Maximum Likelihood Estimates

The maximum likelihood estimates are more efficient than moment estimates, and are the ones to use in practice. For ARMA models, the log likelihood is of the form

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = -\tfrac{n}{2} \log \sigma_a^2 - \tfrac{1}{2} \log D(\boldsymbol{\phi}, \boldsymbol{\theta}) - \tfrac{1}{2} S(\boldsymbol{\phi}, \boldsymbol{\theta})/\sigma_a^2 + C,$$

where $D(\boldsymbol{\phi}, \boldsymbol{\theta})$, $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ are independent of σ_a^2. Maximizing $l(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2)$ w.r.t. σ_a^2, one has $\hat{\sigma}_a^2 = S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})/n$, where $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ maximizes the profile likelihood

$$\tilde{l}(\boldsymbol{\phi}, \boldsymbol{\theta}) = -\tfrac{n}{2} \log S(\boldsymbol{\phi}, \boldsymbol{\theta}) - \tfrac{1}{2} \log D(\boldsymbol{\phi}, \boldsymbol{\theta}) + \tilde{C}.$$

To obtain $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$, iterations are needed, for which the moment estimates may serve as starting values. The ML estimates are also needed for the exact evaluation of AIC or BIC for model selection.

The minimizer of $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ gives the LS estimates of $\boldsymbol{\phi}$, $\boldsymbol{\theta}$.

Conditional Likelihood for ARIMA(p,d,q)

Observing $z_1, \ldots, z_n$ from ARMA(p,q), the log likelihood conditional on $\mathbf{z}_* = (z_0, \ldots, z_{1-p})$ and $\mathbf{a}_* = (a_0, \ldots, a_{1-q})$ is

$$\begin{aligned} l_*(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) &= -\frac{n}{2} \log \sigma_a^2 - \frac{1}{2\sigma_a^2} \sum_{t=1}^{n} a_t^2 + C \\ &= -\frac{n}{2} \log \sigma_a^2 - \frac{1}{2\sigma_a^2} S_*(\boldsymbol{\phi}, \boldsymbol{\theta}) + C, \end{aligned}$$

where $a_t = z_t - \phi_1 z_{t-1} - \cdots - \phi_p z_{t-p} + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}$.

For n large, one may choose to sum from a_{p+1}^2 onward and set previous a's equal to 0. For $q = 0$, this yields the estimation of ϕ_j via the minimization of the LS score,

$$\textstyle\sum_{t=p+1}^{n} (z_t - \phi_1 z_{t-1} - \cdots - \phi_p z_{t-p})^2.$$

For ARIMA(p,d,q), take $w_t = \nabla^d z_t$ and work on w_t.

Exact Likelihood for AR(p)

Observing $z_1, \ldots, z_n$ from AR(p), one has

$$f(\mathbf{z}|\boldsymbol{\phi}, \sigma_a^2) = f(\mathbf{z}_{(p)}|\mathbf{z}_p, \boldsymbol{\phi}, \sigma_a^2) f(\mathbf{z}_p|\boldsymbol{\phi}, \sigma_a^2),$$

where $\mathbf{z}_p = (z_1, \ldots, z_p)^T$ and $\mathbf{z}_{(p)} = (z_{p+1}, \ldots, z_n)^T$. One has

$$f(\mathbf{z}_{(p)}|\mathbf{z}_p, \boldsymbol{\phi}, \sigma_a^2) \propto (\sigma_a^2)^{-(n-p)/2} \exp(-\textstyle\sum_{t=p+1}^n a_t^2/2\sigma_a^2),$$

where $a_t = z_t - \phi_1 z_{t-1} - \cdots - \phi_p z_{t-p}$, and

$$f(\mathbf{z}_p|\boldsymbol{\phi}, \sigma_a^2) \propto (\sigma_a^2)^{-p/2} |\mathbf{M}_p|^{1/2} \exp(-\mathbf{z}_p^T \mathbf{M}_p \mathbf{z}_p / 2\sigma_a^2),$$

