Univariate Time Series Models
Econ 241b, Fall 2007

T. Rothenberg

Contents

1 Discrete-Time Stationary Processes

1.1 Introduction

A discrete-time random process is a doubly infinite sequence of random variables $\{y_t\}$ defined for all integer t. Such a process is said to be *strictly stationary* if, for any integer r, the joint distribution of any finite subset, say $y_a, y_b, ..., y_k$, is the same as that of $y_{a+r}, y_{b+r}, ..., y_{k+r}$. This is often expressed loosely by the phrase "calendar time does not matter." The sequence is said to be *weakly* (or covariance) stationary if each y_t has the same mean and if the covariance between each pair y_t and y_s depends only on the time difference $|t - s|$. Strict stationarity implies weak stationarity as long as the second moments exist. Unless otherwise stated, "stationarity" will be used in these notes to mean weak stationarity.

Stationarity puts a great deal of structure on a time series. To characterize the second-order moments of an arbitrary finite random sequence $y_1, ..., y_T$, we need to specify T means, T variances, and $T(T-1)/2$ covariances. Stationarity reduces this to one mean, one variance, and $T - 1$ covariances. More generally, the second-order moment properties of a weakly stationary process are completely described by its mean and its autocovariances $\gamma_r = E(y_t y_{t+r})$ for all nonnegative integer r. (Note $\gamma_{-r} = \gamma_r$.) It is often more convenient to present this information in the form of the mean, the variance γ_0, and the autocorrelations $\rho_r = \gamma_r/\gamma_0$ for $r > 0$. To simplify the notation, we shall typically assume that the mean is zero.

The weakly stationary process $\{\varepsilon_t\}$ is said to be *white noise* if the ε_t are uncorrelated with mean zero and finite variance σ^2. An infinite sequence $\{y_t\}$ is said to be a *time-invariant linear process* if it can be expressed in the form

$$y_t = \sum_{j=-\infty}^{\infty} c_j \varepsilon_{t-j}$$

for some white noise process $\{\varepsilon_t\}$ and some infinite sequence of real numbers $\{c_j\}$ that are square summable; that is $\sum c_j^2 < \infty$. Then y_t is well defined in the sense of mean square convergence of the infinite sum. Usually we will make the stronger assumption that the $c's$ are absolutely summable: $\sum |c_j| < \infty$. Using the lag operator

$$C(L) = \sum_{j=-\infty}^{\infty} c_j L^j$$

where L is defined by $Lx_t \equiv x_{t-1}$, $L^{-1}x_t \equiv x_{t+1}$, and $L^0 x_t = x_t$, we can write more succinctly $y_t = C(L)\varepsilon_t$. $C(L)$ is sometimes called a *linear filter*. We will usually set $c_0 = 1$. Linear processes with absolutely summable coefficients are always weakly stationary with autocovariances given by

$$\gamma_r = \sigma^2 \sum_{j=-\infty}^{\infty} c_j c_{j+r} \ .$$

These autocovariances will also be absolutely summable. If $c_j = 0$ when $j < 0$ so $C(L)$ is a polynomial in L, $C(L)$ is said to be a *one-sided backwards* filter; $\{y_t\}$ is then called a *moving average process* and is sometimes said to be *causal*. If $C_q(L)$ is a polynomial of finite order q, the moving average process $y_t = C_q(L)\varepsilon_t$ is said to be an MA(q) process. Its autocovariances are zero after lag q.

There are generally many different sequences $\{c_j\}$ that can generate the same valid autocovariance sequence. For example, it is easily verified that the two MA(1) models

$$
\begin{aligned}
y_t &= (1 + \theta L)\varepsilon_t & \{\varepsilon_t\} \text{ white noise with variance } \sigma^2 \\
y_t &= (1 + \frac{1}{\theta}L)\eta_t & \{\eta_t\} \text{ white noise with variance } \sigma^2\theta^2
\end{aligned}
$$

have the same autocovariances, so the MA representation is not unique when $|\theta| \neq 1$. In general we have the following result: As long as $c_0 \neq 0$, the polynomial $C_q(L)$ can always be factored as

$$C_q(L) = (\lambda_1 - L)(\lambda_2 - L) \cdots (\lambda_q - L)\frac{c_0}{\lambda_1 \cdots \lambda_q}$$

where the λ's are solutions of the equation $C_q(\lambda) = 0$. Let $C_q^*(L)$ be the polynomial obtained by replacing one or more of the λ's by its reciprocal. Then there exists a positive scalar ω such that the models

$$
\begin{aligned}
y_t &= C_q(L)\varepsilon_t & \{\varepsilon_t\} \text{ white noise with variance } \sigma^2 \\
y_t &= C_q^*(L)\eta_t & \{\eta_t\} \text{ white noise with variance } \omega\sigma^2
\end{aligned}
$$

have the same autocovariances.

