Ch. 16 Stochastic Model Building

(April 8, 2021)

Unlike linear regression model which usually has already an economic theoretic model built somewhere in economic literature, the parametric time series analysis of a stochastic process needs the ability to relating a stationary $ARMA$ model to real data. It is usually best achieved by a three-stage iterative procedure based on *model identification*, *estimation*, and *diagnostic checking* as suggested by Box and Jenkins (1976).

1 Model Identification

By **identification** we mean the use of the data, and of any information on how the series was generated, to suggest a subclass of **parsimonious** model worthy to be entertained. We usually transform the data, if necessary, so the assumption of covariance stationarity is a reasonable one. We then at this stage make an initial guess of small values of p and q for an $ARMA(p, q)$ model that might describe the transformed data.

1.1 Identifying the degree of Difference

Trend stationary or difference stationary ? See Ch. 19.

1.2 Use of the Autocorrelation and Partial Autocorrelation Function in Identification

1.2.1 Autocorrelation

Recall that if the data really follow an $MA(q)$ process, then its (population) autocorrelation $r_j (= \gamma_j/\gamma_0)$ will be zero for $j > q$. By contrast, if the data follow an $AR(p)$

process, then r_j will gradually decay toward zero as a mixture of exponential or damped sinusoids. On guide for distinguishing MA and AR representation, then, would be the decay properties of r_j. It is useful to have a rough check on whether r_j is effectively zero beyond a certain lag.

A natural estimate of the population autocorrelation r_j is provided by the corresponding sample moment (remember at this stage, you still have no 'model' to estimate, so it is natural to use moment estimator):

$$\hat{r_j} = \frac{\hat{\gamma}_j}{\hat{\gamma}_0},$$

where

$$\hat{\gamma}_j = \frac{1}{T}\sum_{t=j+1}^{T}(Y_t - \bar{Y})(Y_{t-j} - \bar{Y}) \quad for\ j = 0, 1, 2, ..., T-1$$

and

$$\bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t.$$

If the data were really generated by a Gauss $MA(q)$ process, then the covariance of the estimated autocorrelation $\hat{r}_j$, could be approximated by (see Box et al. (1994), p. 33)

$$Var(\hat{r}_j) \cong \frac{1}{T}\left\{1 + 2\sum_{i=1}^{q} r_i^2\right\} \quad for\ j = q+1, q+2, ... \tag{16-1}$$

To use (1) in practice, the estimated autocorrelation $\hat{r}_j$ $(j = 1, 2, ..., q)$ are substituted for the theoretical autocorrelation r_j, and when this is done we shall refer to the square root of (1) as the $large - lag\ standard\ error$. In particular, if we suspect that the data were generated by Gaussian white noise, then $\hat{r}_j \sim N(0, 1/T)$ for $j \neq 0$, that is $\hat{r}_j$ should lie between $\pm 2/\sqrt{T}$ about 95% of the time.

Example:

The following estimated autocorrelations were obtained from a time series of length $T = 200$ observations, generated from a stochastic process for which it was *known* that $r_1 = -0.4$ and $r_j = 0$ for $j \geq 2$:

$\hat{r}_1 = -0.38$, $\hat{r}_2 = -0.08$, $\hat{r}_3 = 0.11$, $\hat{r}_4 = -0.08$, $\hat{r}_5 = 0.02$, $\hat{r}_6 = 0.00$, $\hat{r}_7 = 0.00$,

$\hat{r}_8 = 0.00$, $\hat{r}_9 = 0.07$ and $\hat{r}_{10} = -0.08$.

On the assumption that the series is complete random: $H_0 : r_j = 0, \forall\ j$, then for all j, (1) yields

$$Var(\hat{r}_1) \cong \frac{1}{T} = \frac{1}{200} = 0.005.$$

Under the null hypothesis,

$$\hat{r}_1 \sim N(0, 0.005)$$

or the 95% confidence interval is

$$\begin{aligned} & -2 < \frac{\hat{r}_1}{\sqrt{0.005}} < 2 \\ \equiv & -0.14 < \hat{r}_1 < 0.14. \end{aligned}$$

Since the value of estimated $\hat{r}_1$ is -0.38, which is outside the confidence interval, it can be conclude that the hypothesis that $r_j = 0, \forall j$ is rejected.