where $\mathbf{M}_p = (\sigma_a^2/\gamma_0)\mathbf{P}_p^{-1}$. The exact log likelihood is thus

$$l(\boldsymbol{\phi}, \sigma_a^2) = -\tfrac{n}{2} \log \sigma_a^2 + \tfrac{1}{2} \log |\mathbf{M}_p| - \tfrac{1}{2} S(\boldsymbol{\phi})/\sigma_a^2,$$

where $S(\boldsymbol{\phi}) = \mathbf{z}_p^T \mathbf{M}_p \mathbf{z}_p + \sum_{t=p+1}^n a_t^2$ is the *exact least squares.*

For $z_t = \phi z_{t-1} + a_t$, $\mathbf{M}_1 = |\mathbf{M}_1| = 1 - \phi^2$, so

$$l(\phi, \sigma_a^2) = -\tfrac{n}{2} \log \sigma_a^2 + \tfrac{1}{2} \log(1-\phi^2) - \tfrac{1}{2}\{(1-\phi^2) z_1^2 + \textstyle\sum_{t=2}^n (z_t - \phi z_{t-1})^2\}/\sigma_a^2.$$

Innovations Algorithm

For $\{z_t\}$ with $E[z_t] = 0$ and $E[z_t z_s] = \kappa_{t,s}$, consider the one-step prediction of z_{t+1} given $z_t, \ldots, z_1$, $\hat{z}_{t+1} = E[z_{t+1}|z_t, \ldots, z_1]$. One has $E[(z_{t+1} - \hat{z}_{t+1})z_j] = 0$, for all $j \leq t$. Note that $\hat{z}_1 = 0$. Write

$$\hat{z}_{t+1} = \sum_{j=0}^{t-1} \theta_{t,t-j}(z_{j+1} - \hat{z}_{j+1}) = \sum_{j=0}^{t-1} \theta_{t,t-j} e_{j+1},$$

where the *innovations* $e_{j+1} = z_{j+1} - \hat{z}_{j+1}$ are uncorrelated. Write $\text{var}[e_{t+1}] = v_t$; $v_0 = \kappa_{1,1}$. For $0 \leq j < t$, one has

$$\begin{aligned}
\theta_{t,t-j} &= v_j^{-1} E[z_{t+1}(z_{j+1} - \hat{z}_{j+1})] \\
&= v_j^{-1} E[z_{t+1}(z_{j+1} - \textstyle\sum_{k=0}^{j-1} \theta_{j,j-k}(z_{k+1} - \hat{z}_{k+1}))] \\
&= v_j^{-1}(\kappa_{t+1,j+1} - \textstyle\sum_{k=0}^{j-1} \theta_{j,j-k}\theta_{t,t-k} v_k), \\
v_t &= \kappa_{t+1,t+1} - \textstyle\sum_{j=0}^{t-1} \theta_{t,t-j}^2 v_j.
\end{aligned}$$

Note that $\theta_{t,t-j}$ and v_t can be obtained recursively in the order of v_0; $\theta_{1,1}$, v_1; $\theta_{2,2}$, $\theta_{2,1}$, v_2; $\theta_{3,3}$, $\theta_{3,2}$, $\theta_{3,1}$, v_3; $\ldots$.

