FINC-520
Christiano

Wold Representation Theorem

We have discussed a class of ARMA models and derived restrictions which ensure they are models for covariance stationary time series. We have shown that these ARMA models imply the data are a linear combination of current and past one-step-ahead forecast errors, with weights that decay at a geometric rate.[1] Here, we consider the class of covariance stationary processes and ask whether ARMA models are a strict subset of that class. We start from the assumption that a process is covariance stationary and we study the projection of the process onto its current and past one-step-ahead forecast errors. This decomposition of a covariance stationary process into a projection onto current and past one-step-ahead forecast errors (the 'purely indeterministic part' of the process) and a projection error (the 'purely deterministic part') is called the Wold Representation Theorem.

We conclude that there are two ways in which ARMA models represent a restriction on the class of covariance stationary processes. First, in an ARMA model the purely deterministic part is absent. That is, a researcher working with an ARMA model implicitly assumes both that the process is covariance stationarity and that the process is purely indeterministic. Second, according to the Wold Representation Theorem, covariance stationarity implies that the weights on current and past one-step-ahead forecast errors are square summable. This is weaker than the geometric decay property implied by ARMA models.

The first section below states the Wold Representation Theorem, and then provides an informal proof using the argument in Sargent (1979). I then summarize the implications of the theorem for the ARMA models that we study.

1. The Wold Theorem

Theorem 1.1. *Suppose that $\{x_t\}$ is a covariance stationary process with $Ex_t = 0$ and covariance function, $\gamma(j) = Ex_t x_{t-j}$, $\forall j$. Then*

$$x_t = \sum_{j=0}^{\infty} d_j \varepsilon_{t-j} + \eta_t$$

where

$$\begin{aligned} d_0 &= 1,\ \sum_{j=0}^{\infty} d_j^2 < 0,\ E\varepsilon_t^2 = \sigma_\varepsilon^2,\ E\varepsilon_t\varepsilon_s = 0 \text{ for } t \neq s, && (1.1) \\ E\varepsilon_t &= 0,\ E\eta_t\varepsilon_s = 0 \text{ for all } t, s, && (1.2) \\ P\left[\eta_{t+s} | x_{t-1}, x_{t-2}, ...\right] &= \eta_{t+s},\ s \geq 0. && (1.3) \end{aligned}$$

[1] This linear combination is derived by recursive substitution. Before doing this, one has to make sure that the ARMA model error is the one-step-ahead forecast error, if necessary by 'flipping' moving average roots.

The first part of the representation of x_t looks just like the MA(∞) with square summable moving average terms that we have worked with, while the second part, η_t, is something new. That part is called the deterministic part of x_t because η_t is perfectly predictable based on past observations on x_t.

The style of proof is constructive. We will show that given only covariance stationarity, we can build the Wold representation with the indicated properties. We will not provide a fully rigorous proof and a key result will simply be assumed. The proof is an application of linear projections, and the orthogonality and recursive properties of projections. The proof style follows that in Sargent (1979).

We first find the d_j's and ε_t and establish the required properties. Then, we find the projection error, η_t.

We begin with a preliminary result. Let x_t be a covariance stationary process. Let

$$\hat{x}_t^{(n)} = P\left[x_t | x_{t-1}, ..., x_{t-n}\right],$$

and write

$$x_t = \hat{x}_t^{(n)} + \varepsilon_t^{(n)}.$$

From the orthogonality property of projections we know that

$$\begin{array}{rcl} & & \varepsilon_t^{(n)} \perp \left(x_{t-1}, ..., x_{t-n}\right) \\ E\varepsilon_t^{(n)} & = & \sigma^{2(n)}. \end{array}$$

We assume, without proof, the following result:

$$\begin{array}{rcll} \hat{x}_t^{(n)} & \rightarrow & \hat{x} = P\left[x_t | x_{t-1}, x_{t-2}, ...\right] & (1.4) \\ x_t & = & \hat{x}_t + \varepsilon_t, \ E\varepsilon_t^2 = \sigma^2 & (1.5) \\ & & \varepsilon_t \perp \left(x_{t-1}, x_{t-2}, ...\right). & (1.6) \end{array}$$

The disturbance, ε_t, is known as the 'innovation' in x_t or its 'one-step-ahead forecast error'. It is easy to see that ε_t is a serially uncorrelated process. In particular,

$$\varepsilon_t = x_t - P\left[x_t | x_{t-1}, x_{t-2}, ...\right],$$

so that it is a linear combination of current and past x_t's. It follows that since ε_t is orthogonal to past x_t's, it is also orthogonal to past ε_t's.

