is unstable or noninvertible. Especially in the latter case, modifications are desirable (see Hannan & Kavalieris (1984), Hannan & Deistler (1988)).

To see more clearly what is being done in this preliminary estimation procedure, let us look at an example. Suppose the bivariate VARMA$(1,1)$ echelon form model from (12.1.19) with Kronecker indices $(p_1, p_2) = (1, 0)$ is to be estimated:

$$\begin{aligned} y_{1,t} &= \alpha_{11,1} y_{1,t-1} + u_{1,t} + m_{11,1} u_{1,t-1} + m_{12,1} u_{2,t-1}, \\ y_{2,t} &= \alpha_{21,0} y_{1,t} - \alpha_{21,0} u_{1,t} + u_{2,t} = \alpha_{21,0}(y_{1,t} - u_{1,t}) + u_{2,t}. \end{aligned} \tag{12.3.26}$$

We assume that the sample mean has been removed previously. The parameters in the first equation are estimated by applying LS to

$$\begin{bmatrix} y_{1,1} \\ \vdots \\ y_{1,T} \end{bmatrix} = \begin{bmatrix} y_{1,0} & \widehat{u}_{1,0}(n) & \widehat{u}_{2,0}(n) \\ \vdots & \vdots & \vdots \\ y_{1,T-1} & \widehat{u}_{1,T-1}(n) & \widehat{u}_{2,T-1}(n) \end{bmatrix} \begin{bmatrix} \alpha_{11,1} \\ m_{11,1} \\ m_{12,1} \end{bmatrix} + \begin{bmatrix} u_{1,1} \\ \vdots \\ u_{1,T} \end{bmatrix},$$

or, using obvious notation, to

$$y_{(1)} = X_{(1)} \gamma_1 + u_{(1)}.$$

Here the $\widehat{u}_{i,t}(n)$ are the residuals from the estimated long VAR model of order n. The LS estimator of γ_1 is $\widehat{\gamma}_1 = (X_{(1)}' X_{(1)})^{-1} X_{(1)}' y_{(1)}$.

Similarly, $\alpha_{21,0}$ is estimated by applying LS to

$$\begin{bmatrix} y_{2,1} \\ \vdots \\ y_{2,T} \end{bmatrix} = \begin{bmatrix} y_{1,1} - \widehat{u}_{1,1}(n) \\ \vdots \\ y_{1,T} - \widehat{u}_{1,T}(n) \end{bmatrix} \alpha_{21,0} + \begin{bmatrix} u_{2,1} \\ \vdots \\ u_{2,T} \end{bmatrix}.$$

In this case, it would be possible to use the residuals of the first regression instead of the $\widehat{u}_{1,t}(n)$ which are the residuals from the long VAR. However, we have chosen to use the latter in the preliminary estimation procedure.

In the foregoing, we have so far ignored the problem of choosing presample values for the estimation. Two alternative choices are reasonable. Either all presample values are replaced by zero or some y_t values at the beginning of the sample are set aside as presample values and the presample values for the residuals are replaced by zero.

The initial estimators obtained in the foregoing procedure can be shown to be consistent under general conditions if n goes to infinity with the sample size

(see Hannan & Kavalieris (1984), Hannan & Deistler (1988), Poskitt (1992)). We will discuss the situation, where VAR processes of increasing order are fitted to a potentially infinite order process, in Chapter 15 and therefore we do not give details here.

12.3.5 An Illustration

We illustrate the estimation procedure using the income (y_1) and consumption (y_2) data from File E1. As in previous chapters, we use first differences of logarithms of the data from 1960 to 1978. In this case, we subtract the sample mean at an initial stage and denote the mean-adjusted income and consumption variables by y_{1t} and y_{2t}, respectively. We assume a VARMA(2, 2) model in echelon form with Kronecker indices $\mathbf{p} = (p_1, p_2) = (0, 2)$ [$\text{ARMA}_E(0, 2)$],

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \alpha_{22,1} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \alpha_{22,2} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ m_{21,1} & m_{22,1} \end{bmatrix} \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ m_{21,2} & m_{22,2} \end{bmatrix} \begin{bmatrix} u_{1,t-2} \\ u_{2,t-2} \end{bmatrix}. \tag{12.3.27}$$

In the next chapter, it will become apparent why this model is chosen. It implies that the first variable (income) is white noise ($y_{1t} = u_{1t}$). Given the subset VAR models of Chapter 5 (Table 5.1), this specification does not appear to be totally unreasonable. The second equation in (12.3.27) describes consumption as a function of lagged consumption, lagged income ($u_{1,t-i} = y_{1,t-i}$), and a moving average term involving lagged residuals $u_{2,t}$.

Eventually we use a sample from 1960.2 ($t = 1$) to 1978.4 ($t = 75$), that is, $T = 75$. In the preliminary estimation of the model (12.3.27), we estimate a VAR(8) model first, using 8 presample values. Then, using two more presample values, we run a regression of y_{2t} on its own lags and lagged $\widehat{u}_{it}(8)$. More precisely, the regression model is

$$\begin{bmatrix} y_{2,11} \\ \vdots \\ y_{2,T} \end{bmatrix} = \begin{bmatrix} y_{2,10} & y_{2,9} & \widehat{u}_{1,10}(8) & \widehat{u}_{2,10}(8) & \widehat{u}_{1,9}(8) & \widehat{u}_{2,9}(8) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{2,T-1} & y_{2,T-2} & \widehat{u}_{1,T-1}(8) & \widehat{u}_{2,T-1}(8) & \widehat{u}_{1,T-2}(8) & \widehat{u}_{2,T-2}(8) \end{bmatrix} \boldsymbol{\gamma} + \begin{bmatrix} u_{2,11} \\ \vdots \\ u_{2,T} \end{bmatrix},$$

where $\boldsymbol{\gamma} := (\alpha_{22,1}, \alpha_{22,2}, m_{21,1}, m_{22,1}, m_{21,2}, m_{22,2})'$. In this particular case, we could have substituted y_{1t} for $\widehat{u}_{1t}(8)$ because the model implies $y_{1t} = u_{1t}$.

We have not done so, however, but we have used the residuals from the long autoregression. The resulting preliminary parameter estimates

$$\widetilde{\boldsymbol{\gamma}}_1 = (\widetilde{\alpha}_{22,1}(1), \ldots, \widetilde{m}_{22,2}(1))'$$

are given in Table 12.1.

