

## Lecture 7: Model Building

Bus 41910, Time Series Analysis, Mr. R. Tsay

An effective procedure for building empirical time series models is the Box-Jenkins approach, which consists of three stages: model specification, estimation and diagnostics checking. These three stages are used iteratively until an appropriate model is found. The estimation is accomplished by using mainly the maximum likelihood method. For model checking, there are various methods available in the literature, and we shall discuss some of those methods later. For now, we shall focus on model specification.

Model specification (or identification) is intended to specify, from the data, certain tentative models which are worth a careful investigation. For simplicity, we focus on the class of ARIMA models. However, the three-stage modeling procedure applies equally well to other models. For ARIMA models, there are two main approaches to model specification. The first approach is called the “correlation” approach in which the tentative models are selected via the examination of certain (sample) correlation functions. This approach does not require “full estimation” of any model. However, it is judgemental in the sense that a data analyst must make a decision regarding which models to entertain. The second approach is called the information criterion approach in which an objective function is defined and the model selection is done automatically by evaluating the objective function of possible models. Usually, the model which achieves the minimum of the criterion function is treated as the “most appropriate” model for the data. The evaluation of the criterion function for a given model, however, requires formal estimation of the model.

Suppose that the observed realization is  $\{Z_1, Z_2, \dots, Z_n\}$ . In some cases, certain transformation of  $Z_t$  is needed before model building, e.g. variance stabilization. Thus, one should always plot the data before considering model specification. In what follows, we shall briefly discuss the two model-specification approaches.

A. Correlation approach: The basic tools used in this approach of model specification include (a) sample autocorrelation function (ACF), (b) sample partial autocorrelation function (PACF), (c) extended autocorrelation function (EACF) and (d) the method of smallest canonical correlation (SCAN). The function of these tools can be summarized as

Function	Model	Feature
ACF	MA( $q$ )	Cutting-off at lag $q$
PACF	AR( $p$ )	Cutting-off at lag $p$
EACF	ARMA( $p, q$ )	A triangle with vertex $(p, q)$
SCAN	ARMA( $p, q$ )	A rectangle with vertex $(p, q)$

Illustration: (Some simulated examples are informative).

a. ACF: The lag- $\ell$  sample ACF of  $Z_t$  is defined by

$$\hat{\rho}_\ell = \frac{\sum_{t=\ell+1}^n (Z_t - \bar{Z})(Z_{t-\ell} - \bar{Z})}{\sum_{t=1}^n (Z_t - \bar{Z})^2}$$

where  $\bar{Z} = \frac{1}{n} \sum_{t=1}^n Z_t$  is the sample mean. In the literature, you may see some minor deviation from this definition. However, the above one is close to being a standard. Two main features of sample ACF are particularly useful in model specification. First of all, for a stationary ARMA model,

$$\hat{\rho}_\ell \rightarrow_p \rho_\ell, \quad \text{as } n \rightarrow \infty$$

where  $\rightarrow_p$  denotes convergence in probability. Also,  $\hat{\rho}_\ell$  is asymptotically normal with mean  $\rho_\ell$  and variance being function of the ACF  $\rho_i$ 's. (See Box and Jenkins (1976) and the references therein. Or page 21 of Wei (1990)). Recall that for an MA( $q$ ) process, we have

$$\rho_\ell \begin{cases} \neq 0 & \text{for } \ell = q \\ = 0 & \text{for } \ell > q. \end{cases}$$

Therefore, for moderate and large samples, the sample ACF of an MA( $q$ ) process would show this cutting-off property. In other words, if

$$\hat{\rho}_q \neq 0, \quad \text{but } \hat{\rho}_\ell := 0 \quad \text{for } \ell > q,$$

then the process is likely to follow an MA( $q$ ) model. Here  $:=$  and  $\neq$  denote, respectively, statistically equal to and different from. To judge the significance of sample ACF, we use its asymptotic variance under certain null-hypothesis. It can be shown that for an MA( $q$ ) process, the asymptotic variance of  $\hat{\rho}_\ell$  for  $\ell > q$  is

$$\text{Var}[\hat{\rho}_\ell] = \frac{1 + 2(\rho_1^2 + \dots + \rho_q^2)}{n}.$$

This is referred to as the Bartlett's formula in the literature. See Chapter 6, page 177, of Box and Jenkins (1976). In practice, the  $\rho_i$ 's are estimated by  $\hat{\rho}_i$ 's. In particular, if  $Z_t$  is a white noise process, then  $\text{Var}[\hat{\rho}_\ell] = 1/n$  for all  $\ell > 0$ . See the SCA output of ACF.

