



Università di Pavia

Box & Jenkins Model Selection

Eduardo Rossi



Complicate models can perform very well over the historical period for which parameters are estimated however they can have poor out-of-sample forecasting performances.

1. Transform the data to induce stationarity
2. Make an initial guess for p and q (small values) for an ARMA(p,q) or ARIMA($p,1,q$).
3. Estimate the parameters of $\phi(L)$ and $\theta(L)$.
4. Perform diagnostic analysis to confirm that the model is consistent with the observed features of data.



1. **Stationarity**: Compute and examine the sample ACF and the sample PACF of the original series to further confirm a necessary degree of differencing.

If the sample ACF decays very slowly and the sample PACF cuts off after lag 1, it indicates that differencing is needed:

$$\Delta \log Y_t = \log(Y_t) - \log(Y_{t-1})$$

2. **Selection procedures**: Estimate the autocorrelation function:

$$\hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_0}$$

$$\hat{\gamma}_j = \frac{1}{T} \sum_{t=j+1}^T (y_t - \bar{y})(y_{t-j} - \bar{y}) \quad j = 0, 1, \dots, T - 1$$

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



MODEL SELECTION

MA(q): $\rho_j = 0 \quad j > q$

$$Y_t = \epsilon_t + 0.8\epsilon_{t-1}$$

$$\rho_1 = \frac{0.8}{(1 + 0.8)^2} = \frac{0.8}{1.64} \cong 0.48$$

AR(p): ρ_j gradually decays toward zero as a mixture of exponentials or dumped sinusoids.

We identify the orders p and q by matching the patterns in the sample ACF and PACF with the theoretical patterns of known models.

We are interested in assessing if

$$\rho_j = 0 \quad \text{for } j = q + 1, q + 2, \dots$$



If the data were really generated by a Gaussian MA(q) process, then the variance of the estimate $\hat{\rho}_j$ could be approximated by

$$\text{Var}(\hat{\rho}_j) \cong \frac{1}{T} \left\{ 1 + 2 \sum_{i=1}^q \rho_i^2 \right\} \quad j = q + 1, q + 2, \dots$$

If the data were generated by a Gaussian White Noise, then $\hat{\rho}_j$ for any $j \neq 0$ should lie between $\pm \frac{2}{\sqrt{T}}$ about 95% of the time.

If $|\hat{\rho}_1| > 2/\sqrt{T}$ reject the null, i.e., MA(0) process.



In general if there is autocorrelation in the process that generated the data $\{Y_t\}$, then the estimate of $\hat{\rho}_j$ will be correlated with $\hat{\rho}_i$ for $i \neq j$. Thus patterns in the estimated $\hat{\rho}_j$ may represent sampling error rather than patterns in the true ρ_j .



MEAN SQUARED ERROR

Forecast of Y_{t+1} based on a set of variables observed at date t , \mathbf{X}_t :
 $Y_{t+1|t}^*$. The loss function

$$MSE(Y_{t+1|t}^*) = E[Y_{t+1} - Y_{t+1|t}^*]^2$$

The forecast with the smallest MSE is

$$Y_{t+1|t}^* = E[Y_{t+1} | \mathbf{X}_t]$$

Suppose $Y_{t+1|t}^*$ is a linear function of \mathbf{X}_t :

$$\hat{Y}_{t+1|t} = \boldsymbol{\alpha}' \mathbf{X}_t$$

if

$$E[(Y_{t+1} - \boldsymbol{\alpha}' \mathbf{X}_t) \mathbf{X}_t'] = \mathbf{0}'$$

then $\boldsymbol{\alpha}' \mathbf{X}_t$ is the **linear projection** of Y_{t+1} on \mathbf{X}_t .



The LP projection produces the smallest MSE among the class of *linear* forecasting rule

$$\widehat{P}(Y_{t+1}|\mathbf{X}_t) = \boldsymbol{\alpha}'\mathbf{X}_t$$

$$MSE[\widehat{P}(Y_{t+1}|\mathbf{X}_t)] \geq MSE[E(Y_{t+1}|\mathbf{X}_t)]$$

using

$$E[(Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t)\mathbf{X}_t'] = \mathbf{0}'$$

$$E[Y_{t+1}\mathbf{X}_t'] = \boldsymbol{\alpha}'E[\mathbf{X}_t\mathbf{X}_t']$$

$$\boldsymbol{\alpha}' = E[Y_{t+1}\mathbf{X}_t']E[\mathbf{X}_t\mathbf{X}_t']^{-1}$$



LP is closely related to OLS regression

$$y_{t+1} = \beta' \mathbf{X}_t + u_t$$

$$\hat{\beta} = \left[\frac{1}{T} \sum_{t=1}^T \mathbf{X}_t \mathbf{X}_t' \right]^{-1} \left[\frac{1}{T} \sum_{t=1}^T \mathbf{X}_t Y_{t+1} \right]$$

$\hat{\beta}$ is constructed from the sample moments, while α is constructed from population moments.

If $\{\mathbf{X}_t, Y_{t+1}\}$ is covariance stationary and ergodic for second moments, then the sample moments will converge to the population moments as the sample size T goes to infinity

$$\frac{1}{T} \sum_{t=1}^T \mathbf{X}_t \mathbf{X}_t' \xrightarrow{p} E[\mathbf{X}_t \mathbf{X}_t']$$

$$\frac{1}{T} \sum_{t=1}^T \mathbf{X}_t y_{t+1} \xrightarrow{p} E[\mathbf{X}_t Y_{t+1}]$$



PARTIAL AUTOCORRELATION FUNCTION

In the AR(1) process, y_t and y_{t-2} are correlated even though y_{t-2} does not appear in the model.

