ARMA Estimation Recipes

1. Preliminaries

These notes summarize procedures for estimating the lag coefficients in the stationary ARMA(p,q) model

(1) $y_t = \mu + a_1(y_{t-1}-\mu) + ... + a_p(y_{t-p}-\mu) + \varepsilon_t + b_1\varepsilon_{t-1} + ... + b_q\varepsilon_{t-q}$,

where y_t is observed for t = 1,...,T and the ε_t are unobserved i.i.d. disturbances with mean zero and finite variance σ^2. The mean μ, the lag coefficients $a_1,...,a_p$ and $b_1,...,b_q$, and σ^2 are the parameters of the model. By assumption, $a_p \neq 0$ and $b_q \neq 0$. Define the polynomials $A(z) = a_1z + ... + a_pz^p$ and $B(z) = b_1z + ... + b_qz^q$, and the lag operator L that for any series x_t is defined as $Lx_t = x_{t-1}$. Then the model can be written

(2) $(I - A(L))(y_t - \mu) = (I + B(L))\varepsilon_t$.

The model is *stationary* if and only if the polynomial 1 - A(z) is *stable*; i.e., all its roots lie outside the unit circle. If the model is stationary, then the lag polynomial I - A(L) is invertible, and there is a MA(∞) representation of the model, written formally as

(3) $y_t - \mu = \dfrac{I + B(L)}{I - A(L)} \cdot \varepsilon_t$.

Let $z_1,...,z_p$ denote the roots of 1 - A(z), some of which may be repeated. Then this polynomial can be written $1 - A(z) = (1 - z/z_1)\cdot...\cdot(1 - z/z_p)$. Then $(I - A(L))^{-1} = \prod_{k=1}^{p} \sum_{s=0}^{\infty} z_k^s \cdot L^s$. Alternately, write $(I - A(L))^{-1}(I + B(L)) = \sum_{s=0}^{\infty} \psi_s L^s \equiv \Psi(L)$. The identity

(4) $I + B(L\} = (I - A(L))\cdot \sum_{s=0}^{\infty} \psi_s L^s = \psi_0 I + (\psi_1 - a_1\psi_0)\cdot L + ... + (\psi_s - \sum_{i=1}^{\min(p,s)} a_i\psi_{s-i})\cdot L^s + ...$

implies $\psi_0 = 1$, $\psi_1 = a_1 + b_1$, and $\psi_s = \sum_{i=1}^{\min(p,s)} a_i\psi_{s-i} + b_s$, where $b_s = 0$ for $s > q$. Another derivation of these conditions starts by noting that the covariance of y_t and ε_{t-m} is $\sigma^2\psi_m^2$ for $m \geq 0$, and zero for $m < 0$. Multiplying (1) by ε_{t-m} and taking expectations then gives

(5) $\psi_m = a_1\psi_{m-1} + ... + a_p\psi_{m-p} + b_m$,

with $\psi_{m-s} = 0$ for $s > m$, $b_0 = 1$, and we define $b_m = 0$ for $m > q$. These are called the *Yule-Walker* equations. They can be used recursively to obtain the coefficients ψ_s in the MA(∞) representation.

An implication of the MA(∞) representation is

(6) $\gamma_0 \equiv \mathrm{Var}(y_t) = \sigma^2 \cdot \sum_{s=0}^{\infty} \psi_s^2$ and $\gamma_m \equiv \mathrm{cov}(y_t,y_{t-m}) = \sigma^2 \cdot \sum_{s=0}^{\infty} \psi_s\psi_{s+m}$ for $m > 0$.

Since $\mathrm{cov}(y_{t+m},y_t) = \mathrm{cov}(y_t,y_{t-m})$ for $m > 0$, one has $\gamma_{-m} = \gamma_m$. It is sometimes convenient to summarize the autocovariances of a stationary series x in terms of an *autocovariance generating function* (ACGF)

(7) $g_x(z) = \sum_{s=-\infty}^{\infty} \mathrm{cov}(x_t,x_{t-|s|}) \cdot z^s$.