The second important feature of sample ACF is that for any ARIMA( $p, d, q$ ) model with  $d > 0$ ,

$$\hat{\rho}_\ell \rightarrow_p 1 \quad \text{as } n \rightarrow \infty.$$

This says that the sample ACF is persistent for any ARIMA( $p, d, q$ ) model. In practice, persistent sample ACF is often regarded as an indication of non-stationarity and differencing is used to render the series stationary. See SCA output on differencing.

b. PACF: Recall that ACF of an ARMA( $p, q$ ) model satisfies  $\phi(B)\rho_\ell = 0$  for  $\ell > q$ . In particular, for AR models, the ACF satisfies the difference equation  $\phi(B)X = 0$ , implying that the ACF has infinite non-zero lags and tends to be damped sine (co-sine) function or exponentials. Thus, sample ACF is not particularly useful in specifying pure AR models.

On the other hand, recall that the Yule-Walker equation of an  $AR(p)$  process can be used to obtain the AR coefficients from the ACF. Obviously, for an  $AR(p)$  model, all the AR-coefficients of order higher than  $p$  are zero. Consequently, by examining the estimates of AR coefficients, one can identify the order of an AR process. The  $p$ -th order Yule-walker equation is

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-2} & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-3} & \rho_{p-2} \\ \vdots & & & & & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix}.$$

By the Cramer rule, we have

$$\phi_p = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{2-p} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{3-p} & \rho_2 \\ \vdots & & & & & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & \rho_1 & \rho_p \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-2} & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-3} & \rho_{p-2} \\ \vdots & & & & & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & \rho_1 & 1 \end{vmatrix}}. \quad (1)$$

Let  $\hat{\phi}_{p,p}$  be the estimate of  $\phi_p$  obtained via equation (1) with  $\rho_\ell$  replaced by its sample counterpart  $\hat{\rho}_\ell$ . The function

$$\hat{\phi}_{1,1}, \quad \hat{\phi}_{2,2}, \quad \cdots, \quad \hat{\phi}_{\ell,\ell}, \quad \cdots$$

is called the sample PACF of  $Z_t$ . Based on previous discussion, for an  $AR(p)$  process, we have

$$\hat{\phi}_{p,p} \neq 0, \quad \text{but} \quad \hat{\phi}_{\ell,\ell} := 0 \quad \text{for} \quad \ell > p.$$

This is the cutting-off property of sample PACF by which the order of an AR process can be specified.

Alternatively, the sample PACF  $\hat{\phi}_{\ell,\ell}$  can be defined as the least squares estimates of the following consecutive autoregressions:

$$\begin{aligned} Z_t &= \phi_{1,0} + \phi_{1,1}Z_{t-1} + e_{1t} \\ Z_t &= \phi_{2,0} + \phi_{2,1}Z_{t-1} + \phi_{2,2}Z_{t-2} + e_{2t} \\ Z_t &= \phi_{3,0} + \phi_{3,1}Z_{t-1} + \phi_{3,2}Z_{t-2} + \phi_{3,3}Z_{t-3} + e_{3t} \\ &\vdots = \vdots \end{aligned}$$

This later explanation is more intuitive. It also works better when the process  $Z_t$  is an  $ARIMA(p, d, q)$  process. The first definition of sample PACF via sample ACF is not well-defined in the case of  $ARIMA$  processes. The two definitions, of course, are the same in theory when the series  $Z_t$  is stationary.

In practice, it can be shown that for an AR( $p$ ) process, the asymptotic variance of the sample PACF  $\hat{\phi}_{\ell,\ell}$  is  $\frac{1}{n}$  for  $\ell > p$ . See SCA output.

c. EACF. The model specification of mixed ARMA model is much more complicated than that of pure AR or MA models. We shall consider two methods. The first method to identify the order of a mixed model is the extended autocorrelation function (EACF) of Tsay and Tiao (1984, JASA). [A copy of the paper is in the packet.] The EACF, in fact, applies to ARIMA as well as ARMA models. However, it treats an ARIMA( $p, d, q$ ) model as an ARMA( $p + d, q$ ) model.

The basic idea of EACF is based on the “generalized” Yule-Walker equation. Conceptually, it involves two steps. In the first step, we attempt to obtain consistent estimates of AR coefficients. Given such estimates, we can transform the ARMA series into a pure MA process. The second step then uses the sample ACF of the transformed MA process to identify the MA order  $q$ .

The best way to introduce EACF is to consider some simple examples.