Table 12.1. Iterative estimates of the income/consumption system

	$\widetilde{\boldsymbol{\gamma}}_i$						
i	$\widetilde{\alpha}_{22,1}$	$\widetilde{\alpha}_{22,2}$	$\widetilde{m}_{21,1}$	$\widetilde{m}_{22,1}$	$\widetilde{m}_{21,2}$	$\widetilde{m}_{22,2}$	$\lvert\widetilde{\Sigma}_u(\widetilde{\boldsymbol{\gamma}}_i)\rvert \times 10^8$
1	0.020	0.395	0.296	−0.367	0.181	−0.224	0.872564
2	−0.178	0.492	0.331	−0.527	0.175	−0.015	0.942791
3	0.072	0.117	0.305	−0.589	0.191	0.065	0.779788
4	0.202	0.078	0.311	−0.731	0.146	0.147	0.776107
5	0.219	0.063	0.312	−0.744	0.142	0.158	0.775959
6	0.224	0.062	0.313	−0.748	0.140	0.159	0.775952
⋮							
10	0.225	0.061	0.313	−0.750	0.140	0.160	0.775951

We use these estimates to start the scoring algorithm. For our particular example, the i-th iteration proceeds as follows:

(1) Compute residuals

$$\widetilde{u}_t(i) = y_t - \widetilde{A}_1(i)y_{t-1} - \widetilde{A}_2(i)y_{t-2} - \widetilde{M}_1(i)\widetilde{u}_{t-1}(i) - \widetilde{M}_2(i)\widetilde{u}_{t-2}(i)$$

recursively, for $t = 1, 2, \ldots, T$, with $\widetilde{u}_{-1}(i) = \widetilde{u}_0(i) = y_{-1} = y_0 = 0$ and

$$\widetilde{A}_1(i) = \begin{bmatrix} 0 & 0 \\ 0 & \widetilde{\alpha}_{22,1}(i) \end{bmatrix}, \quad \widetilde{A}_2(i) = \begin{bmatrix} 0 & 0 \\ 0 & \widetilde{\alpha}_{22,2}(i) \end{bmatrix},$$

$$\widetilde{M}_1(i) = \begin{bmatrix} 0 & 0 \\ \widetilde{m}_{21,1}(i) & \widetilde{m}_{22,1}(i) \end{bmatrix}, \quad \widetilde{M}_2(i) = \begin{bmatrix} 0 & 0 \\ \widetilde{m}_{21,2}(i) & \widetilde{m}_{22,2}(i) \end{bmatrix}.$$

(2) Compute the partial derivatives $\widetilde{\partial} u_t / \partial \boldsymbol{\gamma}$ recursively as

$$\begin{aligned} \frac{\widetilde{\partial} u_t}{\partial \boldsymbol{\gamma}'}(i) \;=\; & -\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ y_{2,t-1} & y_{2,t-2} & \widetilde{u}_{2,t-1}(i) & \widetilde{u}_{2,t-1}(i) & \widetilde{u}_{1,t-2}(i) & \widetilde{u}_{2,t-2}(i) \end{bmatrix} \\ & -\widetilde{M}_1(i)\frac{\widetilde{\partial} u_{t-1}}{\partial \boldsymbol{\gamma}'}(i) - \widetilde{M}_2(i)\frac{\widetilde{\partial} u_{t-2}}{\partial \boldsymbol{\gamma}'}(i) \end{aligned}$$

for $t = 1, 2, \ldots, T$, with

$$\frac{\widetilde{\partial} u_{-1}}{\partial \boldsymbol{\gamma}'}(i) = \frac{\widetilde{\partial} u_0}{\partial \boldsymbol{\gamma}'}(i) = 0.$$

(3) Compute

$$\begin{aligned}\widetilde{\Sigma}_u(\widetilde{\boldsymbol{\gamma}}_i) &= \frac{1}{T}\sum_{t=1}^{T}\widetilde{u}_t(i)\widetilde{u}_t(i)',\\ \widehat{\mathcal{I}}(\widetilde{\boldsymbol{\gamma}}_i) &= \sum_{t=1}^{T}\frac{\widetilde{\partial}u_t'}{\partial\boldsymbol{\gamma}}(i)\widetilde{\Sigma}_u(\widetilde{\boldsymbol{\gamma}}_i)^{-1}\frac{\widetilde{\partial}u_t}{\partial\boldsymbol{\gamma}'}(i),\end{aligned}$$

and

$$\frac{\partial(-\ln l_0)}{\partial\boldsymbol{\gamma}'}\Big|_{\widetilde{\boldsymbol{\gamma}}_i} = \sum_{t=1}^{T}\widetilde{u}_t(i)'\widetilde{\Sigma}_u(\widetilde{\boldsymbol{\gamma}}_i)^{-1}\frac{\widetilde{\partial}u_t}{\partial\boldsymbol{\gamma}'}(i).$$

(4) Perform the iteration step

$$\widetilde{\boldsymbol{\gamma}}_{i+1} = \widetilde{\boldsymbol{\gamma}}_i - \widehat{\mathcal{I}}(\widetilde{\boldsymbol{\gamma}}_i)^{-1}\left[\frac{\partial(-\ln l_0)}{\partial\boldsymbol{\gamma}}\Big|_{\widetilde{\boldsymbol{\gamma}}_i}\right].$$

Some estimates obtained in these iterations are also given in Table 12.1 together with $|\widetilde{\Sigma}_u(\widetilde{\boldsymbol{\gamma}}_i)|$. After a few iterations the latter quantity approximately reaches its minimum and, thus, $-\ln l_0$ obtains its minimum. After the tenth iteration there is not much change in the $\widetilde{\boldsymbol{\gamma}}_i$ and $|\widetilde{\Sigma}_u(\widetilde{\boldsymbol{\gamma}}_i)|$ in further steps. We work with $\widetilde{\boldsymbol{\gamma}}_{10}$ in the following.

The determinantal polynomial of the MA operator for $i = 10$ is

$$\begin{aligned}|I_2 + \widetilde{M}_1(10)z + \widetilde{M}_2(10)z^2| &= 1 + \widetilde{m}_{22,1}(10)z + \widetilde{m}_{22,2}(10)z^2\\ &= 1 - .750z + .160z^2\end{aligned}$$

which has roots that are clearly outside the unit circle. Thus, the estimated MA operator is invertible. Also, the determinant of the estimated VAR polynomial,

$$\begin{aligned}|I_2 - \widetilde{A}_1(10)z - \widetilde{A}_2(10)z^2| &= 1 - \widetilde{\alpha}_{22,1}(10)z - \widetilde{\alpha}_{22,2}(10)z^2\\ &= 1 - .225z - .061z^2,\end{aligned}$$

is easily seen to have its roots outside the unit circle. Hence, the estimated VARMA process is stable and invertible.