The ACGF has a useful convolution property: If a linear transformation $C(L) = \sum_{s=-\infty}^{\infty} c_s L^s$ is applied to a stationary series x_t, then $y_t = C(L)x_t$ has $g_y(z) = C(z)C(1/z)g_x(z)$. To verify this, note that $\mathrm{cov}(y_t,y_{t-m}) = \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} c_ic_j \cdot \mathrm{cov}(x_{t-i},x_{t-m-j})$, and $\mathrm{cov}(x_{t-i},x_{t-m-j}) = \mathrm{cov}(x_t,x_{t-m+i-j})$. Then

$$g_y(z) \equiv \sum_{s=-\infty}^{\infty} z^m \mathrm{cov}(y_t,y_{t-m}) = \sum_{m=-\infty}^{\infty}\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} z^m \cdot c_ic_j \cdot \mathrm{cov}(x_{t-i},x_{t-m-j})$$

$$= \sum_{m=-\infty}^{\infty}\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} c_iz^{-i} \cdot c_jz^j \cdot \mathrm{cov}(x_{t-i},x_{t-m-j})z^{m+i-j} = \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} c_iz^{-i} \cdot c_jz^j \cdot g_x(z),$$

with the last equality holding since summing over m for each fixed i and j gives $g_x(z)$.

Let $\eta_t = \varepsilon_t + b_1\varepsilon_{t-1} + ... + b_q\varepsilon_{t-q}$. Then $\mathbf{E}\eta_t = \sigma^2(b_0^2 + ... + b_q^2)$, $\mathbf{E}\eta_t\eta_{t-m} = \sigma^2(b_mb_0 + ... + b_qb_{q-m})$ for $1 < m \le q$, and zero for $m > q$, and $\mathbf{E}\eta_t y_{t-m} = \sigma^2(b_m\psi_0 + ... + b_q\psi_{q-m})$ for $0 \le m \le q$, zero for $m > q$. The i.i.d. series ε_t has the constant ACGF $g_\varepsilon(z) = \sigma^2$ since all its autocovariances are zero. Then, applying the ACGF convolution formula to $\eta_t = (I + B(L))\varepsilon_t$ yields

(8) $g_\eta(z) = \sigma^2\{1+B(z))(1+B(1/z)) = \sigma^2 \sum_{s=-q}^{q} (b_{|s|}b_0 + ... + b_qb_{q-|s|})z^s$.

For example, $q = 1$ yields $g_\eta(z) = \sigma^2((1+b_1^2) + b_1z + b_1z^{-1})$, and $q = 2$ yields $g_\eta(z) = \sigma^2((1+b_1^2+b_2^2) + b_1(1+b_2)z + b_1(1+b_2)z^{-1} + b_2z + b_2z^{-1})$. Applying the convolution formula to $y_t = (I - A(L))^{-1}\eta_t$, it has the ACGF

(9) $g_y(z) = \frac{1}{1-A(z)} \cdot \frac{1}{1-A(1/z)} \cdot g_\eta(z) = \sigma^2 \frac{1+B(z)}{1-A(z)} \cdot \frac{1+B(1/z)}{1-A(1/z)} = \sigma^2 \Psi(z) \cdot \Psi(1/z).$

If all the roots of the polynomial 1 + B(z) lie outside the unit circle, then I + B(L) is invertible, and there is an AR(∞) representation of (1), written formally as

(10) $\frac{I - A(L)}{I + B(L)} \cdot y_t = \varepsilon_t.$

It is not a condition for stationarity that I + B(L) be invertible. However, it is always possible to re-scale the ε's and redefine B(L) so that it has the same ACGF, but all the roots of 1 + B(z) lie outside or on the unit circle. For example, $\eta_t = (I + L/z_1)\varepsilon_t$ has ACGF $\sigma^2((1+z_1^{-2}) + z/z_1 + 1/zz_1)$, while $\eta_t = (I + z_1 L)(\varepsilon_t/z_1)$ has ACGF $(\sigma^2/z_1^2)((1+z_1^2) + zz_1 + z_1/z)$, which is the same. Then one can factor B(L) into an invertible term and a non-invertible term that has unit roots.