1.1. Projection of x_t onto current and past ε_t's

We now consider the projection of x_t on current and past ε_t's:

$$\tilde{x}_t^m = \sum_{j=0}^{m} d_j \varepsilon_{t-j}.$$

The notation, $\tilde{x}_t^m$, is intended to signal that the projection used here is different from the one used to define the ε_t's. The lack of autocorrelation between the ε_t's makes the analysis of the projection coefficients particularly simple. The orthogonality condition associated with the projection is:

$$E\left(x_t-\sum_{j=0}^{m}d_j\varepsilon_{t-j}\right)\varepsilon_{t-k}=0,\ \ k=0,...,m,$$

which, by the lack of correlation in the ε_t's, reduces to:

$$Ex_t\varepsilon_{t-k}-d_kE\varepsilon_{t-k}^2=0,$$

so that

$$d_k=\begin{cases}\frac{Ex_t\varepsilon_{t-k}}{\sigma^2},\ k=1,2,...,m\\ 1,\ k=0\end{cases}$$

That $Ex_t\varepsilon_t=\sigma^2$ follows from (1.5):

$$Ex_t\varepsilon_t=E\left(\hat{x}_t+\varepsilon_t\right)\varepsilon_t=\sigma^2,$$

because $\hat{x}_t$ is a linear function of past x_t's and ε_t is orthogonal to those x_t's. A key property of the projection is that d_k is not a function of m. This reflects the lack of serial correlation in the ε_t's.

We now establish the square summability of the d_j's. Any variance must be non-negative, and this is true of the error in the projection of x_t onto $\varepsilon_t,...,\varepsilon_{t-m}:$

$$E\left(x_t-\sum_{j=0}^{m}d_j\varepsilon_{t-j}\right)^2\geq 0,$$

or,

$$\begin{aligned}&Ex_t^2-2\sum_{j=0}^{m}d_jEx_t\varepsilon_{t-j}+\sum_{j=0}^{m}d_j^2\sigma^2\\ =\ \ &Ex_t^2-\sigma^2\sum_{j=0}^{m}d_j^2\geq 0.\end{aligned}$$

This must be true for all m. Since Ex_t^2 is a fixed number by covariance stationarity, it follows that

$$\lim_{m\to\infty}\sum_{j=0}^{m}d_j^2<\infty.$$

In addition the sum is a non-decreasing sequence because each term (being a square) is non-negative. From this we conclude that the above sum converges to some finite number:

$$\sum_{j=0}^{m}d_j^2\to\sum_{j=0}^{\infty}d_j^2.$$

Given the square summability of the d_j's, it follows that $\tilde{x}_t^m$ forms a Cauchy sequence, so that

$$\tilde{x}_t^m=\sum_{j=0}^{m}d_j\varepsilon_{t-j}\to\tilde{x}_t=\sum_{j=0}^{\infty}d_j\varepsilon_{t-j}.$$

To verify that $\tilde{x}_t^m$ is Cauchy, we establish that for each $e > 0$, there exists an n such that for all $m > n$

$$E\left(\tilde{x}_t^m - \tilde{x}_t^n\right)^2 = \left(\sum_{j=n+1}^{m} d_j^2\right)\sigma^2 < e.$$

1.2. Constructing the η_t's

We define η_t as the difference between x_t and its projection onto the current and past ε_t's:

$$\eta_t = x_t - \tilde{x}_t. \tag{1.7}$$

We first establish that

$$E\eta_t\varepsilon_s = 0 \text{ for all } t, s.$$

That $E\eta_t\varepsilon_s = 0$ for $s > t$ is obvious because η_t is a linear function of x_t and past ε_t's, and ε_s is orthogonal to all these things, $s > t$. That $E\eta_t\varepsilon_s = 0$ for $s \leq t$ follows from the fact that η_t is the error in the projection of x_t on current and past ε_t's. In particular,

$$E\eta_t\varepsilon_{t-k} = Ex_t\varepsilon_{t-k} - d_k\sigma^2 = d_k\sigma^2 - d_k\sigma^2 = 0.$$

