12: AUTOREGRESSIVE AND MOVING AVERAGE PROCESSES

IN DISCRETE TIME

Moving Averages

Recall that a **white noise process** is a series \( \{ \varepsilon_t \}_{t=-\infty}^{\infty} \) of uncorrelated zero mean random variables having variance \( \sigma^2 \). The white noise process has spectral density \( f_\varepsilon(\lambda) = \frac{\sigma^2}{2\pi} \) on \([-\pi, \pi]\). A **moving average process** is the series

\[
X_t = \sum_{j=-\infty}^{\infty} a_j \varepsilon_{t-j} .
\]

Note that \( X_t \) is a linear combination (i.e., "moving average") of past present and future values of the innovations or shocks, \( \varepsilon_t \). A more useful model for forecasting purposes is the **one-sided moving average**

\[
X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} .
\]

Both (1) and (2) are said to be **infinite** moving averages, since there are an infinite number of parameters \( a_j \) needed to determine the process. A **finite** moving average results if only finitely many parameters are used:

\[
X_t = \sum_{j=-m}^{n} a_j \varepsilon_{t-j} .
\]

The moving average \( \{ X_t \} \) (given by (1), say) is the output of a linear filter with input \( \{ \varepsilon_t \} \) and transfer function

\[
B(\lambda) = \sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j} .
\]

Thus,

\[
X_t = \int_{-\pi}^{\pi} e^{i\lambda t} B(\lambda) dZ_\varepsilon(\lambda) ,
\]

and \( \{ X_t \} \) has spectral density
\[ f_X(\lambda) = |B(\lambda)|^2 f_\varepsilon(\lambda) = \frac{\sigma^2}{2\pi} |\sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j}|^2. \]

The variance of \( X_t \) is

\[ \text{Var } X_t = \int_{-\pi}^{\pi} f_X(\lambda) d\lambda = \sigma^2 \sum_{j=-\infty}^{\infty} a_j^2. \]

(Exercise).

**KEY FACT:** Every weakly stationary time series with continuous spectrum can be represented as a (possibly) infinite moving average.

We will prove this important result for the process \( \{Y_t\} \), assumed to have a strictly positive spectral density \( f_Y(\lambda) \). Let \( \sigma^2 \) be a positive number, and define the process

\[ \varepsilon_t = \int_{-\pi}^{\pi} e^{i\lambda t} A(\lambda)dZ(\lambda), \]

where

\[ A(\lambda) = \frac{\sigma}{(2\pi)^{1/2}} \frac{1}{(f_Y(\lambda))^{1/2}}. \]

The matching condition is satisfied (i.e., \( \{\varepsilon_t\} \) is a well-defined weakly stationary process with zero mean and finite variance) since

\[ \int_{-\pi}^{\pi} |A(\lambda)|^2 f_Y(\lambda) = \sigma^2. \]

Now, \( \{\varepsilon_t\} \) has spectrum

\[ f_\varepsilon(\lambda) = |A(\lambda)|^2 f_Y(\lambda) = \frac{\sigma^2}{2\pi}. \]

It follows that \( \varepsilon_t \) is a white noise process. By inverting the filter, \( \{Y_t\} \) can be recovered from \( \{\varepsilon_t\} \):

\[ Y_t = \int_{-\pi}^{\pi} e^{i\lambda t} \frac{1}{A(\lambda)} dZ_\varepsilon(\lambda). \]

Now, \( \frac{1}{A(\lambda)} \) is a square integrable function on \([-\pi, \pi]\) and hence has a Fourier series expansion.
where \{a_j\} are constants with \( \sum a_j^2 < \infty \). (See Example 1.2 of Koopmans for the relevant Hilbert space theory.) Thus,

\[
Y_t = \int \pi f_{\lambda} \left( \sum a_j e^{-i\lambda_j} \right) d\lambda = \sum a_j \int \pi f_{\lambda} e^{-i(j-\lambda)} d\lambda = \sum a_j e_{t-j}.
\]

This is the desired moving average representation of \{Y_t\}.

**Autoregressive Processes**

A discrete time weakly stationary process \{X_t\} is said to be an **autoregression** of order \( p \) (AR(\( p \))) if there exists a white noise process \{\varepsilon_t\} with variance \( \sigma^2 \), and constants \( b_1, \ldots, b_p \) such that

\[
X_t + b_1 X_{t-1} + \cdots + b_p X_{t-p} = \varepsilon_t \quad \text{(all } t \text{)}.
\]

Equation (3) expresses \( X_t \) as a linear regression on its own past; hence the term "autoregression". Equation (4) is a condition to ensure that the "innovation" at time \( t \), namely \( \varepsilon_t \), really represents new information in the sense of being uncorrelated with past values of the series \{\( X_t \)\}. The definition of an **infinite order autoregression** is the same as above except that (3) is replaced by

\[
\sum b_k X_{t-k} = \varepsilon_t, \quad \text{where } b_0 = 1 \text{ and } \sum b_k^2 < \infty.
\]