For some purposes, it is convenient to rewrite the ARMA model (1) by defining the (p+q)×1 vectors $h' = (1,0,...,0)$ with a one in the first component, $r' = (1,0,...,0,1,0,...,0)$ with ones in the first and p+1 components and zeros elsewhere, and $\zeta_t' = (y_t-\mu,...,y_{t-p+1}-\mu,\varepsilon_t,...,\varepsilon_{t-q+1})$. Then

(11) $\zeta_{t+1} = F\zeta_t + r\varepsilon_{t+1}$ and $y_{t+1}-\mu = h'\zeta_{t+1}$,

where

(12) $F = \begin{bmatrix} \boldsymbol{a}' & \boldsymbol{b}' \\ \boldsymbol{I}_{p-1,p} & \boldsymbol{0}_{p-1,q} \\ \boldsymbol{0}_{1,p} & \boldsymbol{0}_{1,q} \\ \boldsymbol{0}_{q-1,p} & \boldsymbol{I}_{q-1,q} \end{bmatrix}$

with $\mathbf{a}' = (a_1,...,a_p)$, $\mathbf{b}' = (b_1,...,b_q)$, $\mathbf{I}_{rs}$ an r×s matrix with ones down the diagonal and zeros elsewhere, and $\mathbf{0}_{rs}$ an r×s matrix of zeros. This is called a *state space representation* of the ARMA(p,q) model, with ζ_t called the *state vector*, $\zeta_{t+1} = F\zeta_t + r\varepsilon_{t+1}$ called the *state equation*, and $y_{t+1}-\mu = h'\zeta_{t+1}$ called the *observation equation*. State space representations are not unique, and the one given here is easy to interpret but not minimal in terms of dimensionality; see Harvey, p. 95-98, for a more compact and commonly used state space representation for the ARMA model.

2. Prediction

Let $\boldsymbol{G}_t$ denote all of history up through t, including the realizations of y_s and ε_s for $s \le t$. Then,

(13) $\mathbf{E}(y_{t+1}|\boldsymbol{G}_t) = a_1E(y_t|\boldsymbol{G}_t) + ... + a_pE(y_{t-p+1}|\boldsymbol{G}_t) + E(\varepsilon_{t-1}|\boldsymbol{G}_t) + b_1E(\varepsilon_{t-2}|\boldsymbol{G}_t) + ... + b_qE(\varepsilon_{t-q-1}|\boldsymbol{G}_t)$

$= a_1y_t + ... + a_py_{t-p+1} + b_1\varepsilon_t + ... + b_q\varepsilon_{t-q+1}.$

As a shorthand, let $y_{t+1|t} = \mathbf{E}(y_{t+1}|\boldsymbol{G}_t)$; this is the forecast of y_{t+1} given $\boldsymbol{G}_t$ that minimizes mean square error. An implication of the formula for $y_{t+1|t}$ is that $\varepsilon_t = y_t - y_{t|t-1}$. Similarly, the minimum-MSE m-period ahead forecast $y_{t+m|t} = \mathbf{E}(y_{t+m}|\boldsymbol{G}_t)$ is obtained using the recursion

(14) $y_{t+m|t} = a_1 \cdot y_{t-1+m|t} + ... + a_p \cdot y_{t-p+m|t} + b_1\varepsilon_{t+m-1|t} + ... + b_q\varepsilon_{t+m-q|t},$

where $y_{t-I+m|t} = y_{t-I+m}$ and $\varepsilon_{t+I-m|t} = \varepsilon_{t+I-m}$ if $I \ge m$, and $\varepsilon_{t+I-m|t} = 0$ if $I < m$. The conditional variance of the forecast error $\lambda_{t+m|t} = y_{t+m} - y_{t+m|t}$ is

(15) $v_{t+m|t} = \mathbf{E}(\lambda_{t+m|t}|\boldsymbol{G}_t).$

A convenient way of summarizing the forecasting formulas is in terms of the state space representation above, where the minimum MSE forecast is

(16) $\zeta_{t+m|t} = F\zeta_{t+m-1|t} = F^m\zeta_t.$

The forecast error is $\eta_{t+m|t} = \zeta_{t+m} - \zeta_{t+m|t} = \sum_{s=0}^{m-1} F^s r\varepsilon_{t+m-s} = r\varepsilon_{t-m} + F\eta_{t-m-1|t}$, implying $\varepsilon_{t+1} = \zeta_{1,t+1} - \zeta_{1,t+1|t}$.