The correlation is $\rho(2) = \rho(1)^2$.

Partial Autocorrelation between y_t and y_{t-s} eliminates the effects of the intervening values y_{t-1} through y_{t-s+1} . The PACF in an AR(1) between y_t and y_{t-2} is equal to zero.



PARTIAL AUTOCORRELATION FUNCTION

$\alpha_m^{(m)}$ \equiv the m -th population partial autocorrelation

is defined as the last coefficient in a linear projection of Y on its most recent values

$$\widehat{Y}_{t+1|t} - \mu = \alpha_1^{(m)}(Y_t - \mu) + \alpha_2^{(m)}(Y_{t-1} - \mu) + \dots + \alpha_m^{(m)}(Y_{t-m+1} - \mu)$$

If the data were really generated by an AR(p) process, only the p most recent values of Y would be useful for forecasting

$$\alpha_m^{(m)} = 0 \quad m = p + 1, p + 2, \dots$$



PARTIAL AUTOCORRELATION FUNCTION

If the data were really generated by an MA(p) process, with $q \geq 1$, $\alpha_m^{(m)}$ asymptotically approaches 0.

A natural estimate of $\alpha_m^{(m)}$ is the OLS estimate $\hat{\alpha}_m^{(m)}$ from an OLS regression of Y on a constant and its m most recent values

$$y_{t+1} = \hat{c} + \hat{\alpha}_1^{(m)} y_t + \hat{\alpha}_2^{(m)} y_{t-1} + \dots + \hat{\alpha}_m^{(m)} y_{t-m+1} + \hat{\epsilon}_{t+1}$$

If the data were really generated by an AR(p) (i.e., $H_0 : \alpha_m^{(m)} = 0, m = p + 1, p + 2, \dots$) then $\hat{\alpha}_m^{(m)}$ would have a variance around the true value (0) that could be approximated by

$$\text{Var}(\hat{\alpha}_m^{(m)}) \cong \frac{1}{T} \quad m = p + 1, p + 2, \dots$$

If the data were really generated by AR(p) process $\alpha_i^{(i)}$ and $\alpha_j^{(j)}$ would be asymptotically independent for $i, j > p$.



PARTIAL AUTOCORRELATION FUNCTION

One can form the PACFs from the autocorrelations as:

$$\begin{aligned}\alpha_1^{(1)} &= \rho(1) \\ \alpha_2^{(2)} &= (\rho(2) - \rho(1)^2) / (1 - \rho(1)^2) \\ \alpha_m^{(m)} &= \frac{\rho(m) - \sum_{j=1}^{m-1} \alpha_j^{(m-1)} \rho(m-j)}{1 - \sum_{j=1}^{m-1} \alpha_j^{(m-1)} \rho(j)} \quad m = 3, 4, 5, \dots \\ \alpha_j^{(m)} &= \alpha_j^{(m-1)} - \alpha_m^{(m)} \alpha_{m-j}^{(m-1)} \quad j = 1, 2, \dots, m-1\end{aligned}$$



MODEL SELECTION CRITERIA

There may be several adequate models that can be used to represent a given data set. Numerous criteria for model comparison have been introduced in the literature for model selection. To avoid the problem of overfitting we impose a cost for increasing the number of parameters in the fitted model.



3. Model selection criteria based on residuals

Assume that a statistical model of M parameters is fitted to data.

To assess the quality of the model fitting, Akaike (1973,1974) introduced an information criterion, *Akaike's Information Criterion*

$$AIC(M) = -2\mathcal{L}(\hat{c}, \hat{\phi}, \hat{\theta}, \hat{\sigma}_\epsilon^2) + 2M$$



For the ARMA model with T effective number of obs, the log-likelihood function:

$$\mathcal{L} = -\frac{T}{2} \log(2\pi\sigma_\epsilon^2) - \frac{S(c, \phi, \theta)}{2\sigma_\epsilon^2}$$

where $\phi = \phi_1, \dots, \phi_p$ and $\theta = \theta_1, \dots, \theta_q$ and $S(c, \phi, \theta)$ is the residual sum of squares.



Maximizing the log-likelihood function with respect to the parameters $c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma_\epsilon^2$

$$\mathcal{L} = -\frac{T}{2} \log(\hat{\sigma}_\epsilon^2) - \frac{T}{2}(1 + \log(2\pi))$$

$$AIC(M) = T \log(\hat{\sigma}_\epsilon^2) + 2M$$

The optimal order is chosen by the value of M , which is a function of p and q , so that $AIC(M)$ is minimum.



Schwartz (1978) suggested the *Schwartz's Information Criterion*

$$SIC(M) = -2\mathcal{L} + M \log(T)$$

$$SIC(M) = T \log \hat{\sigma}_\epsilon^2 + M \log(T)$$



4. Diagnostic Analysis.

The basic assumption is that $\{\epsilon_t\}$ are white noise. **The Box-Pierce Portmanteau Test** (1970) for Residual Autocorrelation (Q test), uses the residual sample ACF's

$$r_j = E[\epsilon_t \epsilon_{t-j}]$$

$$H_0 : r_j = 0 \quad j = 1, \dots, m$$

$$Q = T \sum_{j=1}^m \hat{r}_j^2 \xrightarrow{d} \chi_\nu^2$$

$\nu = m - k$, k is the number of parameters estimated in the model. \hat{r}_j sample autocorrelation of residuals.



Ljung-Box Test (1978) (small-sample correction)

$$Q = T(T + 2) \sum_{j=1}^m \frac{\widehat{r}_j^2}{T - j} \xrightarrow{d} \chi_\nu^2$$

$$\nu = m - p - q.$$