Next we establish that η_t is perfectly predictable from past x_t's. Note

$$P\left[\eta_t | x_{t-1}, x_{t-2}, ...\right] = P\left[x_t | x_{t-1}, x_{t-2}, ...\right] - \sum_{j=0}^{\infty} d_j P\left[\varepsilon_{t-j} | x_{t-1}, x_{t-2}, ...\right], \tag{1.8}$$

where we have used the linearity of projections, $P\left[A + B|\Omega\right] = P\left[A|\Omega\right] + P\left[B|\Omega\right]$. Consider the last term, involving the projections of the ε_t's. In the case, $j = 0$:

$$P\left[\varepsilon_t | x_{t-1}, x_{t-2}, ...\right] = \sum_{j=1}^{\infty} \psi_j x_{t-j}.$$

The ψ_j's satisfy the orthogonality conditions:

$$E\left(\varepsilon_t - \sum_{j=1}^{\infty} \psi_j x_{t-j}\right) x_{t-k} = 0, \;\; k = 1, 2, 3, ... \; .$$

Recall that ε_t is orthogonal to $(x_{t-1}, x_{t-2}...)$ so that $\psi_j = 0$ satisfies the orthogonality conditions. Sufficiency of the orthogonality conditions guarantees that

$$P\left[\varepsilon_t | x_{t-1}, x_{t-2}, ...\right] = 0.$$

Now consider

$$P\left[\varepsilon_t | x_t, x_{t-1}, ...\right]$$

Recall, that ε_t is a linear function of current and past x_t's:

$$\varepsilon_t = x_t - P\left[x_t | x_{t-1}, x_{t-2}, ...\right],$$

so that

$$P\left[\varepsilon_t | x_t, x_{t-1}, ...\right] = \varepsilon_t.$$

To see why this is so, consider the optimization problem that defines a projection:

$$\min_{\{\psi_j\}_{j=0}^{\infty}} E\left(\varepsilon_t - \sum_{j=0}^{\infty} \psi_j x_{t-j}\right)^2.$$

By choosing the ψ_j's to coincide with the linear function, $x_t - P\left[x_t | x_{t-1}, x_{t-2}, ...\right]$, this criterion can be set to zero, which cannot be improved upon. A similar argument establishes

$$P\left[\varepsilon_t | x_{t+j}, x_{t+j-1}, ...\right] = \varepsilon_t, \; j \geq 0.$$

With the previous result, we can write (1.8) as follows:

$$P\left[\eta_t | x_{t-1}, x_{t-2}, ...\right] = P\left[x_t | x_{t-1}, x_{t-2}, ...\right] - \sum_{j=1}^{\infty} d_j \varepsilon_{t-j}.$$

Subtract this from (1.7):

$$\begin{aligned} & \eta_t - P\left[\eta_t | x_{t-1}, x_{t-2}, ...\right] \\ = \; & x_t - P\left[x_t | x_{t-1}, x_{t-2}, ...\right] \\ & - \left(\overbrace{\sum_{j=0}^{\infty} d_j \varepsilon_{t-j}}^{\tilde{x}_t} - \sum_{j=1}^{\infty} d_j \varepsilon_{t-j} \right) \\ = \; & \varepsilon_t - d_0 \varepsilon_t = 0. \end{aligned}$$

This establishes the $s = 0$ part of (1.3). Now consider $s = 1$:

$$\begin{aligned} P\left[\eta_t | x_{t-2}, x_{t-3}, ...\right] &= P\left[x_t | x_{t-2}, x_{t-3}, ...\right] - \sum_{j=0}^{\infty} d_j P\left[\varepsilon_{t-j} | x_{t-2}, x_{t-3}, ...\right] \\ &= P\left[x_t | x_{t-2}, x_{t-3}, ...\right] - \sum_{j=2}^{\infty} d_j \varepsilon_{t-j}, \end{aligned}$$

by an argument similar to the one for $s = 0$. Subtract the above expression from (1.7):

$$\begin{aligned} & \eta_t - P\left[\eta_t | x_{t-2}, x_{t-3}, ...\right] \\ = \; & x_t - P\left[x_t | x_{t-2}, x_{t-3}, ...\right] - \left(\sum_{j=0}^{\infty} d_j \varepsilon_{t-j} - \sum_{j=2}^{\infty} d_j \varepsilon_{t-j} \right) \\ = \; & x_t - P\left[x_t | x_{t-2}, x_{t-3}, ...\right] - \left(\varepsilon_t + d_1 \varepsilon_t\right). \end{aligned}$$