1.2   Autoregressive Models

A common approach to modelling assumes that the observable time series $\{y_t\}$ is a stationary solution of a stochastic difference equation. The AR(1) model where, for all integer t, the y's satisfy $y_t = \alpha y_{t-1} + \varepsilon_t$ for white noise $\{\varepsilon_t\}$ is a simple example. Sometimes the difference equation describes a behavioral or technological relation or has been obtained by algebraic manipulation from such a relation; more commonly, the equation is simply a convenient way to describe a series with a particular autocorrelation pattern. These notes discuss some statistical issues for a general class of difference equation models.

If $|\alpha| < 1$ in the above AR(1) model, we obtain by recursive substitution the moving average

$$y_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}. \tag{1}$$

Since the moving average coefficients are square summable, this is a well-defined stationary solution to the difference equation. This representation can also be obtained using lag operators. Writing the difference equation as $(1 - \alpha L)y_t = \varepsilon_t$, multiplying both sides by $1/(1 - \alpha L)$, and using its power series expansion, we find

$$y_t = \frac{1}{1 - \alpha L}\varepsilon_t = (1 + \alpha L + \alpha^2 L^2 + \cdots)\varepsilon_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}.$$

From difference-equation theory, we know that any time series $\{y_t\}$ of the form

$$y_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j} + A\alpha^t$$

for some constant A is a solution to $y_t = \alpha y_{t-1} + \varepsilon_t$. This is a well-defined stationary process only if $A = 0$ and $|\alpha| < 1$.

If $|\alpha| > 1$, the infinite sum (1) is not the mean-square limit of a sequence of random variables and does not constitute a meaningful solution to the difference equation. But, setting $\beta = 1/\alpha$, we find that the difference equation has the well-defined stationary solution

$$y_t = \frac{1}{1 - \alpha L}\varepsilon_t = -\frac{1}{L}\frac{\beta}{1 - \beta L^{-1}}\varepsilon_t = -\frac{\beta}{L}(1 + \beta L^{-1} + \beta^2 L^{-2} + \cdots)\varepsilon_t = -\sum_{j=1}^{\infty} \beta^j \varepsilon_{t+j},$$

3

a convergent *forward* filter of $\{\varepsilon_t\}$. (This solution could also be obtained by recursive substitution using the implied forward difference equation $y_t = \beta y_{t+1} - \beta \varepsilon_{t+1}$.) Note that now y_{t-1} is correlated with ε_t so the equation $y_t = \alpha y_{t-1} + \varepsilon_t$ does not represent a linear regression.

Thus, as long as $|\alpha| \neq 1$, there exists a unique stationary solution to the difference equation $y_t = \alpha y_{t-1} + \varepsilon_t$. The autocovariances for this stationary AR(1) process are $\gamma_r = \alpha^{|r|} \sigma_\varepsilon^2 / (1 - \alpha^2)$ if $|\alpha| < 1$ and $\gamma_r = \alpha^{-|r|-2} \sigma_\varepsilon^2 / (1 - \alpha^{-2})$ if $|\alpha| > 1$. This implies that, if $|\alpha|$ is neither zero nor one, the two models

$$
\begin{aligned}
y_t &= \alpha y_{t-1} + \varepsilon_t, && \{\varepsilon_t\} \text{ white noise with variance } \sigma^2 \\
y_t &= \alpha^{-1} y_{t-1} + \eta_t, && \{\eta_t\} \text{ white noise with variance } \sigma^2/\alpha^2
\end{aligned}
$$

have stationary solutions with the same second-order moments for the observable data and are in this sense equivalent models. We shall normally work with the model whose autoregressive coefficient is less than one in absolute value so a backward moving average representation exists and the "error" will be uncorrelated with the "regressor." Indeed, many authors reserve the words "stationary AR(1) model" for this invertible case.

Let

$$
A_p(L) = 1 - a_1 L - \cdots - a_p L^p
$$

be a polynomial of order p in the lag operator L so that

$$
A_p(L) y_t = y_t - a_1 y_{t-1} - \cdots - a_p y_{t-p}.
$$

Then, a pth order autoregressive model can be written concisely as $A_p(L) y_t = \varepsilon_t$. If there exists a lag polynomial $C(L)$ (possibly of infinite order as long as it contains only nonnegative powers of L and the c_j are square summable) such that $A_p(L) C(L) = 1$, then $C(L) \equiv A_p(L)^{-1}$ is called the inverse of $A_p(L)$. Since $A_p(L)$ can be factored as

$$
A_p(L) = (\lambda_1 - L)(\lambda_2 - L) \cdots (\lambda_p - L) \frac{1}{\lambda_1 \cdots \lambda_p}
$$

where the λ's are the p (possibly complex) roots of the polynomial equation $A_p(\lambda) = 0$, $A_p(L)$ has an inverse if and only if all these roots lie outside the unit circle in the complex plane. If $A_p(L)$ is invertible, the difference equation $A_p(L) y_t = \varepsilon_t$ can be solved to express y_t as an infinite moving average $y_t = A_p^{-1}(L) \varepsilon_t$, where

$$
A_p^{-1}(L) = \prod_{j=1}^{p} (1 - \lambda_j^{-1} L)^{-1} = \prod_{j=1}^{p} (1 + \lambda_j^{-1} L + \lambda_j^{-2} L^2 + \cdots) = 1 + c_1 L + c_2 L^2 + \cdots
$$

This is the unique stationary solution to the difference equation. (Note that if any of the λ's are inside the unit circle, replacing them with their reciprocals in the factorization yields an invertible polynomial with the same autocorrelation function as before. Since the moving average representation is useful for developing forecasts and estimates, we will always use the invertible version of $A_p(L)$. Again, many authors call this version "the" stationary AR(p) model. Others call it a "causal" AR model. If any root lies on the unit circle, no stationary solution of the difference equation exists.)