Generally, computing the ML estimates is not always easy. Therefore, other estimation methods were also proposed in the literature (e.g., Koreisha & Pukkila (1987), van Overschee & DeMoor (1994), Kapetanios (2003)).

12.4 Asymptotic Properties of the ML Estimators

12.4.1 Theoretical Results

In this section, the asymptotic properties of the ML estimators are given. We will not prove the main result but refer the reader to Hannan (1979), Dunsmuir & Hannan (1976), Hannan & Deistler (1988), and Kohn (1979) for further discussions and proofs.

Proposition 12.1 (*Asymptotic Properties of ML Estimators*)
Let y_t be a K-dimensional, stationary Gaussian process with stable and invertible VARMA(p,q) representation

$$A_0(y_t-\mu) = A_1(y_{t-1}-\mu)+\cdots+A_p(y_{t-p}-\mu)+A_0u_t+M_1u_{t-1}+\cdots+M_qu_{t-q}, \tag{12.4.1}$$

where u_t is Gaussian white noise with nonsingular covariance matrix Σ_u. Suppose the VAR and MA operators are left-coprime and either in final equations form or in echelon form with possibly linear restrictions on the coefficients so that the coefficient matrices $A_0, A_1, \ldots, A_p, M_1, \ldots, M_q$ depend on a set of unrestricted parameters $\boldsymbol{\gamma}$ as in (12.2.21). Let $\widetilde{\mu}, \widetilde{\boldsymbol{\gamma}}$, and $\widetilde{\Sigma}_u$ be the ML estimators of $\mu, \boldsymbol{\gamma}$, and Σ_u, respectively, and denote vech(Σ_u) and vech$(\widetilde{\Sigma}_u)$ by $\boldsymbol{\sigma}$ and $\widetilde{\boldsymbol{\sigma}}$, respectively. Then all three ML estimators are consistent and asymptotically normally distributed,

$$\sqrt{T}\begin{bmatrix} \widetilde{\mu}-\mu \\ \widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma} \\ \widetilde{\boldsymbol{\sigma}}-\boldsymbol{\sigma} \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} \Sigma_{\widetilde{\mu}} & 0 & 0 \\ 0 & \Sigma_{\widetilde{\boldsymbol{\gamma}}} & 0 \\ 0 & 0 & \Sigma_{\widetilde{\boldsymbol{\sigma}}} \end{bmatrix}\right), \tag{12.4.2}$$

where

$$\begin{aligned} \Sigma_{\widetilde{\mu}} &= A(1)^{-1}M(1)\Sigma_u M(1)'A(1)'^{-1}, \\ \Sigma_{\widetilde{\boldsymbol{\gamma}}} &= \mathcal{I}_a(\boldsymbol{\gamma})^{-1} = \operatorname{plim}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial u_t'}{\partial \boldsymbol{\gamma}}\Sigma_u^{-1}\frac{\partial u_t}{\partial \boldsymbol{\gamma}'}\right]^{-1} \end{aligned}$$

with $\partial u_t/\partial\boldsymbol{\gamma}'$ as given in Lemma 12.1, and

$$\Sigma_{\widetilde{\boldsymbol{\sigma}}} = 2\mathbf{D}_K^{+}(\Sigma_u \otimes \Sigma_u)\mathbf{D}_K^{+\prime}$$

with $\mathbf{D}_K^{+} = (\mathbf{D}_K'\mathbf{D}_K)^{-1}\mathbf{D}_K'$ and $\mathbf{D}_K$ is the $(K^2 \times \frac{1}{2}K(K+1))$ duplication matrix. The covariance matrix in (12.4.2) is consistently estimated by replacing the unknown quantities by their ML estimators. ■

Some remarks on this proposition may be worthwhile.

Remark 1 The results of the proposition do not change if the ML estimator $\widetilde{\mu}$ is replaced by the sample mean $\overline{y}$ and $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\sigma}}$ are ML estimators conditional on $\overline{y}$, that is, $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\sigma}}$ are obtained by replacing μ by $\overline{y}$ in the ML algorithm. One consequence of this result is that asymptotically the sample mean is a fully efficient estimator of μ. ■

Remark 2 The proposition is formulated for final equations or echelon form VARMA models. Its statement remains true for other uniquely identified structures. ■

Remark 3 Because the covariance matrix of the asymptotic distribution in (12.4.2) is block-diagonal, the estimators of $\mu, \boldsymbol{\gamma}$, and Σ_u are asymptotically independent. ■

Remark 4 Much of the proposition remains valid even if y_t is not normally distributed. In that case the estimators obtained by maximizing the Gaussian likelihood function are quasi ML estimators. If u_t is independent standard white noise (see Chapter 3, Definition 3.1), $\widetilde{\boldsymbol{\gamma}}$ and $\overline{y}$ maintain their asymptotic properties. The covariance matrix of $\widetilde{\boldsymbol{\sigma}}$ may be different from the one given in Proposition 12.1. ■

Remark 5 The results of the proposition remain valid under general conditions if instead of the ML estimator $\widetilde{\boldsymbol{\gamma}}$ an estimator is used which is obtained from one iteration of the scoring algorithm outlined in Section 12.3.2, starting from the preliminary estimator of Section 12.3.4. Thus, one possible approach to estimating the parameters of a VARMA model is to compute the sample mean $\overline{y}$ first and use that as an estimator of μ. Then the preliminary estimator for $\boldsymbol{\gamma}$ may be computed as described in Section 12.3.4 and that estimator is used as the initial vector in the optimization algorithm of Section 12.3.2. Then just one step of the form (12.3.16) is performed with $s_i = s_1 = 1$. The resulting estimators $\widetilde{\boldsymbol{\gamma}}_2$ and $\widetilde{\Sigma}_u(\overline{y}, \widetilde{\boldsymbol{\gamma}}_2)$ may then be used instead of $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\Sigma}_u$ in Proposition 12.1. Under general conditions, they have the same *asymptotic* distributions as the actual ML estimators. Of course, this possibility is a computationally attractive way to estimate the coefficients of a VARMA model. In general, the small sample properties of the resulting estimators are not the same as those of the ML estimators, however. ■

Remark 6 Because often the final equations form involves more parameters than the echelon form, unrestricted estimation of the former may result in inefficient estimators. Intuitively, if we start from the echelon form and determine the corresponding final equations form, the coefficients of the latter are seen to satisfy restrictions that could be imposed to obtain more efficient estimators. ■