Example 1: Suppose that  $Z_t$  is an ARMA(1,1) model

$$Z_t - \phi Z_{t-1} = a_t - \theta a_{t-1}, \quad |\phi| < 1, \quad |\theta| < 1.$$

For this model, the ACF is

$$\rho_\ell = \begin{cases} \frac{(1-\phi\theta)(\phi-\theta)}{1+\theta^2-2\phi\theta} & \text{for } \ell = 1 \\ \phi\rho_{\ell-1} & \text{for } \ell > 1. \end{cases}$$

For  $p = 1$ , the usual Yule-Walker equation is

$$\rho_1 = \phi\rho_0,$$

and the  $j$ -th generalized Yule-Walker equation is

$$\rho_{j+1} = \phi\rho_j.$$

Denote the solution of the Yule-Walker equation by  $\phi_{1,1} = \phi_{1,1}^{(0)}$  and that of the  $j$ -th generalized Yule-Walker equation by  $\phi_{1,1}^{(j)}$ . Then, we have

$$\phi_{1,1}^{(j)} = \begin{cases} \rho_1 \neq \phi & \text{for } j = 0 \\ \phi & \text{for } j > 0 \end{cases}$$

Thus, the solution of the usual Yule-Walker equation is not consistent with the AR coefficient  $\phi$ . However, **ALL** of the solutions of the  $j$ -th generalized Yule-Walker equations are consistent with the AR coefficient. In sample, these results say that the estimates of  $\phi_{1,1}^{(j)}$  obtained by replacing the ACF by sample ACF have the property:

$$\hat{\phi}_{1,1}^{(j)} \xrightarrow{p} \begin{cases} \rho_1 & \text{for } j = 0 \\ \phi & \text{for } j > 0. \end{cases}$$

Now define the transformed series  $W_{1,t}^{(j)}$  by

$$W_{1,t}^{(j)} = Z_t - \hat{\phi}_{1,1}^{(j)} Z_{t-1} \quad \text{for } j > 0.$$

The above discussion shows that  $W_{1,t}^{(j)}$  for  $j > 0$  is asymptotically a pure MA(1) process. Consequently, by considering the ACF of the  $W_{1,t}^{(j)}$  series, we can identify that the MA order is 1.

Example 2: Suppose now that  $Z_t$  is a stationary and invertible ARMA(1,2) process

$$Z_t - \phi Z_{t-1} = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}.$$

The ACF of  $Z_t$  satisfies

$$\rho_\ell \begin{cases} \neq \phi \rho_1 & \text{for } \ell = 2 \\ = \phi \rho_{\ell-1} & \text{for } \ell > 2 \end{cases}$$

Using this result and considering the solution of the  $j$ -th generalized Yule-Walker equation of order 1

$$\rho_{j+1} = \phi \rho_j,$$

we see that

$$\phi_{1,1}^{(j)} \begin{cases} \neq \phi & \text{for } j \leq 2 \\ = \phi & \text{for } j > 2 \end{cases}$$

Therefore, the  $j$ -th transformed series

$$W_{1,t}^{(j)} = Z_t - \phi_{1,1}^{(j)} Z_{t-1}$$

is an MA(2) series provided that  $j > 2$ .

Compared with the result of Example 1, we see that the difference between ARMA(1,1) and ARMA(1,2) is that we NEED to consider one step further in the generalized Yule-Walker equation. In either case, however, the ACF of the transformed series can suggest the MA order  $q$  once a consistent AR coefficient is used.

In general, the above two simple examples show that for an ARMA(1, $q$ ) model, the  $j$ -th generalized Yule-Walker equation provides consistent AR estimate if  $j > q$ . Thus, the  $j$ -th transform series  $W_{1,t}^{(j)} = Z_t - \phi_{1,1}^{(j)} Z_{t-1}$  is an MA( $q$ ) series for  $j > q$ . In practice, it would be cumbersome to consider ACF of all the transformed series  $W_{1,t}^{(j)}$  for  $j = 1, 2, \dots$ . We are thus led to consider a summary of the ACF. The EACF is a device which is designed to summarize the pattern of ACF of  $W_{1,t}^{(j)}$  for all  $j$ .

**First-order extended ACF:** The first-order extended ACF is defined as

$$\rho_j(1) = \rho_\ell \quad \text{of } W_{1,t}^{(j)}$$

where

$$W_{1,t}^{(j)} = Z_t - \phi_{1,1}^{(j)} Z_{t-1}, \quad \text{with } \phi_{1,1}^{(j)} = \frac{\rho_{j+1}}{\rho_j}, \quad j \geq 0.$$

It is easy to check that for an ARMA(1,q) process, we have

$$\rho_{1,j} \begin{cases} \neq 0 & \text{for } j \leq q \\ = 0 & \text{for } j > q. \end{cases}$$

In summary, the first-order extended autocorrelation function is designed to identify the order of ARMA(1,q) model. It function in an exact manner as that of ACF to an MA model.