The autocorrelation function for an AR(p) process with invertible lag polynomial can be found by solving a difference equation. Suppose

$$
y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \varepsilon_t. \tag{2}
$$

Since we are assuming invertibility, y_t has a moving average representation and $\mathrm{E}y_t\varepsilon_s = 0$ when $t < s$. Then, for positive integer r, multiplying both sides of (2) by y_{t-r}, taking expectations, and dividing by γ_0, we find the $Yule - Walker$ equations

$$\rho_r = \alpha_1\rho_{r-1} + \alpha_2\rho_{r-2} + \cdots + \alpha_p\rho_{r-p}.$$

Thus the autocorrelations satisfy the homogeneous pth order difference equation $A_p(L)\rho_t = 0$. The initial conditions are given by the facts $\rho_0 = 1$, $\rho_1 = \rho_{-1}, \ldots$The high-order autocorrelations ultimately die off exponentially to zero.

For any weakly stationary process $\{y_t\}$, let V_r be the $r \times r$ covariance matrix for $(y_1, \ldots, y_r)$ and let C_r be the r-dimensional column vector $(\gamma_1, \ldots, \gamma_r)'$. Then the rth *partial autocorrelation* is defined as the rth element of $V_r^{-1}C_r$. The first partial autocorrelation coefficient is just the first ordinary autocorrelation coefficient ρ_1. For the AR(p) process (2), the pth partial autocorrelation coefficient is α_p and the higher order coefficients are zero.

1.3   Beveridge-Nelson Decompositions

Any equation relating variables at different dates can be rewritten so that some level variables are replaced by first-differenced variables. Suppose, for example, $\{y_t\}$ is a stationary MA(2) process so $y_t = \varepsilon_t + c_1\varepsilon_{t-1} + c_2\varepsilon_{t-2}$. By rearranging terms, it is easily verified that this equation can be rewritten as

$$y_t = (1 + c_1 + c_2)\varepsilon_t - (c_1 + c_2)\triangle\varepsilon_t - c_2\triangle\varepsilon_{t-1}.$$

More generally, suppose $y_t = C(L)\varepsilon_t$ where $C(L) = \sum_{i=0}^{\infty} c_iL^i$. Then the MA representation can be rewritten as

$$y_t = C(1)\varepsilon_t + C^*(L)\triangle\varepsilon_t$$

where $C^*(L) = \sum_{i=0}^{\infty} c_i^*L^i$ and $c_i^* = -\sum_{j=i+1}^{\infty} c_j$. The coefficients c_i^* are absolutely summable as long as $\sum_{i=0}^{\infty} i|c_i| < \infty$. Defining $v_t = C^*(L)\varepsilon_t$, this implies that y_t has the representation

$$y_t = c\varepsilon_t + \triangle v_t$$

where $c = C(1)$ and v_t is a stationary process. This representation is often called the Beveridge-Nelson decomposition of the MA process and is very useful when deriving asymptotic properties of sample moments. For example, suppose one wanted to show that the standardized sample average of the dependent y data is asymptotically normal. Since

$$T^{-1/2}\sum_{t=1}^{T} y_t = cT^{-1/2}\sum_{t=1}^{T} \varepsilon_t + T^{-1/2}(v_T - v_0)$$

and the final term converges in probability to zero as $T \to \infty$, one only need verify that the standardized average of the ε's is asymptotically normal. If the ε's are i.i.d. with finite variance, this follows from a standard central limit theorem.

A similar decomposition holds for AR processes. Suppose $y_t = \alpha_1y_{t-1} + \alpha_2y_{t-2} + \varepsilon_t$. It is easy to verify that this equation can be rewritten as $\triangle y_t = (\alpha_1 + \alpha_2 - 1)y_{t-1} - \alpha_2\triangle y_{t-1} + \varepsilon_t$. More generally, if $A(L)y_t = \varepsilon_t$, where $A(L) = 1 - \sum_{i=1}^{\infty} a_iL^i$ and $\sum_{i=1}^{\infty} i|a_i| < \infty$, then y_t has the alternative representation

$$\triangle y_t = -A(1)y_{t-1} + A^*(L)\triangle y_{t-1} + \varepsilon_t$$

where $A^*(L) = \sum_{i=1}^{\infty} a_i^*L^i$, $a_i^* = -\sum_{j=i+2}^{\infty} a_j$ and $\sum_{i=1}^{\infty} |a_i^*| < \infty$. The usefulness of this representation will become apparent later when we consider unit root models.