It might be reasonable to ask next whether the series was compatible with the hypothesis that $r_1 \neq 0, r_j = 0, j \geq 2$. Using (1) with $q = 1$, the estimated large-lag variance under this assumption is

$$Var(\hat{r}_2) \cong \frac{1}{200}[1 + 2(-0.38)^2] = 0.0064.$$

Under the null hypothesis $H_0 : r_2 = 0$:

$$\hat{r}_2 \sim N(0, 0.0064)$$

or the 95% confidence interval is

$$\begin{aligned} & -2 < \frac{\hat{r}_2}{\sqrt{0.0064}} < 2 \\ \equiv & -0.16 < \hat{r}_2 < 0.16. \end{aligned}$$

Since the value of estimated $\hat{r}_2$ is -0.08, which is lying in the confidence interval, there is no reason to doubt the adequacy of the model $r_1 \neq 0, r_j = 0,\ j \geq 2$.

Similar approximate expressions as (1) for the covariance between the estimated correlation r_k and r_{k+s} at two different lags k and k_s have been given by Bartlett. In particular, the large-lag approximation reduces to

$$Cov(\hat{r}_k,\ \hat{r}_{k+s}) \cong \frac{1}{T}\sum_{v=-q}^{q} r_v r_{v+s} \quad for\ j = q+1, q+2, ...$$

Bartlett's results show that care is required in the interpretation of individual autocorrelations because large covariance can exist between neighboring values. This effect can sometimes distort the visual appearance of the autocorrelation function, which may fail to damp out according to expectation. Thus patterns in the estimated $\hat{r}_j$ may represent sampling error rather than patterns in the true r_j.

1.2.2 Partial Autocorrelation Function

In the $AR(1)$ process, Y_t and Y_{t-2} are correlated even though Y_{t-2} does not **directly** appear in the model. The correlation between Y_t and Y_{t-2} (i.e., γ_2) is equal to the correlation between Y_t and Y_{t-1} (i.e., γ_1) multiplied by the correlation between Y_{t-1} and Y_{t-2} (i.e., γ_1 again) so that $\gamma_2 = \gamma_1^2$. It is important to note that all such 'indirect' correlation are present in the ACF of any autoregressive process. In contrast, the **partial autocorrelation** between Y_t and Y_{t-s} eliminates the effects of the intervening values Y_{t-1} through Y_{t-s+1}. As such, in an $AR(1)$ process, the partial autocorrelation between Y_t and Y_{t-2} is equal to zero.

𝔇efinition (**Partial Autocovariance**):
The partial correlation between Y_t and Y_{t-m} is the simple covariance between Y_{t-k} and Y_t minus that part explained linear by the intervening lags. That is,

$$\gamma_m^{par} = Cov[Y_t - \hat{P}(Y_t|Y_{t-1}, ..., Y_{t-m+1}), Y_{t-m}].$$

𝔗heorem:
The partial autocorrelation between Y_t and Y_{t-m} is the last coefficient in the linear projection of Y_t on $(Y_{t-1}, Y_{t-2}, ..., Y_{t-m})$.

𝔓roof:
The forecast error ε_t is:

$$\varepsilon_t = Y_t - \hat{P}(Y_t|Y_{t-1}, ..., Y_{t-m+1}) = Y_t - \alpha_1 Y_{t-1} - \alpha_2 Y_{t-2} - ... - \alpha_{m-1} Y_{t-m+1} - r_m^{par} Y_{t-m} \quad (16\text{-}2)$$

which is uncorrelated with $(Y_{t-1}, Y_{t-2}, ..., Y_{t-m})$. From (2) we get

$$(Y_t - \alpha_1 Y_{t-1} - \alpha_2 Y_{t-2} - ... - \alpha_{m-1} Y_{t-m+1}) = r_m^{par} Y_{t-m} + \varepsilon_t. \tag{16-3}$$

Multiply both side of (3) by Y_{t-m} and take expectation we further have

$$E[Y_{t-m} \cdot (Y_t - \alpha_1 Y_{t-1} - \alpha_2 Y_{t-2} - ... - \alpha_{m-1} Y_{t-m+1})] = E[Y_{t-m} \cdot (r_m^{par} Y_{t-m} + \varepsilon_t)] \tag{16-4}$$

That is

$$r_m^{par} = \frac{\gamma_m^{par}}{\gamma_0},$$

as the definition of partial autocorrelation.