In the following sections, we will occasionally be interested in the asymptotic distribution of the coefficients of the standard representation of the process,

$$(y_t - \mu) = A_1(y_{t-1} - \mu) + \cdots + A_p(y_{t-p} - \mu) + u_t + M_1 u_{t-1} + \cdots + M_q u_{t-q}. \tag{12.4.3}$$

The coefficients are functions of $\boldsymbol{\gamma}$ and their asymptotic distributions follow in the usual way. Let

$$\boldsymbol{\alpha} := \operatorname{vec}[A_1, \ldots, A_p] \quad \text{and} \quad \mathbf{m} := \operatorname{vec}[M_1, \ldots, M_q],$$

then

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}(\boldsymbol{\gamma}) \\ \mathbf{m}(\boldsymbol{\gamma}) \end{bmatrix}.$$

The ML estimators are

$$\begin{bmatrix} \widetilde{\boldsymbol{\alpha}} \\ \widetilde{\mathbf{m}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}(\widetilde{\boldsymbol{\gamma}}) \\ \mathbf{m}(\widetilde{\boldsymbol{\gamma}}) \end{bmatrix}.$$

They are consistent and asymptotically normal,

$$\sqrt{T}\left(\begin{bmatrix} \widetilde{\boldsymbol{\alpha}} \\ \widetilde{\mathbf{m}} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{m} \end{bmatrix}\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{\left[\begin{smallmatrix} \widetilde{\boldsymbol{\alpha}} \\ \widetilde{\mathbf{m}} \end{smallmatrix}\right]} = \begin{bmatrix} \dfrac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\gamma}'} \\ \dfrac{\partial \mathbf{m}}{\partial \boldsymbol{\gamma}'} \end{bmatrix} \Sigma_{\widetilde{\boldsymbol{\gamma}}} \left[\frac{\partial \boldsymbol{\alpha}'}{\partial \boldsymbol{\gamma}}, \frac{\partial \mathbf{m}'}{\partial \boldsymbol{\gamma}}\right]\right). \tag{12.4.4}$$

If $A_0 = I_K$, $\boldsymbol{\alpha}$ and $\mathbf{m}$ will often be linearly related to $\boldsymbol{\gamma}$ and we get the following corollary of Proposition 12.1.

Corollary 12.1.1
Under the conditions of Proposition 12.1, if

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{m} \end{bmatrix} = R\boldsymbol{\gamma} + r,$$

$$\sqrt{T}\left(\begin{bmatrix} \widetilde{\boldsymbol{\alpha}} \\ \widetilde{\mathbf{m}} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{m} \end{bmatrix}\right) \xrightarrow{d} \mathcal{N}(0, R\Sigma_{\widetilde{\boldsymbol{\gamma}}}R')$$

and $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\mathbf{m}}$ are asymptotically independent of $\overline{y}$, $\widetilde{\mu}$, and $\widetilde{\boldsymbol{\sigma}}$. ■

The remarks following the proposition also apply for the corollary. For illustrative purposes, consider the bivariate $\text{ARMA}_E(0,1)$ model,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 - \alpha_{22,1}L \end{bmatrix} y_t = \begin{bmatrix} 1 & 0 \\ m_{21,1}L & 1 + m_{22,1}L \end{bmatrix} u_t \tag{12.4.5}$$

or

$$y_t = \begin{bmatrix} 0 & 0 \\ 0 & \alpha_{22,1} \end{bmatrix} y_{t-1} + u_t + \begin{bmatrix} 0 & 0 \\ m_{21,1} & m_{22,1} \end{bmatrix} u_{t-1}.$$

In this case,

$$\boldsymbol{\alpha} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \alpha_{22,1} \end{bmatrix}, \quad \mathbf{m} = \begin{bmatrix} 0 \\ m_{21,1} \\ 0 \\ m_{22,1} \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{bmatrix} \alpha_{22,1} \\ m_{21,1} \\ m_{22,1} \end{bmatrix},$$

$$R = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \text{and} \quad r = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

If the VARMA model is not in standard form originally, we premultiply by A_0^{-1} to get

$$\begin{aligned} y_t - \mu &= A_0^{-1}A_1(y_{t-1}-\mu) + \cdots + A_0^{-1}A_p(y_{t-p}-\mu) \\ &\quad + u_t + A_0^{-1}M_1u_{t-1} + \cdots + A_0^{-1}M_qu_{t-q}. \end{aligned} \tag{12.4.6}$$

In this case, it is more reasonable to assume that

$$\boldsymbol{\beta}_0 := \operatorname{vec}[A_0, A_1, \ldots, A_p, M_1, \ldots, M_q] \tag{12.4.7}$$

is linearly related to $\boldsymbol{\gamma}$, say,

$$\boldsymbol{\beta}_0 = R\boldsymbol{\gamma} + r. \tag{12.4.8}$$

Then it follows for

$$\boldsymbol{\alpha} := \operatorname{vec}[A_0^{-1}A_1, \ldots, A_0^{-1}A_p] = \operatorname{vec}(A_0^{-1}[A_1, \ldots, A_p]) \tag{12.4.9}$$

and

$$\mathbf{m} := \operatorname{vec}(A_0^{-1}[M_1, \ldots, M_q]), \tag{12.4.10}$$

that

$$\begin{bmatrix} \dfrac{\partial\boldsymbol{\alpha}}{\partial\boldsymbol{\gamma}'} \\ \dfrac{\partial\mathbf{m}}{\partial\boldsymbol{\gamma}'} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial\boldsymbol{\alpha}}{\partial\boldsymbol{\beta}_0'} \\ \dfrac{\partial\mathbf{m}}{\partial\boldsymbol{\beta}_0'} \end{bmatrix} \frac{\partial\boldsymbol{\beta}_0}{\partial\boldsymbol{\gamma}'} = \begin{bmatrix} \dfrac{\partial\boldsymbol{\alpha}}{\partial\boldsymbol{\beta}_0'} \\ \dfrac{\partial\mathbf{m}}{\partial\boldsymbol{\beta}_0'} \end{bmatrix} R.$$

Hence, we need to evaluate $\partial\boldsymbol{\alpha}/\partial\boldsymbol{\beta}_0'$ and $\partial\mathbf{m}/\partial\boldsymbol{\beta}_0'$ to obtain the asymptotic covariance matrix of the standard form coefficients.