1.4 Stationary ARMA Models

Let $A_p(L) = 1 - a_1 L - \cdots - a_p L^p$ and $B_q(L) = 1 + b_1 L + \ldots + b_q L^q$ be polynomials in L having no common factors. Let $\{\varepsilon_t\}$ be white noise. A time series $\{y_t\}$ satisfying the difference equation

$$A_p(L)y_t = B_q(L)\varepsilon_t$$

is called a zero-mean ARMA(p,q) process. A unique stationary solution exists as long as none of the roots of the equation $A_p(z) = 0$ lie on the unit circle. If all the roots lie strictly outside the unit circle, then $A_p(L)$ has an inverse and the series has a moving average representation $y_t = A_p(L)^{-1} B_q(L)\varepsilon_t$. If $B_q(L)$ is invertible, y_t also has a (possibly infinite) autoregressive representation $B_q(L)^{-1} A_p(L)y_t = \varepsilon_t$. If some of the roots of either $A_p(z) = 0$ or $B_q(z) = 0$ are inside the unit circle, they can be replaced by their reciprocals in the factorization of $A_p(L)$ or $B_q(L)$ without changing the autocorrelations of the $\{y_t\}$ process. Thus, as long as neither polynomial has a unit root, we will choose the invertible versions; the unique stationary solution to the difference equation will then have both an autoregressive and moving average representation. Indeed, ARMA models can be motivated by the assumption that $\{y_t\}$ has infinite AR and MA representations

$$y_t = C(L)\varepsilon_t \qquad \text{and} \qquad D(L)y_t = \varepsilon_t$$

where the polynomial $C(L)$ can be well approximated by $B_q(L)/A_p(L)$ and the polynomial $D(L)$ can be well approximated by $A_p(L)/B_q(L)$.

If one or more roots of $A_p(L) = 0$ lie on the unit circle, then there is no stationary solution to the difference equation. If those roots are real and positive, then taking differences will produce a stationary series. Unit roots of $B_q(L) = 0$ do not affect the existence of a stationary solution to the difference equation, but typically indicate that the time series $\{y_t\}$ has been overdifferenced. In an influential book, Box and Jenkins argue that, after perhaps taking first or second differences, many real-world time series can be well approximated as stationary invertible ARMA(p,q) processes with small values for p and q. If d differences are needed to produce stationarity, the process is denoted ARIMA(p,d,q). If $\{y_t\}$ is ARIMA (p,d,q), then $\{\Delta^d y_t\}$ is ARMA(p,q); hence, properties of ARIMA processes can easily be derived from the properties of ARMA processes.

The autocovariances for an ARMA(p,q) process with invertible lag polynomials can be found as follows. Since $\mathrm{E}y_t\varepsilon_s = 0$ when $t < s$, if we multiply both sides of

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \tag{3}$$

by the term y_{t-r} and take expectations, we find

$$\gamma_r = \alpha_1 \gamma_{r-1} + \alpha_2 \gamma_{r-2} + \cdots + \alpha_p \gamma_{r-p} \qquad r > q \tag{4}$$

$$\gamma_r = \alpha_1 \gamma_{r-1} + \alpha_2 \gamma_{r-2} + \cdots + \alpha_p \gamma_{r-p} + \beta_r \mathrm{E}y_{t-r}\varepsilon_{t-r} + \cdots + \beta_q \mathrm{E}y_{t-r}\varepsilon_{t-q} \qquad 0 \le r \le q$$

Thus the autocovariances of order $q + 1$ and higher satisfy the same difference equation as the autocovariances of an AR(p) process. The high-order autocovariances ultimately die off exponentially to zero. The low-order autocovariances can be found by multiplying both sides of equation (3) by ε_{t-r} for $r = 0, .., q$ and taking expectations. The resulting equations can be used with (4) above to solve for the γ's.

ARMA models arise quite naturally from the aggregation of AR models. Suppose $A_p(L)x_t = u_t$ and $B_q(L)z_t = v_t$ where u_t and v_t are independent white noise processes. Then, if the lag polynomials are invertible,

$$y_t \equiv x_t + z_t = \frac{u_t}{A_p(L)} + \frac{v_t}{B_q(L)} = \frac{B_q(L)u_t + A_p(L)v_t}{A_p(L)B_q(L)}.$$

But $B_q(L)u_t + A_p(L)v_t$ is a moving average process of order $r = \max(p, q)$ and $A_p(L)B_q(L)$ is a lag polynomial of order $s = p + q$. Hence y_t is an ARMA(s,r) process. More generally, a linear combination of AR processes will be an ARMA process.

1.5   Nonstationary ARMA Models

The discussion in the previous paragraphs assumes the equation $A_p(L)y_t = B_q(L)\varepsilon_t$ holds for all integer t. This is unnecessarily strong given that we often only need to model an observed finite-length series $y_1, ..., y_T$. An alternative is to assume that the difference equation holds only for $t = p + 1, ..., T$ and to make specific assumptions on the initial random variables $y_1, ..., y_p$. It can be shown that, as long as $A(L)$ has an inverse, there always exist initial conditions such that the observed series is stationary and has the same autocovariances (up to order $T - 1$) as that generated by the difference equation with infinite history.