The conditional covariance matrix of the forecast errors is

(17) $V_{t+m|t} = FV_{t+m-1|t}F' + \sigma^2\mathbf{U} = \sum_{s=0}^{m-1} \sigma^2F^sUF'^s,$

where $\mathbf{U}$ is an array that has a one in the northwest corner and zeros elsewhere. The updating formulas $\zeta_{t+m|t} = F\zeta_{t+m-1|t}$ and $V_{t+m|t} = FV_{t+m-1|t}F' + \sigma^2U$ are versions of what is called the *Kalman filter*. This formulation has broader application in state space models for time series analysis, a topic that will be discussed elsewhere.

For some purposes, it is useful to predict backwards to period t from information after t. Multiplying the equations $(I-A(L))y_t = \eta_t$ and $\eta_t = (I+B(L))\varepsilon_t$ by L^{-p} and L^{-q} respectively gives

(18) $E(y_t|y_{t+1},...,\eta_{t+1},...) = (y_{t+p} - a_1y_{t+p-1} + ... + a_{p-1}y_{t+1} - \eta_{t+p})/a_p,$

$E(\eta_t|\varepsilon_{t+1},...,\eta_{t+1},...) = (\eta_{t+q} - \varepsilon_{t+q} - b_1\varepsilon_{t+q-1} + ... + b_{q-1}\varepsilon_{t+1})/b_q.$

3. Method of Moments Estimation

The model (1) can be written as a regression model

(19) $y_t = c + a_1 y_{t-1} + ... + a_p y_{t-p} + \eta_t$

where $c = (1 - a_1 - ... - a_p)\mu$ and the MA(q) disturbances η_t have a covariance matrix given by the formulas following (6). Since η_t is correlated with $y_{t-1},...,y_{t-q}$ but uncorrelated with earlier y's, (19) can be estimated using 2SLS with $1,y_{t-q-1},...,y_{t-q-p}$ as instruments, and observations t = p+q+1,...,T. This method then loses the first p+q observations in order to get the instruments.

To estimate the MA coefficients, first retrieve the residuals η_{tT} from the 2SLS estimation of (19), and form the empirical ACGF,

(20) $g_{\eta T}(z) = \sum_{s=-q}^{q} \kappa_s z^s,$

where $\kappa_s = \frac{1}{T} \sum_{t=p+q+|s|+1}^{T} \eta_{tT}\eta_{t-|s|,T}$ is an empirical estimate of $\mathbf{E}\, \eta_t\eta_{t-|s|}$.

From (8), the theoretical ACGF for η is a bilinear function of the MA coefficients. Then, estimating the MA coefficients so that the theoretical and empirical ACGF coincide is relatively practical. As noted earlier, the solution is not in general unique, and it is preferable to pick estimates that give a stable root rather than an unstable one. This can be achieved by conducting the search over the q possible roots of 1 + B(z) subject to the constraint that these roots lie outside or on the unit circle.

For the example of a MA(1) component in the ARMA model, matching the ACGF terms yields $\kappa_0 = \sigma^2(1 + b_1^2)$ and $\kappa_1 = \sigma^2 b_1$, which yields the quadratic $\kappa_1 - \kappa_0 b_1 + \kappa_1 b_1^2 = 0$. This has the solution $b_1 = \kappa_0/2\kappa_1 \pm (\kappa_0^2 - 4\kappa_1^2)^{1/2}/2\kappa_1$. If $\kappa_0 > 2\kappa_1$, there are two real roots, and $\kappa_0/2\kappa_1 - (\kappa_0^2 - 4\kappa_1^2)^{1/2}/2\kappa_1$ gives the stable root. If $\kappa_0 \le 2\kappa_1$, there is no real root, and the empirical ACGF cannot be matched exactly by a MA(1) model. In this case, $b_1 = \text{sign}(\kappa_1)$ yields the MA(1) model with a root on or outside the unit circle that is closest to the empirical ACGF. For the example of a MA(2) component, writing 1 + B(z) in terms of its roots and matching the ACGF terms yields $\kappa_0 = \sigma^2(1 + 1/z_1^2 + 1/z_2^2)$, $\kappa_1 = \sigma^2(1/z_1)(1 + 1/z_1z_2 + 1/z_2^2)$, and $\kappa_2 = \sigma^2/z_1z_2$. The estimator would then be obtained by a search in σ^2, z_1, and z_2 subject to the restriction that z_1 and z_2 are in the complex plane, on or outside the unit circle, and either are both real, or are convex congugates.

The estimators described above are not the most efficient available among those employing only moment conditions. However, it will be convenient to develop the alternatives as estimators for the case of normal disturbances, and then note that they will be consistent even without normality.