We use the recursive property of projections to evaluate the two-step-ahead forecast error in x_t :

$$\begin{aligned} P\left[x_t | x_{t-1}, x_{t-2}, ...\right] = \; & P\left[x_t | x_{t-2}, x_{t-3}, ...\right] \\ & + P\left[x_t - P\left(x_t | x_{t-2}, x_{t-3} ...\right) | x_{t-1} - P\left(x_{t-1} | x_{t-2}, ...\right)\right]. \end{aligned}$$

In words, the projection of x_t onto x_{t-1} and earlier x_t's is the projection of x_t onto x_{t-2} and earlier plus the best (linear) guess of what that projection error is, given the new information in x_{t-1}. Write the last term in the recursive representation as:

$$P\left[x_t - P\left(x_t|x_{t-2}, x_{t-3}...\right)|x_{t-1} - P\left(x_{t-1}|x_{t-2}, ...\right)\right] = \alpha\varepsilon_{t-1},$$

since $\varepsilon_{t-1} = x_{t-1} - P\left(x_{t-1}|x_{t-2}, ...\right)$. Then,

$$\begin{aligned} \alpha &= \frac{E\left[x_t - P\left(x_t|x_{t-2}, x_{t-3}...\right)\right]\varepsilon_{t-1}}{E\varepsilon_{t-1}^2} \\ &= \frac{Ex_t\varepsilon_{t-1}}{E\varepsilon_{t-1}^2} = d_1. \end{aligned}$$

because ε_{t-1} is orthogonal to $x_{t-2}, x_{t-3}, ...$. So,

$$P\left[x_t|x_{t-1}, x_{t-2}, ...\right] = P\left[x_t|x_{t-2}, x_{t-3}, ...\right] + d_1\varepsilon_{t-1}$$

and

$$\begin{aligned} x_t - P\left[x_t|x_{t-2}, x_{t-3}, ...\right] &= x_t - P\left[x_t|x_{t-1}, x_{t-2}, ...\right] + d_1\varepsilon_{t-1} \\ &= \varepsilon_t + d_1\varepsilon_t. \end{aligned}$$

We conclude that

$$\begin{aligned} &\eta_t - P\left[\eta_t|x_{t-2}, x_{t-3}, ...\right] \\ =\ & x_t - P\left[x_t|x_{t-2}, x_{t-3}, ...\right] - \left(\varepsilon_t + d_1\varepsilon_t\right) \\ =\ & 0. \end{aligned}$$

A continuation of this line of argument establishes that

$$\eta_{t+s} = P\left[\eta_{t+s}|x_{t-2}, x_{t-3}, ...\right],\ s > 0.$$

2. Discussion

The Wold representation says that a covariance stationary process can be represented in the following form:

$$x_t = \underbrace{\sum_{j=0}^{\infty} d_j\varepsilon_{t-j}}_{\text{part of } x_t \text{ that is impossible to predict perfectly}} + \underbrace{\eta_t}_{\text{part of } x_t \text{ that is perfectly predictable}}$$

The two parts of this representation are called the 'purely indeterministic' and the 'deterministic' parts, respectively. It is interesting to evaluate the meaning of η_t. It is not a time trend, for example, because the assumption of covariance stationarity of x_t rules out a time trend.[2] Here is an example of what η_t could be:

$$\eta_t = a\cos\left(\lambda t\right) + b\sin\left(\lambda t\right),$$

[2] The presence of a time trend would imply that the mean of x_t is a function of t.

where λ is a fixed number and

$$Ea = Eb = Eab = 0, \ a, b \perp \{\varepsilon_t\}.$$

To understand this stochastic process for x_t, think of how each realization is constructed. First, draw a and b. Then draw and infinite sequence of ε_t's and generate a realization of η_t and x_t. For the second realization, draw a new a and b, and a new sequence of ε_t's. In this way, all the realizations of the stochastic process may be drawn. Under this representation, the mean and autocovariance function of x_t are not a function of time, and so x_t is covariance stationary.

The idea that η_t is perfectly predictable can be seen as follows. First, a and b can be recovered given only two observations on η_t. Once a and b for a given realization of the stochastic process are in hand, all the η_t's in that realization can be computed. But, how to get the two η_t's? According to the argument in the proof, η_t can be recovered without error from a suitable linear combination of $x_{t-1}, x_{t-2}, ...$ and η_{t+1} can be recovered from a suitable linear combination of $x_t, x_{t-1}, ...$.