Of course, we do not have to assume stationarity. Sometimes it is convenient to assume special initial conditions that may introduce a little nonstationarity but simplify forecasting and estimation. Or, if the difference equation arises from some physical experiment as in many engineering applications, the initial conditions may describe the actual state of the system when the experiment started. With arbitrary initial conditions, the nonstationary ARMA process is well defined even if the autoregressive lag polynomial has unit roots.

1.6   State-Space Models

Suppose the scalar process $\{y_t\}$ is generated as

$$y_t = b'\alpha_t + u_t \qquad t = 1, 2, ...$$

where $\{\alpha_t\}$ is an *unobserved* p-dimensional vector process generated by the first-order difference equation

$$\alpha_t = A\alpha_{t-1} + v_t \qquad t = 1, 2, ...$$

A is a nonrandom square matrix, b is a nonrandom vector, $\{v_t\}$ is a vector white noise process,and $\{u_t\}$ is a scalar white noise process independent of $\{v_t\}$ and the initial value α_0. The α_t are called state variables and the process $\{y_t\}$ is said to have a state-space representation. It turns out that every ARMA process has a state-space representation. Thus the class of state space models includes the class of ARMA models. As we shall show later, some results about ARMA models can be derived most easily as special cases of results for state-space models.

State-space models are sometimes called *hidden Markov models* in the statistical literature. The unobserved state variables α_t are Markovian; that is, the conditional distribution of α_t given its past history depends only on the most recent past history α_{t-1}. Although the observed variables y_t are not Markovian, they inherit some of the simplicity of Markov processes. In particular, general algorithms have been developed for computing the likelihood function for normal state-space models.

1.7   Long Memory Processes

In contrast to ARMA processes whose autocorrelations ultimately drop off to zero at a geometric rate, *long memory processes* have autocorrelations that drop off much more slowly. An example is the fractionally differenced process generated by the stochastic difference equation

$$(1 - L)^d y_t = \varepsilon_t \tag{5}$$

where $\{\varepsilon_t\}$ is white noise, $d \in (-0.5, 0.5)$, and the fractional difference operator $(1 - L)^d$ is defined by its power series expansion

$$(1 - L)^d = 1 - dL + \frac{1}{2}d(d-1)L^2 + \cdots .$$

It can be shown that equation (5) always has a stationary solution with moving average representation of the form

$$y_t = (1 - L)^{-d}\varepsilon_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}$$

where, for large j and $d \neq 0$,

$$c_j \approx \frac{j^{d-1}}{\Gamma(d)} \qquad \text{and} \qquad \rho_j \approx \frac{\Gamma(1-d)}{\Gamma(d)}j^{2d-1} .$$

When $0 < d < 0.5$, the moving average coefficients $\{c_j\}$ are square summable, but not absolutely summable. The autocorrelation coefficients $\{\rho_j\}$ drop off slowly and are not even square summable when $0.25 \leq d < 0.5$.

A generalization of (5) is the fractionally differenced ARIMA(p,d,q) model

$$A_p(L)(1 - L)^d y_t = B_q(L)\varepsilon_t$$

where A_p and B_q are invertible lag polynomials of finite orders and $d \in (-0.5, 0.5)$. In contrast to the case where d is a positive integer, this difference equation has a stationary solution $\{y_t\}$, although with long memory properties.

2 Prediction

Let Y be a scalar random variable and let x be a (possibly infinite dimensional) vector random variable defined on the same probability space. The best predictor of Y given x is defined to be that function $g^*(x)$ such that $E[Y - g(x)]^2$ is minimized. Assuming that second moments exist, we find $g^*(x) = E(Y|x)$. The best *linear* predictor of Y given x (denoted $P(Y|x)$) is defined to be that scalar $a + b'x$ such that $E[Y - a - b'x]^2$ is minimized. Again assuming second moments exist, we find that the solution must satisfy

$$c_{xy} = V_{xx}b^* \quad \text{and} \quad a^* = EY - b^{*'}Ex$$

where c_{xy} is the column vector of covariances between Y and x and V_{xx} is the covariance matrix for x. If V_{xx} is finite dimensional and nonsingular, we find

$$P(Y|x) = EY + c'_{xy}V_{xx}^{-1}(x - Ex).$$

The mean square prediction error using this optimal linear predictor is

$$\mathrm{E}(Y - P(Y|x))^2 = \operatorname{var}(Y) - c'_{xy}V_{xx}^{-1}c_{xy} .$$

When x is a vector of high dimension, inverting V_{xx} is computationally demanding. Sometimes one can compute (or at least approximate) $P(Y|x)$ without explicitly finding the inverse. The best linear predictor has the following useful properties:

1. The prediction error $e \equiv Y - P(Y|x)$ has mean zero and is uncorrelated with every element of x.

2. For constants a, b, and c, $P(aY + bZ + c|x) = aP(Y|x) + bP(Z|x) + c$.

3. If Y is uncorrelated with each element of x, then $P(Y|x) = EY$.

4. If Y is an element of x or a linear combination of elements of x, then $P(Y|x) = Y$.

5. $P(Y|x) = P[P(Y|x, z)|x]$.

