Box-Jenkins Modeling Strategy for Fitting ARMA(p, q) Models

1. Transform the data, if necessary, so that the assumption of covariance stationarity is a reasonable one

2. Make an initial guess for the values of p and q

3. Estimate the parameters of the proposed ARMA(p, q) model

4. Perform diagnostic analysis to confirm that the proposed model adequately describes the data (e.g. examine residuals from fitted model)

Identification of Stationary ARMA(p,q) Processes

Intuition: The mean, variance, and autocorrelations define the properties of an ARMA(p,q) model. A natural way to identify an ARMA model is to match the pattern of the observed (sample) autocorrelations with the patterns of the theoretical autocorrelations of a particular ARMA(p, q) model.

Sample autocovariances/autocorrelations

$$\begin{aligned} \hat{\gamma}_j &= \frac{1}{T}\sum_{t=j+1}^{T}(y_t - \hat{\mu})(y_{t-j} - \hat{\mu}),\ \hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t \\ \hat{\rho}_j &= \frac{\hat{\gamma}_j}{\hat{\gamma}_0} \end{aligned}$$

Sample autocorrelation function (SACF)/correlogram

plot $\hat{\rho}_j$ vs. j

Result: If $Y_t \sim WN(0, \sigma^2)$ then $\rho_j = 0$ for all j and

$$\sqrt{T}\hat{\rho}_j \xrightarrow{d} N(0,1)$$

so that

$$avar(\hat{\rho}_j) = \frac{1}{T}$$

Therefore, a simple $t-$statistic for $H_0 : \rho_j = 0$ is

$$\sqrt{T}\hat{\rho}_j$$

and we reject $H_0 : \rho_j = 0$ if

$$|\hat{\rho}_j| \ > \ 1.96\sqrt{T}$$

Remark: $\hat{\rho}_1, \ldots, \hat{\rho}_k$ are asymptotically independent:

$$\sqrt{T}\hat{\boldsymbol{\rho}}_k \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_k)$$

where $\hat{\boldsymbol{\rho}}_k = (\hat{\rho}_1, \ldots, \hat{\rho}_k)'$.

Box-Pierce and Box-Ljung Q-statistics

The joint statistical significance of $\hat{\rho}_1, \ldots, \hat{\rho}_k$ may be tested using the Box-Pierce Portmanteau statistic

$$Q(k) = T\sum_{j=1}^{k}\hat{\rho}_j^2$$

Assuming $Y_t \sim WN(0, \sigma^2)$ it is straightforward to show that

$$Q(k) \xrightarrow{d} \chi^2(k)$$

Reject $H_0 : \rho_1 = \cdots = \rho_k = 0$ if

$$Q(k) > \chi^2_{0.95}(k)$$

Remark: Box and Ljung show that a simple degrees-of-freedom adjustment improves the finite sample performance:

$$Q^*(k) = T(T+2)\sum_{j=1}^{k}\frac{\hat{\rho}_j^2}{T-j}$$

Partial Autocorrelation Function (PACF)

The kth order partial autocorrelation of $X_t = Y_t - \mu$ is the partial regression coefficient (in the population) ϕ_{kk} in the kth order autoregression

$$X_t = \phi_{1k} X_{t-1} + \phi_{2k} X_{t-2} + \cdots + \phi_{kk} X_{t-k} + \text{error}_t$$

PACF:

$$\text{plot } \phi_{kk} \text{ vs. } k$$

Sample PACF (SPACF)

$$\text{plot } \hat{\phi}_{kk} \text{ vs. } k$$

where ϕ_{kk} is estimated from an AR(k) model for Y_t.

Example: AR(2)

$$\begin{aligned} Y_t - \mu &= \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \varepsilon_t \\ \phi_{11} &\neq 0, \ \phi_{22} = \phi_2, \ \phi_{kk} = 0 \text{ for } k > 2 \end{aligned}$$

Example: MA(1)

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}, \ |\theta| < 1 \text{ (invertible)}$$

Since $|\theta| < 1$, Y_t has an AR(∞) representation

$$\begin{aligned} (Y_t - \mu) &= \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \cdots \\ \phi_j &= -(-\theta)^j \end{aligned}$$

Therefore,

$$\phi_{kk} \neq 0 \text{ for all } k, \text{ and } \phi_{kk} \to 0 \text{ as } k \to \infty$$

Results:

1. AR(p) processes: $\phi_{kk} \neq 0$ for $k \leq p$, $\phi_{kk} = 0$ for $k > p$

2. MA(q) processes: $\phi_{kk} \neq 0$ for all k, and $\to 0$ as $k \to \infty$

Inspection of the SACF and SPACF to identify ARMA models is somewhat of an art rather than a science. A more rigorous procedure to identify an ARMA model is to use formal model selection criteria. The two most widely used criteria are the Akaike information criterion (AIC) and the Bayesian (Schwarz) criterion (BIC or SIC)

$$\begin{aligned}
\mathsf{AIC}(p,q) &= \ln(\hat{\sigma}^2) + \frac{2(p+q)}{T} \\
\mathsf{BIC}(p,q) &= \ln(\hat{\sigma}^2) + \frac{\ln(T)(p+q)}{T} \\
\hat{\sigma}^2 &= \text{estimate of } \sigma^2 \text{ from ARMA}(p,q)
\end{aligned}$$

Intuition: Think of adjusted R^2 in regression:

$\ln(\hat{\sigma}^2)$ measures overall fit

$\frac{2(p+q)}{T}$, $\frac{\ln(T)(p+q)}{T}$ penalty terms for large models

Note: BIC penalizes larger models more than AIC.

How to use AIC, BIC to identify an ARMA(p,q) model

1. Set upper bounds, P and Q for the AR and MA order, respectively

2. Fit all possible ARMA(p,q) models for $p \leq P$ and $q \leq Q$ using a common sample size T

3. The best models satisfy

$$\min_{p \leq P, q \leq Q} \mathsf{AIC}(p,q)$$
$$\min_{p \leq P, q \leq Q} \mathsf{BIC}(p,q)$$

Result: If the true values of p and q satisfy $p \leq P$ and $q \leq Q$ then

1. BIC picks the true model with probability 1 as $T \to \infty$

2. AIC picks largers values of p and q with positive probability as $T \to \infty$

Maximum Likelihood Estimation of ARMA models

For iid data with marginal pdf $f(y_t;\boldsymbol{\theta})$, the joint pdf for a sample $\mathbf{y} = (y_1, \ldots, y_T)$ is

$$f(\mathbf{y};\boldsymbol{\theta}) = f(y_1, \ldots, y_T;\boldsymbol{\theta}) = \prod_{t=1}^{T} f(y_t;\boldsymbol{\theta})$$

The likelihood function is this joint density treated as a function of the parameters $\boldsymbol{\theta}$ given the data $\mathbf{y}$:

$$L(\boldsymbol{\theta}|\mathbf{y}) = L(\boldsymbol{\theta}|y_1, \ldots, y_T) = \prod_{t=1}^{T} f(y_t;\boldsymbol{\theta})$$

The log-likelihood is

$$\ln L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{t=1}^{T} \ln f(y_t;\boldsymbol{\theta})$$

Problem: For a sample from a covariance stationary time series $\{y_t\}$, the construction of the log-likelihood give above doesn't work because the random variables in the sample $(y_1, \ldots, y_T)$ are not iid.

One Solution: Conditional factorization of log-likelihood

Intiution: Consider the joint density of two adjacent observations $f(y_2, y_1;\boldsymbol{\theta})$. The joint density can always be factored as the product of the conditional density of y_2 given y_1 and the marginal density of y_1 :

$$f(y_2, y_1;\boldsymbol{\theta}) = f(y_2|y_1;\boldsymbol{\theta})f(y_1;\boldsymbol{\theta})$$

For three observations, the factorization becomes

$$f(y_3, y_2, y_1;\boldsymbol{\theta}) = f(y_3|y_2, y_1;\boldsymbol{\theta})f(y_2|y_1;\boldsymbol{\theta})f(y_1;\boldsymbol{\theta})$$

In general, the conditional marginal factorization has the form

$$
\begin{aligned}
f(y_T, \ldots, y_1; \boldsymbol{\theta}) &= \left(\prod_{t=p+1}^{T} f(y_t | I_{t-1}, \boldsymbol{\theta}) \right) \cdot f(y_p, \ldots, y_1; \boldsymbol{\theta}) \\
I_t &= \{y_t, \ldots, y_1\} = \text{ info availble at time } t \\
y_p, \ldots, y_1 &= \text{ initial values}
\end{aligned}
$$

The *exact* log-likelihood function may then be expressed as

$$
\ln L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{t=p+1}^{T} \ln f(y_t | I_{t-1}, \boldsymbol{\theta}) + \ln f(y_p, \ldots, y_1; \boldsymbol{\theta})
$$

The *conditional* log-likelihood is

$$
\ln L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{t=p+1}^{T} \ln f(y_t | I_{t-1}, \boldsymbol{\theta})
$$

Two types of maximum likelihood estimates (mles) may be computed. The first type is based on maximizing the conditional log-likelihood function. These estimates are called conditional mles and are defined by

$$
\hat{\boldsymbol{\theta}}_{cmle} = \arg\max_{\theta} \sum_{t=p+1}^{T} \ln f(y_t | I_{t-1}, \boldsymbol{\theta})
$$