Similarly, we can define a 2nd-order EACF to identify the order of an ARMA(2,q) model,

$$Z_t - \phi_1 Z_{t-1} - \phi_2 Z_{t-2} = c + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q}.$$

More specifically, the  $j$ -th generalized Yule-Walker equation of order 2 is defined by

$$\begin{bmatrix} \rho_{j+1} \\ \rho_{j+2} \end{bmatrix} = \begin{bmatrix} \rho_j & \rho_{j-1} \\ \rho_{j+1} & \rho_j \end{bmatrix} \begin{bmatrix} \phi_{2,1}^{(j)} \\ \phi_{2,2}^{(j)} \end{bmatrix}.$$

Obviously, the solution of this equation satisfies

$$\phi_{2,i}^{(j)} = \phi_i \quad i = 1, 2; \quad \text{for } j > q.$$

Define the 2nd-order EACF by

$$\rho_{2,j} = \rho_j \quad \text{of the transformed series } W_{2,t}^{(j)}$$

where

$$W_{2,t}^{(j)} = Z_t - \phi_{2,1}^{(j)} Z_{t-1} - \phi_{2,2}^{(j)} Z_{t-2}.$$

It is clear from the above discussion that

$$\rho_j(2) \begin{cases} \neq 0 & \text{for } j = q \\ = 0 & \text{for } j > q. \end{cases}$$

Here, of course,  $Z_t$  is an ARMA(2,q) process.

You should be able to generalize the EACF to the general ARMA( $p, q$ ) case. (Exercise!)

Model Specification via EACF. To make use of the EACF for model specification, we consider the two-way table:

AR	MA (or $j$ )					
$m$	0	1	2	3	4	...
0	$\rho_1$	$\rho_2$	$\rho_3$	$\rho_4$	$\rho_5$	...
1	$\rho_{1,1}$	$\rho_{1,2}$	$\rho_{1,3}$	$\rho_{1,4}$	$\rho_{1,5}$	...
2	$\rho_{2,1}$	$\rho_{2,2}$	$\rho_{2,3}$	$\rho_{2,4}$	$\rho_{2,5}$	...
3	$\rho_{3,1}$	$\rho_{3,2}$	$\rho_{3,3}$	$\rho_{3,4}$	$\rho_{3,5}$	...
$\vdots$	$\vdots$			$\vdots$		

The EACF Table

In practice, the EACF in the above table is replaced by its sample counterpart. To identify the order of an ARMA model, we need to understand the behavior of the EACF table for a given model. Before giving the theory, I shall illustrate the function of the table. Suppose that  $Z_t$  is an ARMA(1,1) model, then the corresponding EACF table is

AR	MA (or $j$ )						
$m$	0	1	2	3	4	5	...
0	X	X	X	X	X	X	...
1	X	O	O	O	O	O	...
2	*	X	O	O	O	O	...
3	*	*	X	O	O	O	...
4	*	*	*	X	O	O	...

The EACF Table

where “X” and “O” denotes non-zero and zero quantities, respectively, “\*” represents a quantity which can assume any value between  $-1$  and  $1$ .

From the table, we see that there exists a triangle of “O” with vertex at  $(1, 1)$ , which is the order of  $Z_t$ . In practice, the non-zero and zero terms are determined by the sample EACF and its estimated standard error via the Bartlett’s formula for MA models. Of course, we cannot expect to see an exact triangle as that of the above table. However, one can often make a decision based on the pattern of the EACF table.

To understand the triangular pattern, it is best to consider a simple example such as ARMA(1,1) model of the above table. In particular, we shall discuss the reason why  $\rho_{2,2}$  is different from zero for an ARMA(1,1) model. By definition,  $\rho_{2,2}$  is the lag-2 ACF of the transformed series

$$W_{2,t}^{(2)} = Z_t - \phi_{2,1}^{(2)}Z_{t-1} - \phi_{2,2}^{(2)}Z_{t-2}$$

where  $\phi_{2,1}^{(2)}$  and  $\phi_{2,2}^{(2)}$  are the solution of the 2nd generalized Yule-Walker equation of order 2, namely

$$\begin{bmatrix} \rho_3 \\ \rho_4 \end{bmatrix} = \begin{bmatrix} \rho_2 & \rho_1 \\ \rho_3 & \rho_2 \end{bmatrix} \begin{bmatrix} \phi_{2,1}^{(2)} \\ \phi_{2,2}^{(2)} \end{bmatrix}.$$

However, for an ARMA(1,1) model,  $\rho_j = \phi\rho_{j-1}$  for  $j > 1$  so that the above Yule-Walker equation is “singular” in theory. In practice, the equation is not exactly singular, but is ill-conditioned. Therefore, the solution  $\hat{\phi}_{2,1}^{(2)}$  and  $\hat{\phi}_{2,2}^{(2)}$  can assume any real numbers. Consequently, the chance that  $\phi_{2,i}^{(2)} = 0$  is essential zero. More importantly, this implies that the transformed series  $W_{2,t}^{(2)}$  is not an MA(1) series. Therefore,  $\rho_{2,2} \neq 0$ . Intuitively, one can interpret this result as an over-fitting phenomenon. Since the true model is ARMA(1,1) and we are fitting an AR(2) polynomial in the construction of  $W_{2,t}^{(2)}$ , the non-zero  $\rho_{2,2}$  is in effect a result of overfitting of the second AR coefficient.