2.1 Prediction of an ARMA Series

Convenient expressions for $P(Y|x)$ can be found in the special case where Y and x are realizations at different dates of the stationary ARMA(p,q) process $A(L)y_t = B(L)\varepsilon_t$. In particular, for any integer $s > 0$, let $P_T(y_{T+s})$ be the best linear predictor of y_{T+s} given the infinite history $y_T, y_{T-1}, y_{T-2}, ...$. Suppose $A(L)$ and $B(L)$ are both invertible. Then y_{T+s} can be written as a moving average

$$y_{T+s} = \frac{B(L)}{A(L)}\varepsilon_{T+s} = \varepsilon_{T+s} + c_1\varepsilon_{T+s-1} + \cdots + c_s\varepsilon_T + c_{s+1}\varepsilon_{T-1} + \cdots$$

The best forecast of future ε's is zero. Current and past ε's can be perfectly predicted from the autoregressive representation. Thus

$$P_T(y_{T+s}) = c_s\varepsilon_T + c_{s+1}\varepsilon_{T-1} + \cdots = \left(\frac{B(L)}{A(L)L^s}\right)_+ \varepsilon_T = \left(\frac{B(L)}{A(L)L^s}\right)_+ \frac{A(L)}{B(L)}y_T$$

where $D(L)_+$ is defined to be the lag polynomial $D(L)$ with terms containing negative powers of L dropped.

When $s = 1$, this expression simplifies since

$$\left(\frac{B(L)}{A(L)L}\right)_+ = \frac{1}{L}\left[\frac{B(L)}{A(L)} - 1\right] = \frac{B(L) - A(L)}{A(L)L}.$$

We find

$$P_T(y_{T+1}) = \frac{B(L) - A(L)}{B(L)L}y_T$$

which implies $P_T(y_{T+1}) = y_{T+1} - \varepsilon_{T+1}$. That is, if $\{y_t\}$ is the stationary solution of an ARMA model with invertible lag polynomials, ε_t is the difference between y_t and the best linear predictor of y_t given its past history. The ε_t are often called *innovations* and interpreted as new information entering the system at time t.

Of course, in practice we can only make predictions based on a finite past history. If we truncate the infinite autoregression, it can serve as an approximation to the best linear predictor given a finite past history. As we shall see later, the Kalman filter provides an efficient computer algorithm for computing the best one-step-ahead linear predictor $P(y_{T+1}|y_1, ..., y_T)$ exactly.

2.2 Wold Decomposition Theorem

A stationary process $\{y_t\}$ is said to be *deterministic* if $P_T(y_{T+1}) = y_{T+1}$; that is, if it can be linearly predicted without error from its past history. Clearly, ARMA processes are never deterministic unless the innovation variance is zero. The *Wold decomposition theorem* asserts that every weakly stationary process can be written as the sum of a deterministic process and a moving average process. If the moving average coefficients drop off to zero rapidly after some finite lag, the moving average part can be approximated by an ARMA(p,q) process. Hence this theorem suggests that ARMA models should capture the second-order moment properties of many stationary time series after any perfectly predictable component has been removed. The exceptional cases where the moving average coefficients decline only slowly will be discussed later.

2.3 Signal Extraction

An interesting example of prediction is the so-called *signal extraction problem*. Suppose an observed time series y_t is the sum of a *signal* component u_t and a *noise* component v_t

$$y_t = u_t + v_t$$

where u_t and v_t are independent unobserved stationary processes. Given a realization of the time series $\{y_t\}$, we want to find the best linear predictor of u_t. The answer can be expressed simply using *generating functions.* If $\{x_t\}$ is a stationary time series with absolutely summable autocovariance function $\gamma_x(r)$, its autocovariance generating function is defined to be the linear filter

$$G_x(L) = \sum_{r=-\infty}^{\infty} \gamma_x(r) L^r.$$

The best linear predictor $\hat{u}_t = B(L) y_t$ must satisfy

$$E y_{t-r}(u_t - \sum_i b_i y_{t-i}) = 0 \ \text{ for all integer } r.$$

This implies $\gamma_u(r) = \sum_i b_i \gamma_y(r-i)$ which implies

$$G_u(L) = \sum_{r=-\infty}^{\infty} \gamma_u(r) L^r = \sum_{r=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} b_i L^i \gamma_y(r-i) L^{r-i} = B(L) G_y(L).$$

The best linear predictor is

$$\hat{u}_t = \frac{G_u(L)}{G_y(L)} y_t = \frac{G_u(L)}{G_u(L) + G_v(L)} y_t.$$

Recall that, if $y_t = C(L)\varepsilon_t$, then $\gamma_r = \sigma^2 \sum_j c_j c_{j+r}$. It follows that

$$G_y(L) = \sum_{r=-\infty}^{\infty} \gamma_y(r) L^r = \sigma^2 \sum_{r=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} c_j L^{-j} c_{j+r} L^{j+r} = \sigma^2 C(L^{-1}) C(L).$$

Hence if u_t and v_t are independent ARMA processes $P(L) u_t = Q(L) \varepsilon_t$ and $R(L) v_t = S(L)\eta_t$, the best linear predictor is

$$\hat{u}_t = \frac{\sigma_\varepsilon^2 \frac{Q(L)Q(L^{-1})}{P(L)P(L^{-1})}}{\sigma_\varepsilon^2 \frac{Q(L)Q(L^{-1})}{P(L)P(L^{-1})} + \sigma_\eta^2 \frac{S(L)S(L^{-1})}{R(L)R(L^{-1})}} y_t$$

as long as $P(L)$ and $R(L)$ are invertible.