$$\begin{aligned} \frac{\partial\boldsymbol{\alpha}}{\partial\boldsymbol{\beta}_0'} &= (I_{Kp}\otimes A_0^{-1})\frac{\partial\operatorname{vec}[A_1,\ldots,A_p]}{\partial\boldsymbol{\beta}_0'} + \left(\begin{bmatrix} A_1' \\ \vdots \\ A_p' \end{bmatrix}\otimes I_K\right)\frac{\partial\operatorname{vec}(A_0^{-1})}{\partial\boldsymbol{\beta}_0'} \\ &= (I_{Kp}\otimes A_0^{-1})[0 : I_{K^2p} : 0] \\ &\quad - \left(\begin{bmatrix} A_1' \\ \vdots \\ A_p' \end{bmatrix}\otimes I_K\right)((A_0^{-1})'\otimes A_0^{-1})\frac{\partial\operatorname{vec}(A_0)}{\partial\boldsymbol{\beta}_0'} \\ &\qquad\qquad \text{(see Rule 9 of Appendix A.13)} \\ &= (I_{Kp}\otimes A_0^{-1})[0 : I_{K^2p} : 0] \\ &\quad - \left(\begin{bmatrix} (A_0^{-1}A_1)' \\ \vdots \\ (A_0^{-1}A_p)' \end{bmatrix}\otimes A_0^{-1}\right)[I_{K^2} : 0]. \end{aligned} \tag{12.4.11}$$

A similar expression is obtained for $\partial\mathbf{m}/\partial\boldsymbol{\beta}_0'$. This result is summarized in the next corollary.

Corollary 12.1.2
Under the conditions of Proposition 12.1, if $\boldsymbol{\beta}_0$ is as defined in (12.4.7) and satisfies the restrictions in (12.4.8) and $\boldsymbol{\alpha}$ and $\mathbf{m}$ are the coefficients of the standard form VARMA representation defined in (12.4.9) and (12.4.10), respectively, with ML estimators $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\mathbf{m}}$, then

$$\sqrt{T}\left(\begin{bmatrix}\widetilde{\boldsymbol{\alpha}}\\ \widetilde{\mathbf{m}}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\alpha}\\ \mathbf{m}\end{bmatrix}\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{\left[\begin{smallmatrix}\widetilde{\boldsymbol{\alpha}}\\ \widetilde{\mathbf{m}}\end{smallmatrix}\right]} = \begin{bmatrix}H_{\boldsymbol{\alpha}}\\ H_{\mathbf{m}}\end{bmatrix} R\Sigma_{\widetilde{\boldsymbol{\gamma}}}R'[H'_{\boldsymbol{\alpha}} : H'_{\mathbf{m}}]\right),$$

where

$$\begin{aligned} H_{\boldsymbol{\alpha}} &:= \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\beta}_0'} \qquad (K^2p \times K^2(p+q+1)) \\ &= (I_{Kp} \otimes A_0^{-1})[\underbrace{0}_{(K^2p\times K^2)} : I_{K^2p} : \underbrace{0}_{(K^2p\times K^2q)}] \\ &\quad - \left(\begin{bmatrix}(A_0^{-1}A_1)'\\ \vdots \\ (A_0^{-1}A_p)'\end{bmatrix} \otimes A_0^{-1}\right)[I_{K^2} : 0] \end{aligned}$$

and

$$\begin{aligned} H_{\mathbf{m}} &:= \frac{\partial \mathbf{m}}{\partial \boldsymbol{\beta}_0'} \qquad (K^2q \times K^2(p+q+1)) \\ &= (I_{Kq} \otimes A_0^{-1})[0 : I_{K^2q}] - \left(\begin{bmatrix}(A_0^{-1}M_1)'\\ \vdots \\ (A_0^{-1}M_q)'\end{bmatrix} \otimes A_0^{-1}\right)[I_{K^2} : 0]. \end{aligned}$$

■

Again an example may be worthwhile. Consider the following bivariate $\text{ARMA}_E(2,1)$ process with some zero restrictions placed on the coefficients (see also Problem 12.3):

$$\begin{bmatrix} 1-\alpha_{11,1}L-\alpha_{11,2}L^2 & 0 \\ -\alpha_{21,0}-\alpha_{21,1}L & 1-\alpha_{22,1}L \end{bmatrix} y_t = \begin{bmatrix} 1 & 0 \\ -\alpha_{21,0} & 1+m_{22,1}L \end{bmatrix} u_t \tag{12.4.12}$$

or

$$\begin{aligned} \begin{bmatrix} 1 & 0 \\ -\alpha_{21,0} & 1 \end{bmatrix} y_t &= \begin{bmatrix} \alpha_{11,1} & 0 \\ \alpha_{21,1} & \alpha_{22,1} \end{bmatrix} y_{t-1} + \begin{bmatrix} \alpha_{11,2} & 0 \\ 0 & 0 \end{bmatrix} y_{t-2} \\ &\quad + \begin{bmatrix} 1 & 0 \\ -\alpha_{21,0} & 1 \end{bmatrix} u_t + \begin{bmatrix} 0 & 0 \\ 0 & m_{22,1} \end{bmatrix} u_{t-1}. \end{aligned}$$

Hence,

$$\boldsymbol{\beta}_0 = \begin{bmatrix} 1 \\ -\alpha_{21,0} \\ 0 \\ 1 \\ \alpha_{11,1} \\ \alpha_{21,1} \\ 0 \\ \alpha_{22,1} \\ \alpha_{11,2} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ m_{22,1} \end{bmatrix}, \quad R = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad r = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \alpha_{21,0} \\ \alpha_{11,1} \\ \alpha_{21,1} \\ \alpha_{22,1} \\ \alpha_{11,2} \\ m_{22,1} \end{bmatrix}.$$

Furthermore,

$$A_0^{-1} = \begin{bmatrix} 1 & 0 \\ -\alpha_{21,0} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ \alpha_{21,0} & 1 \end{bmatrix}.$$

Thus,

$$\begin{aligned} \boldsymbol{\alpha} &= \operatorname{vec}[A_0^{-1}A_1, A_0^{-1}A_2] \\ &= \operatorname{vec}\left[\begin{array}{cc|cc} \alpha_{11,1} & 0 & \alpha_{11,2} & 0 \\ \alpha_{11,1}\alpha_{21,0} + \alpha_{21,1} & \alpha_{22,1} & \alpha_{11,2}\alpha_{21,0} & 0 \end{array}\right] \\ &= \begin{bmatrix} \alpha_{11,1} \\ \alpha_{11,1}\alpha_{21,0} + \alpha_{21,1} \\ 0 \\ \alpha_{22,1} \\ \alpha_{11,2} \\ \alpha_{11,2}\alpha_{21,0} \\ 0 \\ 0 \end{bmatrix} \end{aligned}$$

and

$$\mathbf{m} = \text{vec}[A_0^{-1} M_1] = \text{vec}\begin{bmatrix} 0 & 0 \\ 0 & m_{22,1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ m_{22,1} \end{bmatrix}.$$

Consequently,

$$\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\gamma}'} = H_{\boldsymbol{\alpha}} R = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ \alpha_{11,1} & \alpha_{21,0} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ \alpha_{11,2} & 0 & 0 & 0 & \alpha_{21,0} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{12.4.13}$$

and

$$\frac{\partial \mathbf{m}}{\partial \boldsymbol{\gamma}'} = H_{\mathbf{m}} R = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{12.4.14}$$

(see also Problem 12.7).