Exact Likelihood for MA(q)

Consider an MA(q) process with autocovariance γ_k. For $t-j>q$, from $\kappa_{t+1,j+1}=\gamma_{t-j}=0$ and the recursion formula for $\theta_{t,j}$, one has $\theta_{t,t-j}=0$. It follows that

$$\hat{z}_{t+1}=\begin{cases}\sum_{j=1}^{t}\theta_{t,j}(z_{t-j+1}-\hat{z}_{t-j+1}), & t=1,\dots,q-1,\\ \sum_{j=1}^{q}\theta_{t,j}(z_{t-j+1}-\hat{z}_{t-j+1}), & t\geq q.\end{cases}$$

Write $v_t=\tilde{v}_t\sigma_a^2$, where $\tilde{v}_t$ do not depend on σ_a^2, one has

$$l(\boldsymbol{\theta},\sigma_a^2)=-\frac{n}{2}\log\sigma_a^2-\frac{1}{2}\sum_{t=1}^{n}\log\tilde{v}_t-\frac{1}{2\sigma_a^2}\sum_{t=1}^{n}\frac{(z_t-\hat{z}_t)^2}{\tilde{v}_{t-1}}.$$

For $z_t=a_t-\theta a_{t-1}$, one has $\theta_{t,1}=-\theta/\tilde{v}_{t-1}$, $\theta_{t,j}=0$, $j>1$, where $\tilde{v}_0=1+\theta^2$, $\tilde{v}_t=1+\theta^2(\tilde{v}_{t-1}-1)/\tilde{v}_{t-1}\rightarrow 1$. It follows that

$$\hat{z}_1=0,\qquad \hat{z}_{t+1}=-\theta(z_t-\hat{z}_t)/\tilde{v}_{t-1},\quad t\geq 1.$$

Exact Likelihood for ARMA(p,q) – I

For $\phi(B)z_t = \theta(B)a_t$ with autocovariance γ_k, define

$$u_t = z_t\, I_{[t\le m]} + \phi(B)z_t\, I_{[t>m]},$$

where $m = \max(p,q)$. $E[u_t u_s] = \kappa_{t,s}$ is given by

$$\kappa_{t,s} = \begin{cases} \gamma_{t-s}, & s \le t \le m, \\ \gamma_{t-s} - \sum_{k=1}^{p} \phi_j \gamma_{t-s-k}, & s \le m < t, \\ \sigma_a^2 \sum_{k=0}^{q} \theta_k \theta_{k+t-s}, & m < s \le t \end{cases}$$

where $\theta_0 = -1$ and $\theta_k = 0$, $k > q$. When $t \ge m$ and $t - j > q$, $\kappa_{t+1,j+1} = 0$, so $\theta_{t,t-j} = 0$. It follows that

$$\hat{u}_{t+1} = \begin{cases} \sum_{j=1}^{t} \theta_{t,j}(u_{t-j+1} - \hat{u}_{t-j+1}), & t < m, \\ \sum_{j=1}^{q} \theta_{t,j}(u_{t-j+1} - \hat{u}_{t-j+1}), & t \ge m. \end{cases}$$

Note that for $t > m$, $\hat{u}_t = \hat{z}_t - \phi_1 z_{t-1} - \cdots - \phi_p z_{t-p}$.

Exact Likelihood for ARMA(p,q) – II

It is easily seen that $u_t - \hat{u}_t = z_t - \hat{z}_t$, hence the log likelihood has the same expression as for MA(q), but

$$\hat{z}_{t+1} = \begin{cases} \sum_{j=1}^{t} \theta_{t,j}(z_{t-j+1} - \hat{z}_{t-j+1}), & t < m, \\ \sum_{j=1}^{p} \phi_j z_{t-j} + \sum_{j=1}^{q} \theta_{t,j}(z_{t-j+1} - \hat{z}_{t-j+1}), & t \geq m. \end{cases}$$

Consider $z_t = \phi z_{t-1} + a_t - \theta a_{t-1}$. One has

$$\kappa_{t,s} = \begin{cases} \sigma_a^2(1 + \theta^2 - 2\phi\theta)/(1 - \phi^2), & t = s = 1, \\ \sigma_a^2(1 + \theta^2), & t = s \geq 2, \\ \sigma_a^2(-\theta), & t - s = 1, \\ 0, & \text{otherwise.} \end{cases}$$

The same as for MA(q), $\theta_{t,1} = -\theta/\tilde{v}_{t-1}$, $\theta_{t,j} = 0$, $j > 1$, and $\tilde{v}_t = 1 + \theta^2(\tilde{v}_{t-1} - 1)/\tilde{v}_{t-1}$. Setting $\theta = 0$, one has $\tilde{v}_0 = (1 - \phi^2)^{-1}$, and for $t \geq 2$, $\hat{z}_t = \phi z_{t-1}$ and $\tilde{v}_{t-1} = 1$.