3 ARMAX Models

The ARMA model can be generalized to allow for the mean of y_t to depend on the exogenous variables $x_{1t}, x_{2t}, ..., x_{Kt}$, possibly with some lag. Many dynamic economic relations can be expressed in the so-called ARMAX form

$$A(L) y_t = \mu + F_1(L) x_{1t} + ... + F_k(L) x_{kt} + B(L)\varepsilon_t$$

where A, B, and the F's are low-order lag polynomials and μ is a scalar intercept.

3.1 Distributed Lags

ARMAX models are commonly employed in the context of distributed lag estimation. For example, suppose some economic variable y moves over time in response to some variable x, but perhaps with some lag. One might write

$$y_t = \alpha + \beta C(L) x_t + u_t, \quad D(L) u_t = \varepsilon_t \tag{6}$$

where the ε_t are white noise and D is an invertible lag polynomial. The first equation represents the (causal) distributed lag relation; the second equation models the autocorrelation structure of the errors (capturing omitted effects). The variable x is assumed to be exogenous; that is, $E[\varepsilon_t|\text{past } y\text{'s and all } x\text{'s}] = 0$. Although such models can be estimated by OLS or GLS, the number of parameters in $C(L)$ is likely to be large and the resulting collinearity will result in large standard errors.

The *rational distributed lag* approach uses ratios of low-order polynomials to approximate the (possibly high order) polynomials $C(L)$ and $D(L)$. If

$$C(L) = \frac{P(L)}{Q(L)} \quad \text{and} \quad D(L) = \frac{A(L)}{B(L)},$$

then we obtain the ARMAX model

$$A(L)Q(L)y_t = \overline{\alpha} + \beta A(L)P(L)x_t + Q(L)B(L)\varepsilon_t. \tag{7}$$

The simplest version of the rational lag model (first proposed by Koyck) assumes no autocorrelation in the residuals and geometrically declining weights in $C(L)$ so $A(L) = B(L) = P(L) = 1$ and $Q(L) = 1 - \lambda L$. Then (7) becomes

$$y_t = \overline{\alpha} + \lambda y_{t-1} + \beta x_t + w_t$$

where $w_t = \varepsilon_t - \lambda \varepsilon_{t-1}$. A least squares regression is not appropriate because y_{t-1} is correlated with w_t. Instead, the parameters are commonly estimated by the method of instrumental variables, using x_{t-1} as instrument for y_{t-1}. This ignores the constraint that the error distribution depends on the coefficient λ.

In general, models like (6) can be estimated by the method of instrumental variables or, more efficiently, by Gaussian maximum likelihood. Unfortunately, consistent estimation of the ARMAX model by either method requires that the lag orders are known. Box and Jenkins suggest possible ways to estimate these orders, but it is not clear that they are very successful.

In practice, we are often unwilling to pretend we know the orders of the lag polynomials. In that case, we might just run an OLS regression of y_t on $(x_t, x_{t-1}, ..., x_{t-p})$ for some fairly large value of p. If the true order of the lag polynomial is less than p, this estimate of $C(L)$ will be unbiased even if the errors are autocorrelated. A look at the residuals might suggest a second-stage GLS estimator. Estimating a long distributed lag by OLS or GLS is likely to lead to collinearity problems. Although this will typically not seriously affect the precision of the estimate of the long-run multiplier, it will lead to large standard errors for individual lag parameters. To lessen collinearity, one can impose constraints on the lag coefficients. Forcing the c_j to lie on a polynomial in j leads to a linear regression with fewer parameters. For example, if $p = 4$ and

$$C(L) = 1 + c_1 L + c_2 L^2 + c_3 L^3 + c_4 L^4,$$

the linear constraint that $c_j = (4-j)/4$ yields a regression (with intercept) of y_t on the single explanatory variable

$$x_t^* = \frac{4x_t + 3x_{t-1} + 2x_{t-2} + x_{t-3}}{4}.$$

Since this polynomial distributed lag approach never introduces lagged y's as regressors, standard linear model theory holds. The main cost is that the first p observations needed to construct x^* are lost; with 50 years of annual data and a lag length of 10, one would lose 20% of the observations.