Using exactly the same reasoning, one can deduce the triangular pattern of the EACF table. Thus, it can be said that the triangular pattern of EACF is related to the overfitting of AR polynomials in constructing the transformed series  $W_{m,t}^{(j)}$ .

Illustration:

D. SCAN. Next we consider the SCAN method which is closely related to the EACF approach as both methods rely on the generalized moment equations of a time series. However, the SCAN approach utilizes the generalized moment equations in a different way so that it does not encounter the overfitting problem of EACF. In practice, my experience indicates that EACF tends to specify mixed ARMA models whereas SCAN prefers AR type of models.

Although the SCAN approach applies to the non-stationary ARIMA models, we shall only consider the stationary case in this introduction. The moment equations of an ARMA( $p, q$ ) process is

$$\rho_\ell - \phi_1\rho_{\ell-1} - \cdots - \phi_p\rho_{\ell-p} = f(\boldsymbol{\theta}, \boldsymbol{\phi}, \sigma_a^2), \quad \ell \geq 0,$$

where  $f(\cdot)$  is a function of its arguments. In particular, for  $\ell > q$ , we have

$$\rho_\ell - \phi_1\rho_{\ell-1} - \cdots - \phi_p\rho_{\ell-p} = 0. \quad (2)$$

Obviously, Yule-Walker equations and their generalizations are ways to exploit the above moment equation. An alternative approach to make use of the equation (2) is to consider the singularity of the matrices  $\mathbf{A}(m, j)$  for  $m \geq 0$  and  $j \geq 0$ , where

$$\mathbf{A}(m, j) = \begin{vmatrix} \rho_{j+1} & \rho_j & \cdots & \rho_{j+2-m} & \rho_{j+1-m} \\ \rho_{j+2} & \rho_{j+1} & \cdots & \rho_{j+3-m} & \rho_{j+2-m} \\ \vdots & & & & \vdots \\ \rho_{j+1+m} & \rho_{j+m} & \cdots & \rho_{j+2} & \rho_{j+1} \end{vmatrix}_{(m+1) \times (m+1)}.$$

For example, suppose that  $Z_t$  is ARMA(1,1), then

$$\rho_\ell - \phi_1\rho_{\ell-1} = 0 \quad \text{for } \ell > 1.$$

Consequently, by arranging the  $\mathbf{A}(m, j)$  in a two-way table

AR	MA					
	$j$					
$m$	0	1	2	3	4	...
0	$\mathbf{A}(0, 0)$	$\mathbf{A}(0, 1)$	$\mathbf{A}(0, 2)$	$\mathbf{A}(0, 3)$	$\mathbf{A}(0, 4)$	...
1	$\mathbf{A}(1, 0)$	$\mathbf{A}(1, 1)$	$\mathbf{A}(1, 2)$	$\mathbf{A}(1, 3)$	$\mathbf{A}(1, 4)$	...
2	$\mathbf{A}(2, 0)$	$\mathbf{A}(2, 1)$	$\mathbf{A}(2, 2)$	$\mathbf{A}(2, 3)$	$\mathbf{A}(2, 4)$	...
...						

we obtain the pattern

	$j$					
$m$	0	1	2	3	4	...
0	N	N	N	N	N	...
1	N	S	S	S	S	...
2	N	S	S	S	S	...
3	N	S	S	S	S	...
$\vdots$	$\vdots$					

where N and S denote, respectively, singular and non-singular matrix.

From the table, we see that the order (1,1) corresponds exactly to the vertex of a rectangle of singular matrices.

Mathematically, there are many ways to show singularity of a matrix. For instance, one can use determinant or the smallest eigenvalue. An important consideration here is, of course, the statistical properties of the test statistic used to check singularity of a sample matrix. The SCAN approach makes use of the idea of “canonical correlation analysis”, which is a standard technique in multivariate analysis. See, for instance, Anderson (1984). It turns out that there are other advantages in using canonical correlation analysis. For instance, the approach also applies to multivariate time series analysis, see Tiao and Tsay (1989, JRSSB).

For a time series  $Z_t$ , the matrix  $\mathbf{A}(m, j)$  is the covariance matrix between the vectors  $\mathbf{Y}_{m,t} = (Z_t, Z_{t-1}, \dots, Z_{t-m})'$  and  $\mathbf{Y}_{m,t-j-1} = (Z_{t-j-1}, Z_{t-j-2}, \dots, Z_{t-j-1-m})'$ . The singularity of  $\mathbf{A}(m, j)$  means that a linear combination of  $\mathbf{Y}_{m,t}$  is uncorrelated with the vector  $\mathbf{Y}_{m,t-j-1}$ . Thinking in this way, it is then easy to understand the SCAN approach.