Non-recursive Likelihood of ARMA(p,q) – I

Setting $\theta_0 = -1$ and writing $z_{t-k} = \sum_j \psi_j a_{t-k-j} = \sum_j \psi_{j-k} a_{t-j}$, where $\psi_j = 0$ for $j < 0$, one has

$$\begin{aligned}\gamma_k = E[z_t z_{t-k}] &= E\big[\sum_{j=1}^p \phi_j z_{t-j} - \sum_{j=0}^q \theta_j a_{t-j})z_{t-k}\big] \\ &= \sum_{j=1}^p \phi_j \gamma_{k-j} - \sigma_a^2 \sum_{j=0}^q \theta_j \psi_{j-k} = \sum_{j=1}^p \phi_j \gamma_{k-j} - \sigma_a^2\, b_k,\end{aligned}$$

where $b_k = \sum_{j=0}^q \theta_j \psi_{j-k}$. The covaviance of $(z_1, \ldots, z_n)$ is seen to be, for $\tilde{\gamma}_k = \gamma_k / \sigma_a^2$,

$$\sigma_a^2 \begin{pmatrix} \tilde{\gamma}_0 & \tilde{\gamma}_1 & \tilde{\gamma}_2 & \cdots & \tilde{\gamma}_{n-1} \\ \tilde{\gamma}_1 & \tilde{\gamma}_0 & \tilde{\gamma}_1 & \cdots & \tilde{\gamma}_{n-2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \tilde{\gamma}_{n-1} & \tilde{\gamma}_{n-2} & \tilde{\gamma}_{n-3} & \cdots & \tilde{\gamma}_0 \end{pmatrix}.$$

Non-recursive Likelihood of ARMA(p,q) – II

Given $(\tilde{\gamma}_0, \ldots, \tilde{\gamma}_{p-1})$, $\tilde{\gamma}_k$, $k \geq p$ can be obtained recursively, and one has $p+1$ linear equations involving $(\tilde{\gamma}_0, \ldots, \tilde{\gamma}_p)$,

$$\begin{aligned}
\tilde{\gamma}_0 &= \phi_1 \tilde{\gamma}_1 + \cdots + \phi_p \tilde{\gamma}_p - b_0, \\
\tilde{\gamma}_1 &= \phi_1 \tilde{\gamma}_0 + \cdots + \phi_p \tilde{\gamma}_{p-1} - b_1, \\
&\cdots \\
\tilde{\gamma}_p &= \phi_1 \tilde{\gamma}_{p-1} + \cdots + \phi_p \tilde{\gamma}_0 - b_p.
\end{aligned}$$

For example, with $p = 3$, one solves

$$\begin{pmatrix} 1 & -\phi_1 & -\phi_2 & -\phi_3 \\ -\phi_1 & 1-\phi_2 & -\phi_3 & 0 \\ -\phi_2 & -\phi_1-\phi_3 & 1 & 0 \\ -\phi_3 & -\phi_2 & -\phi_1 & 1 \end{pmatrix} \begin{pmatrix} \tilde{\gamma}_0 \\ \tilde{\gamma}_1 \\ \tilde{\gamma}_2 \\ \tilde{\gamma}_3 \end{pmatrix} = - \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

Parameterization of ARMA(p,q)

Recall the Durbin-Levinson algorithm for recursive solutions to the Yule-Walker equations of an AR(p) model: for $k = 1, \ldots, p$, $\phi_{kk} = \ldots$, $\boldsymbol{\phi}_{k,k-1} = \boldsymbol{\phi}_{k-1} - \phi_{kk}\tilde{\boldsymbol{\phi}}_{k-1}$, where $\boldsymbol{\phi}_k^T = (\boldsymbol{\phi}_{k,k-1}^T, \phi_{kk})$, $\tilde{\boldsymbol{\phi}}_h$ is $\boldsymbol{\phi}_h$ in reverse order, and ϕ_{kk} is $\mathrm{corr}(z_0, z_k | z_1, \ldots, z_{k-1})$.