The partial autocorrelation is a device to exploits the fact that whereas an $AR(p)$ has an autocorrelation function which is infinite in extent, it can by its very nature be described in terms of p nonzero **functions** of the autocorrelations. The mth population partial autocorrelation (denoted $\alpha_m^{(m)}$) is defined as the last coefficient in a linear projection of Y on its m most recent value:

$$\hat{Y}_{t+1|t} - \mu = \alpha_1^{(m)}(Y_t - \mu) + \alpha_2^{(m)}(Y_{t-1} - \mu) + ... + \alpha_m^{(m)}(Y_{t-m+1} - \mu). \tag{16-5}$$

We saw in (16) of Chapter 15 that the vector $\boldsymbol{\alpha}^{(m)}$ can be calculated from

$$\begin{bmatrix} \alpha_1^{(m)} \\ \alpha_2^{(m)} \\ . \\ . \\ . \\ \alpha_m^{(m)} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & . & . & . & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & . & . & . & \gamma_{m-2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \gamma_{m-1} & \gamma_{m-2} & . & . & . & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ . \\ . \\ . \\ \gamma_m \end{bmatrix}.$$

Recall that if the data were really generated by an $AR(p)$ process, only the p most recent values of Y would be useful for forecasting. In this case, the projection coefficients on Y's more than p periods in the past are equal to zeros:

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

By contrast, if the data really were generated by an $MA(q)$ process with $q \geq 1$, then the partial autocorrelation $\alpha_m^{(m)}$ asymptotically approaches zero instead of cutting off abruptly.

Since forecast error ε_{t+1} is uncorrelated with $\mathbf{x}_t$, we could rewrite (2) as

$$Y_{t+1} - \mu = \alpha_1^{(m)}(Y_t - \mu) + \alpha_2^{(m)}(Y_{t-1} - \mu) + ... + \alpha_m^{(m)}(Y_{t-m+1} - \mu) + \varepsilon_{t+1}, \quad t \in \mathcal{T}$$

or

$$Y_t - \mu = \alpha_1^{(m)}(Y_{t-1} - \mu) + \alpha_2^{(m)}(Y_{t-2} - \mu) + ... + \alpha_m^{(m)}(Y_{t-m} - \mu) + \varepsilon_t, \quad t \in \mathcal{T} \quad (16\text{-}6)$$

The reason why the quantity $\alpha_m^{(m)}$ defined through (2) is called the *partial autocorrelation* of the process $\{Y_t\}$ at lag m is clear from (3), since it is actually equal to the partial correlation between the variable Y_t and Y_{t-m} adjusted for the intermediate variables $Y_{t-1}, Y_{t-2}, ..., Y_{t-m+1}$, and $\alpha_m^{(m)}$ measures the correlation between Y_t and Y_{t-m} after adjusting for the effect of $Y_{t-1}, Y_{t-2}, ..., Y_{t-m+1}$ (or the correlation between Y_t and Y_{t-m} not account for by $Y_{t-1}, Y_{t-2}, ..., Y_{t-m+1}$). See the counterpart-result from sample on p.6 of Chapter 6.

A natural estimate of the mth partial autocorrelations is the last coefficients in an OLS regression of Y on a constant and its m most recent values:

$$Y_t = \hat{c} + \hat{\alpha}_1^{(m)} Y_{t-1} + \hat{\alpha}_2^{(m)} Y_{t-2} + ... + \hat{\alpha}_m^{(m)} Y_{t-m} + \hat{e}_t, \quad (16\text{-}7)$$

where $\hat{e}_t$ denotes the OLS regression residual. If the data were really generated by an $AR(p)$ process, then the sample estimate $\hat{\alpha}_m^{(m)}$ would have a variance around the true value (0) that could be approximated by (see Box et al. 1994, p.68)

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

Example:
See Example 4.1 on p. 112 of Hamilton. (The standard error of Figure 4.2 (a) may be wrong for it is only the standard error of $H_0 : r_j = 0\ j \geq 1$.)