It is interesting to compare the purely indeterministic part of the Wold representation with the MA(∞) representations we have discussed in class. The models of MA(∞) representations are in their most general form, ARMA(p,q) representations:

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \nu_t + \theta_1 \nu_{t-1} + ... + \theta_q \nu_{t-q},$$

where ν_t is an *iid* process. As long as the roots of the autoregressive part of this process are less than unity in absolute value, y_t has an MA(∞) representation with square summable moving average terms. Still, there are two possible differences between this and a Wold representation. First, only if the roots of the moving average part, i.e., the zeros of

$$\lambda^q + \theta_1 \lambda^{q-1} + ... + \theta_q$$

are less than unity in absolute value is ν_t the one-step-ahead forecast error in y_t (to see this, note than only in this case can recursive substitution be done to represent ν_t as a function of current and all past y_t's).

Second, the ARMA(p,q) form, while it generates an MA(∞) with square summable weights, it is not the only form that does this. This is perhaps obvious when we observe that the rate of decay of the moving average coefficients in the models we have considered is geometric. This is a faster rate of decay than is required for square summability. For example, with geometric decay absolute and square summability are the same thing. But, in general, a process that is square summable is not necessarily absolutely summable.[3]

We can think of the weight on distant past ε_t's of the MA(∞) representation as corresponding to the amount of 'memory' in the process. Thus, the ARMA(p,q) models have 'short memory' relative to the entire class representations envisioned by the Wold representation. It has been argued that there is evidence of long-memory in economic time series, and that this warrants investigating a class of time series models different from ARMA models. See, for example, Parke (1999). [4] We will not be studying long memory processes in this course.

[3] For example, consider $\psi_j = 1/j$. The rate of decay of ψ_j^2 is fast enough that $\{\psi_j\}$ satisfies square summability, but $\{\psi_j\}$ does not satisfy absolute summability.

[4] Here are some sources cited in Parke (1999), footnote 2. Evidence of long memory has been found in

References

[1] Andersen, Torben G. and Tim Bollerslev, “Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long-Run in High Frequency Returns,” Journal of Finance, 52 (1997), 975-1005.

[2] Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys, “The Distribution of Exchange Rate Volatility,” working paper (1999).

[3] Baillie, Richard T., "Long Memory Processes and Fractional Integration in Economics," Journal of Econometrics, 73 (1996), 5-59.

[4] Baillie, Richard T., Tim Bollerslev, and Hans Ole Mikkelsen, “Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity,” Journal of Econometrics, 74 (1996), 3-30.

[5] Baillie, Richard T., Ching-Fan Chung, and Margie A. Tieslau, “Analyzing Inflation by the Fractionally Integrated ARFIMA-GARCH Model,” Journal of Applied Econometrics, 11 (1996), 23-40.

[6] Breidt, F. Jay, Nuno Crato, and Pedro de Lima, “The Detection and Estimation of Long Memory in Stochastic Volatility,” Journal of Econometrics, 83 (1998), 325-348.

[7] Diebold, Francis X. and Glenn D. Rudebusch, "Long Memory and Persistence in Aggregate Output," Journal of Monetary Economics, 24 (1989), 189-209.

[8] Ding, Zhuanxin, Clive W.J. Granger, and Robert F. Engle, “A Long Memory Property of Stock Market Returns and a New Model,” Journal of Empirical Finance, 1 (1993), 83-106.

[9] Geweke, John and Susan Porter-Hudak, "The Estimation and Application of Long Memory Time Series Models," Journal of Time Series Analysis, 4 (1983), 221-238.

[10] Parke, William, 1999, ‘What is Fractional Integration?’, Working Paper 99-01, Department of Economics, University of North Carolina, Chapel Hill.

[11] Sargent, Thomas, 1979, Macroeconomic Theory.

[12] Sowell, Fallow, “Modeling Long-Run Behavior with the Fractional ARIMA Model,” Journal of Monetary Economics, 29 (1992), 277-302.

traditional business cycle indicators such as aggregate economic activity [Diebold and Rudebusch (1989), Sowell (1992)] and prices indices [Geweke and Porter-Hudak (1983), Baillie, Chung, and Tieslau (1996)]. There is also strong evidence of long memory in asset price and exchange rate volatility [Andersen and Bollerslev (1997), Andersen, Bollerslev, Diebold, and Labys (1999), Baillie, Bollerslev, and Mikkelsen (1996), Breidt, Crato, and Lima (1998), Ding, Granger, and Engle (1993)]. Baillie (1996) provides an excellent survey of the literature on fractional integration and long memory.