As long as $\phi_{kk} \in (-1, 1)$, $\boldsymbol{\phi}_p$ resulting from the Durbin-Levinson algorithm ensures a $\phi(B)$ with roots outside of the unit circle. Numerically, one may transform via $\phi_{kk} = (1 - e^{-\beta_k})/(1 + e^{-\beta_k})$ and work with $\beta_k \in (-\infty, \infty)$; this allows the use of unconstrained optimization.

To ensure invertibility, one may parameterize $\theta(B)$ similarly.

Numerical Optimization

Using the proceeding techniques, one is able to calculate the log likelihood function $l(\boldsymbol{\gamma})$ for given parameters $\boldsymbol{\gamma}$, but analytical derivatives are in general *not* available.

A standard approach to optimization using only function evaluations is quasi-Newton methods with numerical derivatives. Quasi-Newton builds up Hessian using gradients, and numerical differentiation approximates $f'(x)$ by $\big(f(x+\delta)-f(x-\delta)\big)/2\delta$ with δ of adequate size.

Assuming convexity with continuous Hessian near the bottom, the algorithms converge to the minimizer reasonably fast, and also return the Hessian at the minimizer if needed.

In `R`, one may use `optim` or `nlm`.

Asymptotic Properties, Regression

With correct model identification and sufficiently large sample size, the ML estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ are consistent and asymptotically normal with mean $(\boldsymbol{\phi}, \boldsymbol{\theta})$. The approximate covariance matrix of $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ is given by $I^{-1}(\hat{\boldsymbol{\gamma}})$, where $I(\boldsymbol{\gamma}) = -\partial^2 l / \partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T$ is the (observed) information matrix.

For regression models with ARIMA errors,

$$z_t = \mathbf{x}_t^T \boldsymbol{\beta} + \epsilon_t, \quad \varphi(B)\epsilon_t = \theta(B)a_t,$$

the likelihood is easily evaluated with $z_t' = z_t - \mathbf{x}_t^T \boldsymbol{\beta}$ replacing z_t. Note that with $d > 0$, $\nabla \mathbf{x}_t$ is taken along with ∇z_t, so coefficients for monomials in t up to order $d-1$ are not estimable.

To explore the proper form of $\mathbf{x}_t^T \boldsymbol{\beta}$, one may use the standard regression tools and count on the robustness of LS estimates of $\boldsymbol{\beta}$.

Residual Analysis

Statistical model building is typically an iterative process, in which one cycles through model identification, model fitting, and model checking. As in a standard regression analysis, a major source of information for model checking is in the residuals.

For an invertible ARMA model $\phi(B)z_t = \theta(B)a_t$, one may simply take as residuals $\hat{a}_t = \hat{\theta}^{-1}(B)\hat{\phi}(B)z_t$, where the needed starting values $z_0, z_{-1}, z_{-2}, \ldots$, can be obtained through "back-forecasting" using the reverse-time model. Alternatively, one may take $\hat{a}_t = e_t/\sqrt{v_{t-1}}$, where $e_t = z_t - \hat{z}_t$ are the innovations with variances v_{t-1}. The latter is suspected to be the ones from `arima`.

For a good fit, $\hat{a}_t$ should behave as white noise. Among simple tools for checking the "whiteness" of the residuals are the raw residual plot, the ACFs and the PACFs.