Exercise:
Let $Y_t = -0.7Y_{t-1} + \varepsilon_t - 0.7\varepsilon_{t-1}$. Calculate the first 9 autocorrelations, $r_j, j = 1, 2, .., 9$ and the first 8 partial autocorrelations, $r_i^{par}, i = 1, 2, ..., 8$ of Y_t.[1]

[1]Hint: The partial autocorrelation r_2^{par} can be calculated from multiplying Y_{t-1} and Y_{t-2} on

$$Y_t = \alpha_1^{(2)} Y_{t-1} + \alpha_2^{(2)} Y_{t-2} + \varepsilon_t$$

1.3 Use of Model Selection Criteria

Another approach to model selection is the use of information criteria such as the Akaike Information Criterion (AIC) proposed by Akaike (1974), the Schwartz Bayesian Criterion (SBC) of Schwartz (1978) and the Hurvich and Tsai (1989) Criterion (AIC_c) which is of particular use when the sample size is small. In the implementation of this approach, a range of potential $ARMA$ models is estimated by maximum likelihood methods to be discussed in Chapter 17, and for each, a criterion such as AIC (normalized by sample size T, given by

$$AIC = \frac{-2\ln(maximized\ likelihood) + 2m}{T} \approx \ln(\hat{\sigma}^2) + \frac{2m}{T} \tag{16-10}$$

or the related SBC given by

$$BIC = \ln(\hat{\sigma}^2) + \frac{m\ln(T)}{T} \tag{16-11}$$

and the AIC_c is[2]

$$AIC_c = AIC + \frac{2(m+1)(m+2)}{T-m-2}, \tag{16-12}$$

is evaluated, where $\hat{\sigma}^2$ denotes the maximum likelihood estimate of σ^2, and $m = p+q+1$ denotes the number of parameters estimated in the model, including a constant term.

---

and take expectation to obtain

$$\begin{aligned} E(Y_{t-1}Y_t) &= \alpha_1^{(2)}E(Y_{t-1}Y_{t-1}) + \alpha_2^{(2)}E(Y_{t-1}Y_{t-2}) + E(Y_{t-1}\varepsilon_t) \\ E(Y_{t-2}Y_t) &= \alpha_1^{(2)}E(Y_{t-2}Y_{t-1}) + \alpha_2^{(2)}E(Y_{t-2}Y_{t-2}) + E(Y_{t-2}\varepsilon_t), \end{aligned}$$

i.e.

$$\begin{aligned} \gamma_1 &= \alpha_1^{(2)}\gamma_0 + \alpha_2^{(2)}\gamma_1 \\ \gamma_2 &= \alpha_1^{(2)}\gamma_1 + \alpha_2^{(2)}\gamma_0 \end{aligned}$$

or

$$r_1 = \alpha_1^{(2)} + \alpha_2^{(2)}r_1 \tag{16-8}$$
$$r_2 = \alpha_1^{(2)}r_1 + \alpha_2^{(2)}. \tag{16-9}$$

Solving (8) and (9) we have

$$r_2^{par} = \alpha_2^{(2)} = \frac{r_2 - r_1^2}{1 - r_1^2}.$$

[2]Thus, the AIC_c is the sum of AIC and an additional non-stochastic penalty term $\frac{2(m+1)(m+2)}{T-m-2}$.

In the criteria above, the first term essentially corresponds to minus $2/T$ times the log of the maximized likelihood, while the second term is a 'penalty factor' for inclusion of additional parameters in the model.

In the information criteria approach, models that yield a **minimum value** for the criterion are to be preferred, and the AIC or SBC or AIC_c (when small sample size) values are compared among various model as the basis for selection of the models. However, one immediate disadvantage of this approach is that several models may have to be estimated by MLE, which is computationally time consuming and expensive. For this reason, Hannan and Rissanen (1982) propose an alternative model selection procedure. See Box and Jenkins (1994), p. 201 for details.

2 Model Estimation

By **estimation** we mean efficient use of the data to make inference about parameters conditional on the adequacy of the model entertained. See Chapter 17 for details.

3 Model Diagnostic Checking

By **diagnostic checking** we mean checking the fitted model in its relation to the data with intent to reveal model inadequacies and so to achieve model improvement.

Suppose that using a particular time series, the model has been identified and the parameters estimated using the methods described in Chapter 17. The question remains (unlike the regression analysis where an economic or finance model is provided by theoretical literature) of deciding whether this model is adequate. If there should be evidence of serious inadequacy, we shall need to know how the model should be modified. By reference to familiar procedures outside time series analysis, the scrutiny of residuals for the analysis of variance would be called diagnostic checks.

3.1 Diagnostic Checks Applied to residuals

It cannot be too strongly emphasized that *visual inspection of a plot of the residual* is an indispensable first step in the checking process.