12.4.2 A Real Data Example

In their general form, the results may look more complicated than they usually are. Therefore, considering our income/consumption example from Section 12.3.5 again may be helpful. For the VARMA(2, 2) model with Kronecker indices (0, 2) given in (12.3.27), the parameters are

$$\boldsymbol{\gamma} = (\alpha_{22,1}, \alpha_{22,2}, m_{21,1}, m_{22,1}, m_{21,2}, m_{22,2})'.$$

The ML estimates are given in Table 12.1. Using $\widetilde{\boldsymbol{\gamma}} = \widetilde{\boldsymbol{\gamma}}_{10}$ from that table, an estimate of $\mathcal{I}(\boldsymbol{\gamma})$ is obtained from the iterations described in Section 12.3.5, that is, we use $\widehat{\mathcal{I}}(\widetilde{\boldsymbol{\gamma}}_{10}) = \widehat{\mathcal{I}}(\widetilde{\boldsymbol{\gamma}})$. The square roots of the diagonal elements of $\widehat{\mathcal{I}}(\widetilde{\boldsymbol{\gamma}})^{-1}$ are estimates of the standard errors of the elements of $\widetilde{\boldsymbol{\gamma}}$. Giving the estimated standard errors in parentheses, we get

$$\widetilde{\boldsymbol{\gamma}} = \begin{bmatrix} .225 & (.252) \\ .061 & (.166) \\ .313 & (.090) \\ -.750 & (.274) \\ .140 & (.141) \\ .160 & (.233) \end{bmatrix}. \tag{12.4.15}$$

As mentioned in Remark 5 of Section 12.4.1, an alternative, *asymptotically* equivalent estimator is obtained by iterating just once. In the present example that leads to estimates

$$\widetilde{\boldsymbol{\gamma}}_2 = \begin{bmatrix} -.178 & (.165) \\ .492 & (.133) \\ .331 & (.099) \\ -.527 & (.172) \\ .175 & (.127) \\ -.015 & (.152) \end{bmatrix}. \tag{12.4.16}$$

These estimates are somewhat different from those in (12.4.15). However, given the sampling variability reflected in the estimated standard errors, the differences in most of the parameter estimates are not substantial.

Under a two-standard error criterion, only two of the coefficients in (12.4.15) are significantly different from zero. As a consequence, one may wish to restrict some of the coefficients to zero and thereby further reduce the parameter space. We will not do so at this stage but consider the estimates of $\boldsymbol{\alpha}$ and $\mathbf{m}$ implied by $\widetilde{\boldsymbol{\gamma}}$ given in (12.4.15) (see, however, Problem 12.10):

$$\widetilde{\boldsymbol{\alpha}} = \text{vec}[\widetilde{A}_1, \widetilde{A}_2] = \begin{bmatrix} 0 \\ 0 \\ 0 \\ .225(.252) \\ 0 \\ 0 \\ 0 \\ .061(.166) \end{bmatrix},$$

$$\widetilde{\mathbf{m}} = \text{vec}[\widetilde{M}_1, \widetilde{M}_2] = \begin{bmatrix} 0 \\ .313(.090) \\ 0 \\ -.750(.274) \\ 0 \\ .140(.141) \\ 0 \\ .160(.233) \end{bmatrix}. \tag{12.4.17}$$

The standard errors are, of course, not affected by adding a few zero elements. A more elaborate but still simple computation becomes necessary to obtain the standard errors if $A_0 \neq I_K$ (see Corollary 12.1.2).

12.5 Forecasting Estimated VARMA Processes

With respect to forecasting with estimated processes, in principle, the same arguments apply for VARMA models that have been put forward for pure

VAR models. Suppose that the generation process of a multiple time series of interest admits a VARMA(p, q) representation,

$$y_t - \mu = A_1(y_{t-1} - \mu) + \cdots + A_p(y_{t-p} - \mu) + u_t + M_1 u_{t-1} + \cdots + M_q u_{t-q}, \tag{12.5.1}$$

and denote by $\widehat{y}_t(h)$ the h-step ahead forecast (with nonzero mean) at origin t given in Section 11.5, based on estimated rather than known coefficients. For instance, using the pure VAR representation of the process,

$$\widehat{y}_t(h) = \widehat{\mu} + \sum_{i=1}^{h-1} \widehat{\Pi}_i(\widehat{y}_t(h-i) - \widehat{\mu}) + \sum_{i=h}^{\infty} \widehat{\Pi}_i(y_{t+h-i} - \widehat{\mu}). \tag{12.5.2}$$

For practical purposes, one would, of course, truncate the infinite sum. For the moment we will, however, consider the infinite sum. For this predictor, the forecast error is

$$y_{t+h} - \widehat{y}_t(h) = [y_{t+h} - y_t(h)] + [y_t(h) - \widehat{y}_t(h)],$$

where $y_t(h)$ is the optimal forecast based on known coefficients and the two terms on the right-hand side are uncorrelated as the first one can be written in terms of u_s with $s > t$ and the second one contains y_s with $s \leq t$, if the parameter estimators are based on y_s with $s \leq t$ only. Thus, the forecast MSE becomes

$$\begin{aligned} \Sigma_{\widehat{y}}(h) &= \mathrm{MSE}[y_t(h)] + \mathrm{MSE}[y_t(h) - \widehat{y}_t(h)] \\ &= \Sigma_y(h) + E[y_t(h) - \widehat{y}_t(h)][y_t(h) - \widehat{y}_t(h)]'. \end{aligned} \tag{12.5.3}$$