The second type is based on maximizing the exact log-likelihood function. These estimates are called exact mles, and are defined by

$$
\hat{\boldsymbol{\theta}}_{mle} = \arg\max_{\theta} \sum_{t=p+1}^{T} \ln f(y_t | I_{t-1}, \boldsymbol{\theta}) + \ln f(y_p, \ldots, y_1; \boldsymbol{\theta})
$$

Result: For stationary models, $\hat{\boldsymbol{\theta}}_{cmle}$ and $\hat{\boldsymbol{\theta}}_{mle}$ are consistent and have the same limiting normal distribution. In finite samples, however, $\hat{\boldsymbol{\theta}}_{cmle}$ and $\hat{\boldsymbol{\theta}}_{mle}$ are generally not equal and my differ by a substantial amount if the data are close to being non-stationary or non-invertible.

Example: MLE for stationary AR(1)

$$\begin{aligned} y_t &= c + \phi y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim iid\ N(0, \sigma^2),\ t = 1, \ldots, T \\ \boldsymbol{\theta} &= (c, \phi, \sigma^2)',\ |\phi| < 1 \end{aligned}$$

Conditional on I_{t-1}

$$y_t | I_{t-1} \sim N(c + \phi y_{t-1}, \sigma^2),\ t = 2, \ldots, T$$

which only depends on y_{t-1}. The conditional density $f(y_t|I_{t-1}, \theta)$ is then

$$\begin{aligned} f(y_t|y_{t-1}, \boldsymbol{\theta}) &= (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y_t - c - \phi y_{t-1})^2\right), \\ t &= 2, \ldots, T \end{aligned}$$

To determine the marginal density for the initial value y_1, recall that for a stationary AR(1) process

$$\begin{aligned} E[y_1] &= \mu = \frac{c}{1-\phi} \\ var(y_1) &= \frac{\sigma^2}{1-\phi^2} \end{aligned}$$

It follows that

$$y_1 \sim N\left(\frac{c}{1-\phi}, \frac{\sigma^2}{1-\phi^2}\right)$$

$$f(y_1; \boldsymbol{\theta}) = \left(\frac{2\pi\sigma^2}{1-\phi^2}\right)^{-1/2} \exp\left(-\frac{1-\phi^2}{2\sigma^2}\left(y_1 - \frac{c}{1-\phi}\right)^2\right)$$

The conditional log-likelihood function is

$$\begin{aligned}\sum_{t=2}^{T} \ln f(y_t|y_{t-1}, \boldsymbol{\theta}) &= \frac{-(T-1)}{2}\ln(2\pi) - \frac{(T-1)}{2}\ln(\sigma^2) \\ &\quad -\frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2\end{aligned}$$

Notice that the conditional log-likelihood function has the form of the log-likelihood function for a linear regression model with normal errors

$$\begin{aligned} y_t &= c + \phi y_{t-1} + \varepsilon_t, t = 2, \ldots, T \\ &\quad \varepsilon_t \sim \text{iid } N(0, \sigma^2)\end{aligned}$$

It follows that

$$\begin{aligned}\hat{c}_{cmle} &= \hat{c}_{ols} \\ \hat{\phi}_{cmle} &= \hat{\phi}_{ols} \\ \hat{\sigma}^2_{cmle} &= (T-1)^{-1}\sum_{t=2}^{T}(y_t - \hat{c}_{cmle} - \hat{\phi}_{cmle} y_{t-1})^2\end{aligned}$$

The marginal log-likelihood for the initial value y_1 is

$$\begin{aligned}\ln f(y_1; \boldsymbol{\theta}) &= -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\frac{\sigma^2}{1-\phi^2}\right) \\ &\quad -\frac{1-\phi^2}{2\sigma^2}\left(y_1 - \frac{c}{1-\phi}\right)^2\end{aligned}$$

The exact log-likelihood function is then

$$\begin{aligned}\ln L(\boldsymbol{\theta}|\mathbf{y}) &= -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\frac{\sigma^2}{1-\phi^2}\right) \\ &\quad -\frac{1-\phi^2}{2\sigma^2}\left(y_1 - \frac{c}{1-\phi}\right)^2 \\ &\quad -\frac{(T-1)}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2\end{aligned}$$

Remarks

1. The exact log-likelihood function is a non-linear function of the parameters θ, and so there is no closed form solution for the exact mles.