Let  $F_t$  denote the information available up to and including  $Z_t$ . In other words,  $F_t$  is the  $\sigma$ -field generated by  $\{Z_t, Z_{t-1}, Z_{t-2}, \dots\}$ . Then, the equation of an ARMA( $p, q$ ) model

$$Z_t - \phi_1 Z_{t-1} - \dots - \phi_p Z_{t-p} = a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q}$$

says, essentially, that the linear combination

$$Z_t - \phi_1 Z_{t-1} - \dots - \phi_p Z_{t-p} \stackrel{def}{=} (1, -\phi_1, -\phi_2, \dots, -\phi_p) \mathbf{Y}_{p,t}$$

is uncorrelated with  $F_{t-j-1}$  for all  $j \geq q$ . Therefore, for an ARMA( $p, q$ ) series, a linear combination of  $\mathbf{Y}_{p,t}$  is uncorrelated with  $\mathbf{Y}_{p,t-j-1}$  for all  $j \geq q$ .

In practice, to test that a linear combination of  $\mathbf{Y}_{m,t}$  is uncorrelated with  $\mathbf{Y}_{m,t-j-1}$ , the SCAN approach uses the test statistic

$$c(m, j) = -(n - m - j) \log\left(1 - \frac{\lambda^2(m, j)}{d(m, j)}\right)$$

where  $n$  is the sample size,  $\lambda^2(m, j)$  is the square of the smallest canonical correlation between  $\mathbf{Y}_{m,t}$  and  $\mathbf{Y}_{m,t-j-1}$  and  $d(m, j)$  is defined by

$$d(m, 0) = 1, \quad d(m, j) = 1 + 2 \sum_{k=1}^j \hat{\rho}_k^2(W), \quad j > 0$$

where  $W_t$  is a transformed series of  $Z_t$  based on the eigenvector of  $\mathbf{A}(m, j)$  corresponding to  $\lambda^2(m, j)$ . This statistic  $c(m, j)$  follows asymptotically a chi-square distribution with 1 degree of freedom for (a)  $m = p$  and  $j \geq q$  or (b)  $m \geq p$  and  $j = q$ . For further details, see Tsay and Tiao (1985, *Biometrika*).

Illustration:

Remark: I assume that most of you have the idea of canonical correlation analysis. If you don't, please consult any textbook of multivariate analysis. For example, Anderson (1984) and Mardia, Kent, and Bibby (1979). Roughly speaking, consider two vector variables  $\mathbf{X}$  and  $\mathbf{Y}$ . Canonical correlation analysis is a technique intended to answer the following questions:

- Q1: Can you find a linear combination of  $\mathbf{X}$ , say  $x_1 = \boldsymbol{\alpha}'_1 \mathbf{X}$ , and a linear combination of  $\mathbf{Y}$ , say  $y_1 = \boldsymbol{\beta}'_1 \mathbf{Y}$ , such that the correlation between  $x_1$  and  $y_1$  is the maximum among all possible linear combinations of  $\mathbf{X}$  and all possible linear combinations of  $\mathbf{Y}$ ?
- Q2: Can you find a linear combination of  $\mathbf{X}$ , say  $x_2 = \boldsymbol{\alpha}'_2 \mathbf{X}$ , which is orthogonal to  $x_1$ , and a linear combination of  $\mathbf{Y}$ , say  $y_2 = \boldsymbol{\beta}'_2 \mathbf{Y}$ , which is orthogonal to  $y_1$ , such that the correlation between  $x_2$  and  $y_2$  is the maximum among all linear combinations of  $\mathbf{X}$  and all linear combinations of  $\mathbf{Y}$  that satisfy the orthogonality condition?

Obviously, one can continue the question until the dimension of  $\mathbf{X}$  or that of  $\mathbf{Y}$  is reached. The solutions of the above questions for  $\mathbf{X}$  turn out to be the eigenvalues and their corresponding eigenvectors of the matrix:

$$[V(\mathbf{X})]^{-1} Cov(\mathbf{X}, \mathbf{Y}) [V(\mathbf{Y})]^{-1} Cov(\mathbf{Y}, \mathbf{X})$$

with the maximum eigenvalue giving rise to the maximum correlation. By interchanging  $\mathbf{X}$  and  $\mathbf{Y}$ , we obtain the linear combinations of  $\mathbf{Y}$ .

We now consider the problem of model selection via information criteria. There are several information criteria proposed in the literature. Basically, they are in the form

$$\text{crit}(m) = -2 \ln(\text{maximized likelihood}) + f(n, m)$$

where  $m$  denotes a model,  $n$  is the sample size, and  $f(n, m)$  is a function of  $n$  and the number of independent parameters in the model  $m$ . Roughly speaking, the first term on the right hand side is a measure of fidelity of the model to the data (or goodness of fit) and the second term is a "penalty function" which penalizes higher dimensional models. Given a set of candidate models, the selection is typically made by choosing the model that minimizes the adopted criterion function among all the models in the set.