4. Maximum Likelihood Estimation

Suppose the disturbances ε_t in (1) are normal. Then, vector $(y_1,...,y_T)$ is then multivariate normal with mean μ and a covariance matrix that is a band matrix Σ, with all coefficients s places off the diagonal equal to γ_s:

$$(21)\quad \Sigma = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{T-2} & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{T-3} & \gamma_{T-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{T-4} & \gamma_{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{T-2} & \gamma_{T-3} & \gamma_{T-4} & \cdots & \gamma_0 & \gamma_1 \\ \gamma_{T-1} & \gamma_{T-2} & \gamma_{T-3} & \cdots & \gamma_1 & \gamma_0 \end{bmatrix}.$$

The autocovariances γ_s in Σ are functions of the deep parameters of the model, σ^2 and the coefficients of A(L) and B(L); these are given by the coefficients of the ACGF (9). A brute force approach to estimating the model, which is efficient under the assumptions of normality, is then to maximize the log likelihood function

$$(22)\quad L = -(T/2)\log(2\pi) - (\tfrac{1}{2})\log(\det(\Sigma)) - (\tfrac{1}{2})(y_1-\mu,...,y_T-\mu)\Sigma^{-1}(y_1-\mu,...,y_T-\mu)'$$

in the parameters μ, σ^2, $a_1,...,a_p$, $b_1,...,b_q$. The likelihood function is minimized in μ at the sample mean μ_T. Using the formulas for differentiation of determinants and inverses, the first-order-condition for a parameter θ is

$$(23)\quad \partial L/\partial\theta = -\tfrac{1}{2}\Sigma^{-1}(I - uu'\Sigma^{-1})\cdot\partial\Sigma/\partial\theta,$$

where $u' = (y_1-\mu_T,..,y_T-\mu_T)$ is the vector of deviations from sample mean.

The practical problem with the brute force approach is that the parameters of the ARMA process appear in L non-linearly, deep within the T×T matrix Σ. Therefore, considerable effort in time-series analysis goes to reformulating the maximum likelihood problem in ways that are more tractable computationally.

The model (1) can be rewritten as

$$(24)\quad \varepsilon_t = y_t - a_1y_{t-1} - ... - a_py_{t-p} - b_1\varepsilon_{t-1} - ... - b_q\varepsilon_{t-q}.$$

Taking expectations conditioned on the information G_{t-1}, this equation implies $0 = y_{t|t-1} - a_1y_{t-1} - ... - a_py_{t-p} - b_1\varepsilon_{t-1} - ... - b_q\varepsilon_{t-q}$, and hence

(25) $\varepsilon_t = y_t - y_{t|t-1}$.

The mapping from ε's to y's is linear, with a transformation matrix that is triangular (i.e., y_t depends only on ε's at or before t) with ones on the diagonal. Then, the Jacobean of the transformation from $(y_1,...,y_T)$ to $(\varepsilon_1,...,\varepsilon_T)$ is one, and the log density of $(\varepsilon_1,...,\varepsilon_T)$ conditioned on earlier ε's is

(26) $L = -(T/2)\log(2\pi) - (½)\log(\sigma^2) - ½ \sum_{t=1}^{T} \varepsilon_t^2/\sigma^2$,

with the ε_t defined (recursively) as functions of the lag parameters from (24). This is called the *predictive error decomposition* of the likelihood. One approach to estimation is to condition on $(y_1,...,y_p)$ and $(\varepsilon_{p-q},...,\varepsilon_{p-1})$, so that (24) gives the ε's for t = p+1,...,T, and then to maximize the conditional log likelihood, which is equivalent to minimizing the *conditional sum of squares*,

(27) $CSS = \sum_{t=p+1}^{T} \varepsilon_t^2$,

in the lag parameters. This gives a pre-estimator (because of dependence on ε_{p-1},..) that is equivalent to the solution at convergence obtained by iteratively applying least squares to the equation below, with the lagged ε's computed using (24) and the lag coefficient estimates from the previous round:

(28) $y_t = a_1 y_{t-1} + ... + a_p y_{t-p} + b_1 \varepsilon_{t-1} + ... + b_q \varepsilon_{t-q} + \nu_t$ for t = p+1,...,T.