2. The exact mles must be determined by numerically maximizing the exact log-likelihood function. Usually, a Newton-Raphson type algorithm is used for the maximization which leads to the interative scheme

$$\hat{\boldsymbol{\theta}}_{mle,n} = \hat{\boldsymbol{\theta}}_{mle,n-1} - \hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_{mle,n-1})^{-1}\hat{\mathrm{s}}(\hat{\boldsymbol{\theta}}_{mle,n-1})$$

where $\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}})$ is an estimate of the Hessian matrix (2nd derivative of the log-likelihood function), and $\hat{\mathrm{s}}(\hat{\boldsymbol{\theta}})$ is an estimate of the score vector (1st derivative of the log-likelihood function).

3. The estimates of the Hessian and Score may be computed numerically (using numerical derivative routines) or they may be computed analytically (if analytic derivatives are known).

Prediction Error Decomposition of Log-Likelihood

To illustrate this algorithm, consider the simple AR(1) model. Recall,

$$y_t|I_{t-1} \sim N(c + \phi y_{t-1}, \sigma^2),\ t = 2, \ldots, T$$

from which it follows that

$$\begin{aligned} E[y_t|I_{t-1}] &= c + \phi y_{t-1} \\ var(y_t|I_{t-1}) &= \sigma^2 \end{aligned}$$

The 1-step ahead prediction errors may then be defined as

$$v_t = y_t - E[y_t|I_{t-1}] = y_t - c + \phi y_{t-1},\ t = 2, \ldots T$$

The variance of the prediction error at time t is

$$f_t = var(v_t) = var(\varepsilon_t) = \sigma^2,\ t = 2, \ldots T$$

For the initial value, the first prediction error and its variance are

$$\begin{aligned} v_1 &= y_1 - E[y_1] = y_1 - \frac{c}{1-\phi} \\ f_1 &= var(v_1) = \frac{\sigma^2}{1-\phi^2} \end{aligned}$$

Using the prediction errors and the prediction error variances, the exact log-likelihood function may be re-expressed as

$$\ln L(\boldsymbol{\theta}|\mathbf{y}) = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\ln f_t - \frac{1}{2}\sum_{t=1}^{T}\frac{v_t^2}{f_t}$$

which is the prediction error decomposition.

Remarks

1. A further simplification may be achieved by writing

$$\begin{aligned} var(v_t) &= \sigma^2 f_t^* \\ &= \sigma^2 \cdot \frac{1}{1-\phi^2} \text{ for } t = 1 \\ &= \sigma^2 \cdot 1 \text{ for } t > 1 \end{aligned}$$

That is $f_t^* = 1/(1-\phi^2)$ for $t = 1$ and $f_t^* = 1$ for $t > 1$. Then the log-likelihood becomes

$$\ln L(\boldsymbol{\theta}|\mathbf{y}) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2}\sum_{t=1}^{T}\ln f_t^* - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\frac{v_t^2}{f_t^*}$$

2. With above simplification, σ^2 may be concentrated out of the log-likelihood. That is,

$$\frac{\partial \ln L(\boldsymbol{\theta}|\mathbf{y})}{\partial\sigma^2} = 0 \Rightarrow \hat{\sigma}^2_{mle}(c,\phi) = \frac{1}{T}\sum_{t=1}^{T}\frac{v_t^2}{f_t^*}$$

Substituting $\hat{\sigma}^2_{mle}(c,\phi)$ back into $\ln L(\boldsymbol{\theta}|\mathbf{y})$ gives the concentrated log-likelihood

$$\ln L^c(c,\phi|\mathbf{y}) = -\frac{T}{2}\ln(2\pi+1) - \frac{T}{2}\ln\hat{\sigma}^2_{mle}(c,\phi) - \frac{1}{2}\sum_{t=1}^{T}\ln f_t^*$$

Maximizing $\ln L^c(c,\phi|\mathbf{y})$ gives the mles for c and ϕ. Maximizing $\ln L^c(c,\phi|\mathbf{y})$ is faster than maximizing $\ln L(\boldsymbol{\theta}|\mathbf{y})$ and is more numerically stable.

3. For general time series models, the prediction error decomposition may be conveniently computed as a by product of the *Kalman filter algorithm* if the time series model can be cast in *state space form*.

Diagnostics of Fitted ARMA Model

1. Compare theoretical ACF, PACF with SACF, SPACF

2. Examine autocorrelation properties of residuals

SACF, SPACF of residuals

LM test for serial correlation in residuals

Remarks:

1. If the Box-Ljung Q^*-statistic is used on the residuals from a fitted ARMA(p,q) model, the degrees of freedom of the limiting chi-square distribution must be adjusted for the number of estimated parameters:

$$Q^*(k) \xrightarrow{d} \chi^2_{k-(p+q)}$$

Note that this test is only valid for $k > (p+q)$. In finite samples, it may perform badly if k is close to $p+q$.