Formally, this is the same expression that was obtained for finite order VAR processes and, using the same arguments as in that case, we approximate the $\mathrm{MSE}[y_t(h) - \widehat{y}_t(h)]$ by $\Omega(h)/T$, where

$$\Omega(h) = E\left[\frac{\partial y_t(h)}{\partial \boldsymbol{\eta}'} \Sigma_{\widetilde{\boldsymbol{\eta}}} \frac{\partial y_t(h)'}{\partial \boldsymbol{\eta}}\right], \tag{12.5.4}$$

$\boldsymbol{\eta}$ is the vector of estimated coefficients, and $\Sigma_{\widetilde{\boldsymbol{\eta}}}$ is its asymptotic covariance matrix. If ML estimation is used and

$$\boldsymbol{\eta} = \begin{bmatrix} \mu \\ \boldsymbol{\alpha} \\ \mathbf{m} \end{bmatrix},$$

where $\boldsymbol{\alpha} = \mathrm{vec}[A_1, \ldots, A_p]$ and $\mathbf{m} = \mathrm{vec}[M_1, \ldots, M_q]$, we have from Proposition 12.1 and Corollaries 12.1.1 and 12.1.2,

$$\Sigma_{\widetilde{\boldsymbol{\eta}}} = \begin{bmatrix} \Sigma_{\widetilde{\mu}} & 0 \\ 0 & \Sigma_{\left[\begin{smallmatrix} \widetilde{\boldsymbol{\alpha}} \\ \widetilde{\mathbf{m}} \end{smallmatrix}\right]} \end{bmatrix}.$$

Thus,

$$\frac{\partial y_t(h)}{\partial \boldsymbol{\eta}'} \Sigma_{\tilde{\boldsymbol{\eta}}} \frac{\partial y_t(h)'}{\partial \boldsymbol{\eta}} = \frac{\partial y_t(h)}{\partial \mu'} \Sigma_{\tilde{\mu}} \frac{\partial y_t(h)'}{\partial \mu} + \frac{\partial y_t(h)}{\partial [\boldsymbol{\alpha}', \mathbf{m}']} \Sigma_{\left[\begin{smallmatrix} \tilde{\boldsymbol{\alpha}} \\ \tilde{\mathbf{m}} \end{smallmatrix}\right]} \frac{\partial y_t(h)'}{\partial \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{m} \end{bmatrix}}.$$

Hence, in order to get an expression for $\Omega(h)$ we need the partial derivatives of $y_t(h)$ with respect to $\mu, \boldsymbol{\alpha}$, and $\mathbf{m}$. They are given in the next lemma.

Lemma 12.2
If y_t is a process with stable and invertible VARMA(p, q) representation (12.5.1) and pure VAR representation

$$y_t = \mu + \sum_{i=1}^{\infty} \Pi_i (y_{t-i} - \mu) + u_t,$$

we have

$$\frac{\partial y_t(h)}{\partial \mu'} = \begin{cases} \left(I_K - \sum_{i=1}^{\infty} \Pi_i \right), & \text{for } h = 1, \\ \left(I_K - \sum_{i=1}^{\infty} \Pi_i \right) + \sum_{i=1}^{h-1} \Pi_i \frac{\partial y_t(h-i)}{\partial \mu'}, & h = 2, 3, \ldots, \end{cases}$$

$$\begin{aligned} \frac{\partial y_t(h)}{\partial [\boldsymbol{\alpha}', \mathbf{m}']} = & \sum_{i=1}^{h-1} [(y_t(h-i) - \mu)' \otimes I_K] \frac{\partial \operatorname{vec}(\Pi_i)}{\partial [\boldsymbol{\alpha}', \mathbf{m}']} + \sum_{i=1}^{h-1} \Pi_i \frac{\partial y_t(h-i)}{\partial [\boldsymbol{\alpha}', \mathbf{m}']} \\ & + \sum_{i=h}^{\infty} [(y_{t+h-i} - \mu)' \otimes I_K] \frac{\partial \operatorname{vec}(\Pi_i)}{\partial [\boldsymbol{\alpha}', \mathbf{m}']}, \quad \text{for } h = 1, 2, \ldots, \end{aligned}$$

with

$$\frac{\partial \operatorname{vec}(\Pi_i)}{\partial [\boldsymbol{\alpha}', \mathbf{m}']} = -\sum_{j=0}^{i-1} [H'(\mathbf{M}')^{i-1-j} \otimes J \mathbf{M}^j] \begin{bmatrix} 0 & I_{Kq} \otimes J' \\ I_{Kp} \otimes J' & 0 \end{bmatrix},$$

where H, $\mathbf{M}$, and J are as defined in Chapter 11, Section 11.3.2, (11.3.13). In other words, H, $\mathbf{M}$, and J are defined so that $-\Pi_i = J \mathbf{M}^i H$. ■

The proof of this lemma is left as an exercise (see Problem 12.8). The formulas given in this lemma can be used for recursively computing the partial derivatives of $y_t(h)$ with respect to the VARMA coefficients for $h = 1, 2, \ldots$.

An estimator of $\Omega(h)$ is obtained by replacing all unknown quantities by their respective estimators and truncating the infinite sum or, equivalently, replacing $y_t - \mu$ by zero for $t \leq 0$. Denoting the resulting estimated partial derivatives by

$$\frac{\widehat{\partial} y_t(h)}{\partial \mu'} \qquad \text{and} \qquad \frac{\widehat{\partial} y_t(h)}{\partial [\boldsymbol{\alpha}', \mathbf{m}']},$$

an estimator for $\Omega(h)$ is

$$\widehat{\Omega}(h) = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\widehat{\partial} y_t(h)}{\partial \mu'}\widehat{\Sigma}_{\widetilde{\mu}}\frac{\widehat{\partial} y_t(h)'}{\partial \mu} + \frac{\widehat{\partial} y_t(h)}{\partial[\boldsymbol{\alpha}', \mathbf{m}']}\widehat{\Sigma}_{\left[\begin{smallmatrix}\tilde{\boldsymbol{\alpha}}\\ \tilde{\mathbf{m}}\end{smallmatrix}\right]}\frac{\widehat{\partial} y_t(h)'}{\partial\left[\begin{smallmatrix}\boldsymbol{\alpha}\\ \mathbf{m}\end{smallmatrix}\right]}\right], \tag{12.5.5}$$

where $\widehat{\Sigma}_{\widetilde{\mu}}$ and

$$\widehat{\Sigma}_{\left[\begin{smallmatrix}\tilde{\boldsymbol{\alpha}}\\ \tilde{\mathbf{m}}\end{smallmatrix}\right]}$$

are estimators of $\Sigma_{\widetilde{\mu}}$ and

$$\Sigma_{\left[\begin{smallmatrix}\tilde{\boldsymbol{\alpha}}\\ \tilde{\mathbf{m}}\end{smallmatrix}\right]},$$

respectively (see Corollaries 12.1.1 and 12.1.2 for the latter matrix). An estimator of the forecast MSE matrix (12.5.3) is then

$$\widehat{\Sigma}_{\widetilde{y}}(h) = \widehat{\Sigma}_y(h) + \frac{1}{T}\widehat{\Omega}(h), \tag{12.5.6}$$

where the estimator $\widehat{\Sigma}_y(h)$ is again obtained by replacing unknown quantities by their respective estimators.