3.1.1 Autocorrelation Check

Suppose we have identified and fitted a model

$$\phi(L)Y_t = \theta(L)\varepsilon_t$$

with MLE estimator $(\hat{\phi}, \hat{\theta})$ obtained for the parameters. Then we shall refer the quantities

$$\hat{\varepsilon}_t = \hat{\theta}^{-1}(L)\hat{\phi}(L)Y_t$$

as the *residuals*. The residuals are computed recursive from $\hat{\theta}(L)\hat{\varepsilon}_t = \hat{\phi}(L)Y_t$ as

$$\hat{\varepsilon}_t = Y_t - \sum_{j=1}^{p} \hat{\phi}_j Y_{t-j} + \sum_{j=1}^{q} \hat{\theta}_j \hat{\varepsilon}_{t-j} \quad t = 1, 2, ..., T$$

using either zero initial values (conditional method) or back-forecasted initial value (exact method) for the initial $\hat{\varepsilon}'s$ and $Y's$.

Now it is possible to show that if the model is adequate,

$$\hat{\varepsilon}_t = \varepsilon_t + O\left(\frac{1}{\sqrt{T}}\right). \text{ (read as big O } T^{-1/2}\text{, it means this term has to multiply } T^{1/2} \text{ to be bounded. That is, it converges to zero itself !)}$$

As the series length increase, the $\hat{\varepsilon}_t$'s become close to the white noise ε_t's. Therefore, one might expect that study of the $\hat{\varepsilon}_t$'s could indicate the existence and nature of model adequacy. In particular, recognizable patterns in the estimated autocorrelations function of the $\hat{\varepsilon}_t$'s, $\hat{r}_j(\hat{\varepsilon})$, and using (1), could point out to appropriate modification in the model.

3.1.2 Portmanteau Lack-of-Fit Test

Rather than consider the $\hat{r}_j(\hat{\varepsilon})$'s individually, an indication is often needed of whether, say, the first 20 autocorrelations of the $\hat{\varepsilon}_t$'s taken as a whole, indicating inadequacy of the model. Suppose we have the first k autocorrelation [3] $\hat{r}_j(\hat{\varepsilon}), j = 1, 2, ..., k$ form any

[3] Here, k is chosen sufficiently large so that the weight φ_j in the model written in the form

$$Y_t = \phi(L)^{-1}\theta(L)\varepsilon_t = \varphi(L)\varepsilon_t$$

will be negligible small after $j = k$.

$ARMA(p,q)$ process; then it is possible to show that if the model is appropriate, the Box-Pierce (1970) Q statistics

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

is approximately distributed as χ_k^2. On the other hand, if the model is inappropriate, the average value of Q will be inflated. A refinement that appears to have better finite-sample properties is the Ljung-Box (1979) statistics:

$$Q' = T(T+2)\sum_{j=1}^{k} \frac{\hat{r}_j^2(\hat{\varepsilon})}{T-k}.$$

The limiting distribution of Q' is the same as that of Q.

3.1.3 Normality Test

The section considers the general problem of using the moments of the fitted residuals to make inference about the distribution of the true disturbances.

The natural estimator of

$$\mu_r = E(\varepsilon^r)$$

would be

$$m_r = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_t^r.$$

The normal distribution is symmetric and mesokurtic. The symmetry implies that the third moment $E(\varepsilon^3)$ is zero. The standard measure of symmetry of a distribution is the skewness coefficient,

$$\sqrt{\alpha_1} = \frac{E(\varepsilon^3)}{(\sigma^2)^{3/2}}.$$

Kurtosis is a measure of the thickness of the tails of a distribution. The measure is

$$\alpha_2 = \frac{E(\varepsilon^4)}{(\sigma^2)^2},$$

which is 3 for a normal distribution. Therefore, we might compare a distribution with the normal distribution by comparing this skewness with zero and its kurtosis to three.

In practice, the usual measure is the **degree of excess**, $(\alpha_2 - 3)$. Bera and Jarque (1980) use this device in a Wald test statistics. Under the hypothesis of normality, the test statistics would be

$$W = T\left[\frac{a_1}{6} + \frac{(a_2-3)^2}{24}\right] \xrightarrow{L} \chi^2_2,$$

where a_1 and a_2 are replacing μ_r with m_r in α_1 and α_2, respectively. In large samples, W is distributed as chi-squared with two degrees of freedom.

Exercise:
Build up a stochastic model to the data set I give to you from Box-Jenkins procedure.