Since \{\( \varepsilon_t \)\} is the result of passing \{\( X_t \)\} through a linear filter with transfer function \( B(\lambda) = \sum b_k e^{-i\lambda k} \), it follows that \( dZ_\varepsilon(\lambda) = B(\lambda) dZ_X(\lambda) \), and hence the spectrum of the AR process is

\[
f_X(\lambda) = \frac{\sigma^2}{2\pi} \left( \sum b_k e^{-i\lambda k} \right)^2.
\]

For AR(\( p \)) models, a key question is whether there exists a weakly stationary solution \{\( X_t \)\} to the stochastic difference equations (3) and (4). These equations define the AR(\( p \)) process implicitly, and a
solution cannot be guaranteed to exist unless an additional condition is imposed on the parameter values \( b_1, \ldots, b_p \). The condition involves the zeros (roots) of the function

\[
\beta(z) = 1 + b_1 z + b_2 z^2 + \cdots + b_p z^p
\]

considered as a polynomial in the complex variable \( z \):

**For a finite autoregression to exist, it is sufficient that all the zeros of \( \beta(z) \) lie outside the unit circle.** The solution will be a one-sided moving average of the defining white noise process with coefficients equal to those of the power series expansion of \( 1/\beta(z) \).

This is an important result, and a sketch of the proof follows. First, note that the transfer function is \( B(\lambda) = \beta(e^{-i\lambda}) \). If \( \beta(z) \) has no zeros on the unit circle, then

\[
X_t = \int_{-\pi}^{\pi} e^{i\lambda t} [1/B(\lambda)] dZ(\lambda)
\]

satisfies the matching condition, and hence is a weakly stationary finite-variance solution to (3). It follows that \( X_t \) has the moving average representation (see previous section)

\[
X_t = \sum_{j=-\infty}^{\infty} a_j \varepsilon_{t-j}
\]

(5)

where the \( a_j \) are the Fourier coefficients of \( 1/B(\lambda) \):

\[
1/B(\lambda) = \sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j}
\]

(6)

Equation (5) gives an explicit form for the solution \( X_t \) to (3). The question that remains is whether this solution satisfies (4). Assuming that the zeros of \( \beta(z) \) lie outside the unit circle, there must be some real number \( \rho > 1 \) such that \( 1/\beta(z) \) is analytic for \( |z| < \rho \). It follows that for \( |z| < \rho \), \( 1/\beta(z) \) can be represented as a power series (Laurent series) involving *positive powers* of \( z \) only. On the unit circle (i.e., if \( |z| = 1 \)), this power series must be identical to the Fourier series (6). It follows that \( a_j = 0 \) for \( j < 0 \), and hence

\[
X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}
\]

i.e., \( X_t \) can be written as a *one-sided* moving average. Hence, if \( s < t \),
\[ E[X_t \varepsilon_t] = E[\sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \varepsilon_t] = \sum_{j=0}^{\infty} a_j E[\varepsilon_{t-j} \varepsilon_t] = 0. \]

Thus, \( \{X_t\} \) satisfies condition (4), and the proof is complete.

We have just shown that the condition that the zeros of \( \beta(z) \) lie outside the unit circle is sufficient for the existence of a finite autoregression. It can be shown that this condition is also necessary. Thus, a finite autoregression exists if and only if the zeros of \( \beta(z) \) lie outside the unit circle.

**Parameter Estimation**

Suppose we are willing to assume that the process \( \{X_t\} \) obeys some finite-parameter model of known dimension. Two examples are the \( p \)'th order autoregressive (AR(\( p \))) model,

\[
\sum_{j=0}^{p} b_j X_{t-j} = \varepsilon_t \quad (b_0 = 1, \text{ var } \varepsilon_t = \sigma^2),
\]

and the \( q \)'th order moving average (MA(\( q \))) model,

\[
X_t = \sum_{j=0}^{q} a_j \varepsilon_{t-j} \quad (a_0 = 1, \text{ var } \varepsilon_t = \sigma^2).
\]

Since in practice we will have only a limited amount of data \( x_0, \ldots, x_{n-1} \) from the stochastic process \( \{X_t\} \), the parameters \( (b_1, \ldots, b_p, \sigma^2) \) for the AR(\( p \)), \( (a_1, \ldots, a_q, \sigma^2) \) for the MA(\( q \)) will typically be unknown, and hence must be estimated from the available data. If we assume that the white noise process is Gaussian, then we may estimate the parameters by maximum likelihood. Let \( \theta \) denote the vector of model parameters, and let \( x = (x_0, \ldots, x_{n-1})^T \). It can be shown that \( \theta \) determines the covariance function \( \{ c_r \} \) \( (r = 0, 1, \ldots) \) of the process. Let \( \Sigma_0 \) be the resulting \( n \times n \) covariance matrix of \( X_0, \ldots, X_{n-1} \). The matrix \( \Sigma_0 \) has \( (i, j) \) entry \( \Sigma_0(i, j) = c_{i-j} \) for \( i, j = 0, \ldots, n-1 \). Then the (Gaussian) likelihood function is the joint multivariate Gaussian density function

\[
L(x) = (2\pi)^{-n/2} |\Sigma_0|^{-1/2} \exp(-x^T \Sigma_0^{-1} x / 2).
\]