Some of the most commonly used criterion functions for selecting ARMA( $p, q$ ) models are

- AIC: Akaike's information criterion (Akaike, 1973)

$$\text{AIC}(p, q) = n \ln(\hat{\sigma}_a^2) + 2(p + q)$$

where  $\hat{\sigma}_a^2$  is the MLE of the variance of the innovational noises. Note that for an ARMA( $p, q$ ) model, the number of independent parameters is  $p + q + 2$ . However, since 2 is a constant for all models, it is omitted from the above criterion function.

- BIC: Schwarz's information criterion (Schwarz, 1978, Ann. Statist.)

$$\text{BIC}(p, q) = n \ln(\hat{\sigma}_a^2) + (p + q) \ln(n).$$

- HQ: Hannan and Quinn (1979, JRSSB)

$$\text{HQ}(p, q) = n \ln(\hat{\sigma}_a^2) + c(p + q) \ln[\ln(n)], \quad c > 2,$$

For AR( $p$ ) models, there are other criteria available:

- Akaike's final prediction error (FPE):

$$\text{FPE}(p) = \frac{n + p}{n - p} \hat{\sigma}_p^2$$

where  $\hat{\sigma}_p^2$  is the MLE of residual variance when an AR( $p$ ) model is fitted to the data.

- Akaike's Bayesian information criterion (Bic):

$$\text{Bic}(p) = n \ln(\hat{\sigma}_p^2) - (n - p) \ln(1 - p/n) + p \ln(n) + p \ln[p^{-1}(\hat{\sigma}_z^2/\hat{\sigma}_p^2 - 1)]$$

where  $\hat{\sigma}_z^2$  is the sample variance of observations. This approach is very close to the BIC of Schwarz (1978). In fact, we have

$$\text{Bic}(p) \approx \text{BIC}(p) + O(p)$$

where  $O(p)$  denotes a term which is functionally independent of  $n$ .

- Parzen's CAT:

$$\text{CAT}(p) = \begin{cases} -(1 + (1/n)) & \text{if } p = 0 \\ (\frac{1}{n} \sum_{j=1}^p \frac{1}{\hat{\sigma}_j^2}) - \frac{1}{\hat{\sigma}_p^2} & \text{for } p > 0 \end{cases}$$

Recently, Hurvich and Tsai (1989, 1991, BKA) consider a bias-corrected AIC for AR( $p$ ) models as

$$\text{AICc}(p) = n \ln(\hat{\sigma}_a^2) + n \frac{1 + p/n}{1 - (p + 2)/n}.$$

This criterion function is asymptotically equivalent to AIC( $p$ ). In fact, we can write

$$\text{AICc}(p) = \text{AIC}(p) + \frac{2(p + 1)(p + 2)}{n - p - 2}.$$

This result can easily be shown by rewriting  $AIC(p)$  as

$$AIC(p) = n \ln(\hat{\sigma}_a^2) + n + 2(p + 1)$$

in which  $n$  and 2 are added. Since these two numbers are constant for all models, they do not affect the model selection. Simulation study indicates that AICc outperforms AIC in the small samples.

**Discussion:** Among the above criteria, BIC and HQ(.) are consistent in the sense that if the set of candidate models contains the “true” model, then these two criteria select the true model with probability 1 asymptotically. All the other criteria are inconsistent. On the other hand, since there is no “true” model in practice, “consistency” might not be a relevant property in application. Shibata (1980, Ann. Statist.) shows that AIC is asymptotically efficient in the sense that it selects the model which is closest to the unknown true model asymptotically. Here the unknown true model is assumed to be of infinite dimension.

There are advantages and disadvantages in using criterion functions in model selection. For instance, one possible disadvantage is that the selection is fully based on the data and the adopted information criterion. It is conceivable that certain substantive information is important in model selection, e.g. model interpretation. The information criterion does not incorporate such information in model selection.

In what follows, I briefly sketch a derivation of AIC information criterion. Let  $f(\cdot)$  and  $g(\cdot)$  be two probability density functions. A measure of goodness of fit by using  $g(\cdot)$  as an estimate of  $f(\cdot)$  is the entropy defined by

$$B(f; g) = - \int f(z) \ln\left(\frac{f(z)}{g(z)}\right) dz.$$

It can be shown that  $B(f; g) \leq 0$  and that  $B(f; g) = 0$  if and only if  $f(\cdot) = g(\cdot)$ . Thus, a maximum  $B(f; g)$  indicates  $g$  is close to  $f$ . Akaike (1973) argues that  $-B(f; g)$  can be used as a discrepancy between  $f(\cdot)$  and  $g(\cdot)$ . Since

$$\begin{aligned} -B(f; g) &= \int f(z) \ln\left(\frac{f(z)}{g(z)}\right) dz = \int \ln(f(z))f(z) dz - \int \ln(g(z))f(z) dz \\ &= \text{constant} - E_f[\ln(g(z))], \end{aligned}$$

where  $E_f$  denotes the expectation with respect to  $f(\cdot)$ , we define the discrepancy between  $f(\cdot)$  and  $g(\cdot)$  as

$$d(f; g) = E_f[-\ln(g(z))].$$

The objective then is to choose  $g$  which minimizes this discrepancy measure.