With these results in hand, forecast intervals can be set up, under Gaussian assumptions, just as in the finite order VAR case discussed in Chapters 2 and 3.

12.6 Estimated Impulse Responses

As mentioned in Section 11.7.2, the impulse responses of a VARMA(p, q) process are the coefficients of pure MA representations. For instance, if the process is in standard form, the forecast error impulse responses are

$$\Phi_i = J\mathbf{A}^i H \tag{12.6.1}$$

with J, $\mathbf{A}$, and H as defined in Section 11.3.2 (see (11.3.10)). Other quantities of interest may be the elements of $\Theta_i = \Phi_i P$, where P is a lower triangular Choleski decomposition of Σ_u, the white noise covariance matrix. Also forecast error variance components and accumulated impulse responses may be of interest. All these quantities are estimated in the usual way from the estimated coefficients of the process. For example, $\widehat{\Phi}_i = J\widehat{\mathbf{A}}^i H$, where $\widehat{\mathbf{A}}$ is obtained from $\mathbf{A}$ by replacing the A_i and M_j by estimators $\widehat{A}_i$ and $\widehat{M}_j$. The asymptotic distributions of the estimated quantities follow immediately from Proposition 3.6, which is formulated for the finite order VAR case. The only modifications that we have to make to accommodate the VARMA(p, q) case are to replace $\boldsymbol{\alpha}$ by

$$\boldsymbol{\beta} := \text{vec}[A_1, \ldots, A_p, M_1, \ldots, M_q] = \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{m} \end{bmatrix},$$

replace $\Sigma_{\widetilde{\boldsymbol{\alpha}}}$ by $\Sigma_{\widetilde{\boldsymbol{\beta}}}$ and specify

$$\begin{aligned} G_i &= \frac{\partial \,\text{vec}(\Phi_i)}{\partial \boldsymbol{\beta}'} = (H' \otimes J) \frac{\partial \,\text{vec}(\mathbf{A}^i)}{\partial \boldsymbol{\beta}'} \\ &= (H' \otimes J) \left[\sum_{m=0}^{i-1} (\mathbf{A}')^{i-1-m} \otimes \mathbf{A}^m \right] \frac{\partial \,\text{vec}(\mathbf{A})}{\partial \boldsymbol{\beta}'} \\ &= \sum_{m=0}^{i-1} H'(\mathbf{A}')^{i-1-m} \otimes J\mathbf{A}^m J'. \end{aligned} \tag{12.6.2}$$

With these modifications of Proposition 3.6, the asymptotic distributions of all the quantities of interest are available. Of course, all the caveats of Proposition 3.6 apply here too. In principle, structural impulse responses, as discussed in Chapter 9, may be of interest as well. They are typically not based on VARMA models, however.

12.7 Exercises

Problem 12.1
Are the operators

$$\begin{bmatrix} 1 - 0.5L & 0.3L \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 - 0.2L & 1.3L - 0.44L^2 \\ 0.5L & 1 + 0.2L \end{bmatrix}$$

left-coprime? (Hint: Show that the first operator is a common factor.)

Problem 12.2
Write the bivariate process

$$\begin{bmatrix} 1 - \beta_1 L & 0 \\ \beta_2 L^2 & 1 - \beta_3 L \end{bmatrix} y_t = \begin{bmatrix} 1 - \beta_1 L & 0 \\ \beta_4 L & 1 \end{bmatrix} u_t$$

in final equations form and in echelon form.

Problem 12.3
Show that (12.4.12) is an echelon form representation.

Problem 12.4
Derive the likelihood function for a general Gaussian VARMA(p, q) model given fixed but not necessarily zero initial vectors $y_{-p+1}, \ldots, y_0$. Do not assume that $u_{-q+1} = \cdots = u_0 = 0$!

Problem 12.5
Identify the rules from Appendix A that are used in deriving the partial derivatives in (12.3.8).

Problem 12.6
Suppose that $\ln|\widetilde{\Sigma}_u(\mu,\boldsymbol{\gamma})|$ given in (12.3.11) is to be minimized with respect to $\boldsymbol{\gamma}$. Show that the resulting normal equations are

$$\frac{\partial \ln|\widetilde{\Sigma}_u(\mu,\boldsymbol{\gamma})|}{\partial \boldsymbol{\gamma}'} = \frac{2}{T}\sum_{t=1}^{T} u_t'\Sigma_u^{-1}\frac{\partial u_t}{\partial \boldsymbol{\gamma}'}.$$

Thus, the normal equations are equivalent to those obtained from the log-likelihood function.

Problem 12.7
Consider the bivariate VARMA$(2,1)$ process given in (12.4.12) and set up the matrices $H_{\boldsymbol{\alpha}}$ and $H_{\mathbf{m}}$ according to their general form given in Corollary 12.1.2. Show that $H_{\boldsymbol{\alpha}}R$ and $H_{\mathbf{m}}R$ are identical to the matrices specified in (12.4.13) and (12.4.14), respectively.

Problem 12.8
Prove Lemma 12.2. (Hint: Use Rule (8) of Appendix A.13.)

Problem 12.9
Derive the asymptotic covariance matrices of the impulse responses and forecast error variance components obtained from an estimated VARMA process. (Hint: Use the suggestion given in Section 12.6.)

Problem 12.10
Consider the income/consumption example of Section 12.3.5 and determine preliminary and full ML estimates for the parameters of the model

$$\begin{aligned}\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} &= \begin{bmatrix} 0 & 0 \\ 0 & \alpha_{22,2} \end{bmatrix}\begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} \\ &\quad + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ m_{21,1} & m_{22,1} \end{bmatrix}\begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix} \\ &\quad + \begin{bmatrix} 0 & 0 \\ m_{21,2} & 0 \end{bmatrix}\begin{bmatrix} u_{1,t-2} \\ u_{2,t-2} \end{bmatrix}.\end{aligned}$$