The maximum likelihood estimate (MLE) \( \hat{\theta} \) is obtained by minimizing the quantity

\[
-2 \log L(x) = n \log (2\pi) + \log |\Sigma_0| + x^T \Sigma_0^{-1} x,
\]

where the right hand side is treated as a function of \( \theta \), with \( x \) held fixed at its observed value.
The chief drawback of the MLE is that it may be quite difficult to compute. An explicit formula for $\hat{\theta}$ does not exist, so the estimator must be calculated by numerical optimization methods. These methods are inherently difficult to use unless the model dimension is small. Furthermore, the likelihood function itself is often very rough, and contains many local optima. As a result, the computation of the MLE may require a great deal of computer time.

The computation of the likelihood function even for a single given value of $\theta$ is a non-trivial task itself, since in principle it requires the inversion of the $n \times n$ matrix $\Sigma$. This problem has largely been solved, however, by taking advantage of the Toeplitz nature of the covariance matrix, or by using so-called State Space or Kalman Filter methods.

To save time, computationally efficient methods of parameter estimation have been developed. Examples include the Burg and Yule-Walker methods for AR models, and the least squares and Hannan-Rissanen estimators for more general models. Unfortunately, some of the computationally efficient estimators do not perform as well as the MLE in terms of generally accepted performance criteria, such as (statistical) asymptotic efficiency. Still, these estimators are useful as starting values in the iterative search for the MLE.

Model Selection

Another important practical problem arising in the application of finite-parameter models is that of model selection. It is clear that in most practical situations both the model type (for example, $AR$ or $MA$) as well as the corresponding model dimension ($p$ or $q$) will be unknown. Instead, we have a number of candidate models we are willing to try, and we want to select the one which is "best". Box and Jenkins refer to this as the model identification problem, a terminology which seems to imply that one of the candidate models is in fact the true model (i.e., the model which actually generated the data.)

Box and Jenkins propose to identify $AR$ or $MA$ models by examining plots of the sample autocorrelation function $\{\hat{\rho}_r\} = \{\hat{\rho}_r / \sigma^2\}$, and another quantity called the sample partial autocorrelation function. It can be shown that if the process is $MA(q)$, then the population autocorrelation function $\{\rho_r\} = \{\rho_r / \sigma^2\}$ will be zero for all lags exceeding $q$. Thus, aside from sampling variability, the sample
autocorrelation function should display this same "cut-off" at lag $q$. Similarly, if the process is $AR(p)$, then the sample partial autocorrelation function should cut off at lag $p$, aside from sampling variability.

We, on the other hand, do not wish to assume that any of the candidate models is necessarily the true model. In fact, we do not think it is usually reasonable to assume that the true model is even finite-dimensional. Nevertheless, it may be quite useful (especially for rapid on-line forecasting) to fit finite-dimensional models to the data, as long as it is clearly recognized that such models can only provide _approximations_ to the truth. This point of view is shared by many time series analysts, including Tukey, Bloomfield, Akaike, Parzen and Shibata. In this context, an important and useful method for model selection has been developed by Akaike. The method is called the Akaike Information Criterion (AIC). The method proceeds by computing the _criterion function_

$$AIC = -2 \log \text{(maximized likelihood)} + 2(\#\text{parameters})$$

for each of the candidate models, where (\#parameters) denotes the number of parameters estimated in the model. (For example, the $AR(p)$ model has $p+1$ parameters and the $MA(q)$ model has $q+1$ parameters, since the innovation variance $\sigma^2$ is included separately in these models.) The model which minimizes AIC is then selected for final use. If the models are not estimated by maximum likelihood, then the first term of AIC may be replaced by $n \log \hat{\sigma}^2$, where $\hat{\sigma}^2$ is the estimated white noise variance.

It has been shown by Shibata that if the true model is Gaussian $AR(\infty)$ and does not degenerate to a finite-order autoregression, then AIC provides an asymptotically optimal model selection, assuming that the candidate models are finite-dimensional Gaussian $AR$ models. Other model selection methods include the CAT criterion of Parzen (for AR models only), the $AIC_C$ criterion of Hurvich and Tsai (this is a correction to AIC to provide improved performance in small samples) and the Bayesian Information Criterion, $BIC$, of Schwarz (which assumes that one of the candidate models is the true model.) If $m$ denotes the number of parameters in the candidate model, then $AIC_C$ and $BIC$ are given by

$$AIC_C = n \log \hat{\sigma}^2 + 2(m + 1) \frac{n}{n - m - 2} , \quad BIC = n \log \hat{\sigma}^2 + m \log n .$$