Suppose that  $\mathbf{x}$  is a set of  $n$  data points and the statistical analysis of  $\mathbf{x}$  is to predict  $y$  whose distribution is identical to that of the elements of  $\mathbf{x}$ . Such a prediction is made by

using the predictive distribution of  $y$  given  $\mathbf{x}$ . Denote the true distribution of  $y$  by  $f(y)$  and the predictive density of  $y$  given  $\mathbf{x}$  by  $g(y|\mathbf{x})$ . Then, the discrepancy is

$$d(f; g) = E_f[-\ln(g(y|\mathbf{x}))] = E_y[-\ln(g(y|\mathbf{x}))],$$

where we change the index  $f$  to  $y$  as  $f(\cdot)$  is the true density function of  $y$ . This discrepancy, of course, depends on the data realization  $\mathbf{x}$ . Therefore, the expected discrepancy is

$$D(f; g) = E_x[E_y(-\ln g(y|\mathbf{x}))]$$

where  $E_x$  denotes the expectation over the joint distribution of  $\mathbf{x}$ . The question then is how to estimate this expected discrepancy.

Here  $f(\cdot)$  is the true model and  $g(y|\mathbf{x})$  is an entertained model. Suppose now that the entertained models  $g(y|\mathbf{x})$  are indexed by the parameter  $\theta$  and that the true model  $f(\cdot)$  of  $y$  is within this class of candidate models, say  $f(y) = g(y|\theta_0)$ . Also, assume that the usual regularity conditions of MLE hold. Let  $\hat{\theta}(\mathbf{x})$  be the MLE of  $\theta$  given the data  $\mathbf{x}$ , i.e.

$$g(\mathbf{x}|\hat{\theta}(\mathbf{x})) = \max_{\theta} g(\mathbf{x}|\theta).$$

The following two results are well-known:

- As  $n \rightarrow \infty$ , the likelihood ratio statistic  $2 \ln g(\mathbf{x}|\hat{\theta}(\mathbf{x})) - 2 \ln g(\mathbf{x}|\theta_0)$  is asymptotically chi-square with degrees of freedom  $r = \dim(\hat{\theta}(\mathbf{x}))$ .
- By Taylor expansion and asymptotic normality of MLE,

$$2 \ln g(y|\theta_0) - 2 \ln g(y|\hat{\theta}(\mathbf{x})) \approx n(\hat{\theta}(\mathbf{x}) - \theta_0)' \mathbf{I}(\hat{\theta}(\mathbf{x}) - \theta_0) \sim \chi_r^2,$$

where  $\mathbf{I}$  is the Fisher information matrix of  $\theta$  evaluated at  $\theta_0$ .

Consequently, we have

$$2E_x \ln g(\mathbf{x}|\hat{\theta}(\mathbf{x})) - 2E_x \ln g(\mathbf{x}|\theta_0) = r$$

and

$$2E_x E_y \ln g(y|\theta_0) - 2E_x E_y \ln g(y|\hat{\theta}(\mathbf{x})) = r.$$

Summing over the above two equations and dividing the result by 2, we have

$$E_x \ln g(\mathbf{x}|\hat{\theta}(\mathbf{x})) - E_x E_y \ln g(y|\hat{\theta}(\mathbf{x})) = r.$$

Therefore,

$$E_x E_y [-\ln g(y|\hat{\theta}(\mathbf{x}))] = E_x [-\ln g(\mathbf{x}|\hat{\theta}(\mathbf{x}))] + r.$$

Since  $E_x \ln g(\mathbf{x}|\hat{\theta}(\mathbf{x}))$  is the expectation of the logarithm of the maximized likelihood of  $\mathbf{x}$ , Akaike proposes his AIC, based on the above equation, by estimating the expected discrepancy by

$$\hat{D}(f; g) = E_x [-\ln g(\mathbf{x}|\hat{\theta}(\mathbf{x}))] + r = -\ln g(\mathbf{x}|\hat{\theta}(\mathbf{x})) + r.$$

For Gaussian time series,  $-\ln g(\mathbf{x}|\hat{\theta}(\mathbf{x})) = \frac{n}{2} \ln(\hat{\sigma}_a^2) + C$ , where  $C$  is a function of  $n$  and  $2\pi$ . Therefore, dropping the constant  $C$  and multiplying by 2, we have

$$\text{AIC}(m) = n \ln(\hat{\sigma}_a^2) + 2r$$

where  $r$  is the dimension of  $\hat{\theta}(\mathbf{x})$  and  $m$  denotes the model corresponding to the density  $g(\cdot|\theta)$  entertained.

**Some examples.**