Ljung-Box-Pierce Portmanteau Tests

The sample ACFs $r_k(\hat{a})$ of the residuals $\hat{a}_t$ are correlated, and for small lags, their variances can be substantially smaller than the nominal n^{-1} calculated for the "true" white noise a_t. Individual assessment of $r_k(\hat{a})$ thus can mislead.

To collectively digest the information contained in the leading lag $r_k(\hat{a})$'s, *portmanteau tests* were developed by Box-Pierce and Ljung-Box. Under the null that the ARMA(p,q) model is correct, one has the asymptotic χ^2_{K-p-q} statistics,

$$Q = n \sum_{k=1}^{K} r_k^2(\hat{a}), \quad \tilde{Q} = n(n+2) \sum_{k=1}^{K} (n-k)^{-1} r_k^2(\hat{a}).$$

The null distribution of $\tilde{Q}$ is closer to χ^2_{K-p-q} than that of Q. The choice of K is somewhat arbitrary, but a larger K tends to lower the power of the test. `Box.test` in `ts` implements these tests.

Cumulative Periodogram

Recall the periodogram of $z_1, \ldots, z_n$ from a stationary process,

$$I(\omega_j) = \frac{1}{n}\left(\sum_{t=1}^{n} z_t \cos 2\pi t\omega_j\right)^2 + \frac{1}{n}\left(\sum_{t=1}^{n} z_t \sin 2\pi t\omega_j\right)^2,$$

where $\omega_j = j/n$. The *cumulative periodogram* on $(0, 1/2)$,

$$C(\omega) = \sum_{0<\omega_j\leq\omega} I(\omega_j) / \sum_{0<\omega_j\leq 1/2} I(\omega_j),$$

is the empirical version of $P(\omega) = \int_{-\omega}^{\omega} p(\lambda)d\lambda = 2\int_0^{\omega} p(\lambda)d\lambda$, where $p(\lambda)$ is the spectral density. To test the hypothesis that the spectral distribution is given by some known $P(\omega)$, one may use the Kolmogorov-Smirnov statistic, $\sup |C(\omega) - P(\omega)|$.

For white noise, the spectral density $p(\omega) = 1$ is uniform, and $P(\omega) = 2\omega$. Tolerance band for $C(\omega)$ under the null can be constructed from the Kolmogorov-Smirnov distribution. The cumulative periodogram check is implemented in `cpgram`.

Model Modification and Testing

Suppose that a model $\varphi_0(B)z_t = \theta_0(B)a_t$ is fitted and the residuals $\hat{a}_t$ are obtained. After examining the $\hat{a}_t$'s, it is suggested that a_t may not be white, but rather is of the form $\tilde{\varphi}(B)a_t = \tilde{\theta}(B)b_t$ with b_t white. Straightforward algebra yields

$$\varphi(B)z_t = \tilde{\varphi}(B)\varphi_0(B)z_t = \tilde{\theta}(B)\theta_0(B)b_t = \theta(B)b_t.$$

Hence, to modify the model, one simply increases the respective orders of $\varphi(B) = \phi(B)\nabla^d$ and $\theta(B)$.

To validate the modification, one may use the likelihood ratio test of nested models. Let $l_0(\hat{\boldsymbol{\varphi}}_0, \hat{\boldsymbol{\theta}}_0, \hat{\sigma}_a^2)$, $l(\hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}_b^2)$ be the maximized log likelihoods. Under the null of $\varphi_0(B)z_t = \theta_0(B)a_t$,

$$2(l(\hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}_b^2) - l_0(\hat{\boldsymbol{\varphi}}_0, \hat{\boldsymbol{\theta}}_0, \hat{\sigma}_a^2)) \sim \chi^2_{\tilde{p}+\tilde{q}},$$

where $\tilde{p}$, $\tilde{q}$ are the orders of $\tilde{\phi}(B)$, $\tilde{\theta}(B)$.