These approaches to distributed lag estimation assume that the x's are exogenous. The equation is usually given a causal interpretation as the response of some economic variable to

changes in its determinants. How do we know that the direction of causality is not reversed? or that both variables are caused by a third variable? Sims has argued that, before attempting to fit a distributed lag, one should first regress y_t on some past, current and future x's. If the future x's appear with significant coefficients, the assumption that x is exogenous is suspect. We shall discuss such exogeneity tests in more detail when we consider multivariate dynamic models.

3.2 Problems in Interpreting Dynamic Regression Models

Even if we are sure that x is exogenous, there can be conceptual problems in interpreting dynamic economic models. Some of the problems can be illustrated by the following simple example. Suppose one theory is that people set y_t based on its past value y_{t-1} and on the past value of some exogenous variable x_{t-1}. Given y_{t-1} and x_{t-1}, the past is irrelevant. The model is

$$y_t = \alpha y_{t-1} + \beta x_{t-1} + \varepsilon_t$$

where ε_t represent other determinants of y_t and is assumed to be white noise. Another theory is that people set y_t as a geometrically declining weighted average of past x's:

$$y_t = \frac{\beta}{1 - \alpha L} x_{t-1} + u_t$$

where u_t represents other determinants of y_t . If $(1-\alpha L)u_t$ is white noise, the two models are equivalent: it is not possible to tell from the data which theory is correct. This does not matter if our goal is forecasting in a world where the process generating the data has not changed. In both cases, $E(y_t|\text{past data}) = \alpha y_{t-1} + \beta x_{t-1}$. But suppose the system were interrupted (say, due to war) and started over again. Then the two models would predict very different behavior in the second period. Prior knowledge about the form of autocorrelation in the errors can be used to distinguish between economically different models. But in the absence of such knowledge, it is hard to distinguish among alternative behavioral models containing lagged endogenous variables. In the distributed lag model (6), the problem was assumed away by the specification that only lagged x's enter the behavioral equation. In the absence of such assumptions, dynamic ARMAX models are open to many possible interpretations.

4 ARCH and GARCH Models

The models investigated so far are based on second-order moments and are designed primarily to help forecast future values of y_t from its past history. That is, the goal was to specify a simple model for the conditional mean of y_t given its past. Sometimes, however, we may want to base the analysis on higher-order moments. This is particularly the case when the goal is to model the conditional *variance* of y_t given the past. For example, when analyzing stock market returns for the purpose of pricing derivative securities, the Black-Scholes theory requires modelling both the conditional mean and the conditional variance of returns.

An AR(p) model with i.i.d. innovations specifices that the conditional expectation of y_t given past values is a linear function of the most recent p past values but the conditional variance of y_t given past values is constant. If we permit the conditional variance to vary with t we get a model with conditional heteroskedasticity. Suppose, for example,

$$C_p(L)y_t = u_t$$

where $E(u_t|\text{past } y\text{'s}) = 0$ and

$$\text{Var}(y_t|\text{past } y\text{'s}) = \beta + D_q(L)u_{t-1}^2$$

so that past deviations of y_t from its conditional mean help predict future variability of y_t. Then we have a model exhibiting *autoregressive conditional heteroskedasticity* (ARCH). The parameters of this model could be estimated simply (but generally inefficiently) by a two-step procedure. The u_t could be estimated by the residuals from a regression of y_t on p lagged y's; then the remaining parameters could be estimated from a regression of $\hat{u}_t^2$ on q of its lagged values.

More efficient estimates can be obtained if we are willing to make the somewhat stronger assumption that the u_t's can be expressed in terms of an underlying i.i.d. process. Specifically, if we assume that

$$u_t = \sqrt{h_t}\varepsilon_t$$

where $h_t = Var(y_t|\text{past } y\text{'s}) = \beta + D_q(L)u_{t-1}^2$ and the ε_t are i.i.d. with mean zero and variance one, then we get the ARCH process described in the previous paragraph. Moreover, if $f(\varepsilon)$ is the density function for each of the ε_t, the joint density of the observations $y_{p+q+1}, ..., y_T$ (conditional on observations up to period $p + q$) is

$$\prod_{t=p+q+1}^{T} f\left(\frac{C_p(L)y_t}{\sqrt{\beta + D_q(L)[C_p(L)y_{t-1}]^2}}\right) \frac{1}{\sqrt{\beta + D_q(L)[C_p(L)y_{t-1}]^2}}.$$

If the density f is known and we are willing to condition on the first $p + q$ values, the parameters of the model can be estimated by maximum likelihood.

Estimation in a high dimensional parameter space can be avoided by approximating $C_p(L)$ and $D_q(L)$ by ratios of low order polynomials. If, for example, $C_p(L) = A(L)/B(L)$ and $D_q(L) = Q(L)/P(L)$, then we have a generalized ARCH (or GARCH) process for the conditional variance added on to an ARMA model for the mean:

$$
\begin{aligned}
A(L)y_t &= B(L)u_t \\
u_t &= \sqrt{h_t}\varepsilon_t \\
P(L)h_t &= \bar{\beta} + Q(L)u_{t-1}^2
\end{aligned}
$$

Again, under appropriate initial conditions, the parameters can be estimated by maximum likelihood if the density for ε_t is known.