The final step is to get rid of the dependence of the estimator on the initial ε's. This is often done in the computationally expedient way of replacing them by their unconditional expectations, which are zero. A superior procedure, only moderately more difficult, is described later.

For AR(p) models, a useful simplification of (22) comes from noting that the density of $(y_1,...,y_T)$ can be written as the product of the p-dimensional multivariate normal density of $(y_1,...,y_p)$ and the 1-dimensional conditional densities of y_t given $y_{t-1},...,y_{t-p}$ for t = p+1,...,T. In this formulation, the log likelihood is

(29) $L = -(T/2)\log(2\pi) - ½\log(\det(\Sigma_p)) - ½(y_1-\mu,...,y_p-\mu)\Sigma_p^{-1}(y_1-\mu,...,y_p-\mu)$
$- ½\log(\sigma^2) - (1/2\sigma^2) \sum_{t=p+1} (y_t - a_1 y_{t-1} - ... - a_p y_{t-p})^2$,

where Σ_p is the covariance matrix of the first p observations. If, further, one conditions on $(y_1,...,y_p)$, the resulting log likelihood is maximized when the quadratic form $\sum_{t=p+1}^{T} (y_t - a_1 y_{t-1} - ... - a_p y_{t-p})^2$ is

minimized; this is exactly the same as the regression (19) in the case q = 0. Conditioning on the first p observations, maximization of this likelihood function is equivalent to (non-iterative) least squares applied to the model (28), or to minimizing CSS in (27).

Now consider the full stationary ARMA(p,q) model (1), and its representation (11) in state space form. Let z_t denote the conditional expectation of ζ_t given $y_t,y_{t-1},...,y_1$. This is the optimal predictor of ζ_t given this information. Given normality, this is a linear function of the y's. Let P_t denote the conditional MSE of the deviation $\zeta_t - z_t$; i.e., $P_t = \mathbf{E}\{(\zeta_t - z_t)(\zeta_t - z_t)' \,|\, y_t,y_{t-1},...,y_1\}$. Similarly, define $z_{t|t-1} = \mathbf{E}(\zeta_t|y_{t-1},...,y_1)$ and $P_{t|t-1} = \mathbf{E}\{(\zeta_t - z_t)(\zeta_t - z_t)' \,|\, y_{t-1},...,y_1\}$. Given the state equation $\zeta_t = F\zeta_{t-1} + r\varepsilon_t$ and the observation equation $y_t = h'\zeta_t$, and taking conditional expectations, one obtains the formulas

(30) $z_{t|t-1} = Fz_{t-1}$, $P_{t|t-1} = FP_{t-1}F' + \sigma^2 rr'$, $y_{t|t-1} = h'z_{t|t-1}$,

and from these the updating relationships for projections,

(31) $\nu_t \equiv y_t - y_{t|t-1} = h'(\zeta_t - z_{t|t-1})$,

$f_t = h'P_{t|t-1}h$

$z_t = z_{t|t-1} + P_{t|t-1}h(y_t - h'z_{t|t-1})/f_t$

$P_t = P_{t|t-1} - P_{t|t-1}hh'P_{t|t-1}/f_t$.

These formulas (with a different notation) are derived and discussed in Harvey, p. 85-86, for a more general model that includes stationary ARMA as a special case.

The formulas in (31) can be employed, with an initialization for z_0 and P_0, to calculate the exact joint normal density function of $(y_1,...,y_T)$:

(32) $L = -(T/2)\log(2\pi) - \tfrac{1}{2} \sum_{t=1}^{T} \log(f_t) - \tfrac{1}{2} \sum_{t=1}^{T} \nu_t^2/f_t$

This is a *predictive error decomposition* form of the log likelihood. The maximum likelihood estimates can also be given an interpretation of minimizing a CSS; see Harvey, p. 90.

Harvey, p. 88, also describes the construction of starting values. For the stationary ARMA model, they are $z_0 = (I - F)^{-1}c$, where c is a vector with μ in the first p components and 0 in the remaining q components, and $\mathrm{vec}(P_0) = \sigma^2(I - F\otimes F)^{-1}\mathrm{vec}(rr')$. Absent normality, the predictors above remain best linear predictors, and the estimators continue to have an interpretation of minimum CSS estimators that use all of the sample information on the first two moments, with a Bayesian interpretation of the starting values.