Vector Time-Series Models

1 Introduction

The n-dimensional, mean-zero vector process $\{z_t\}$ is said to be weakly stationary if second-order moments exist and only depend on time differences. These moments are completely described by the doubly-infinite sequence of $n \times n$ autocovariance matrices

$$\Gamma_r = E(z_t z'_{t-r}) \qquad\qquad r = 0, 1, -1, 2, -2, ...$$

Note that $\Gamma_r = \Gamma'_{-r}$ so $\Gamma_0, \Gamma_1, ...$ are sufficient to describe the process.

The weakly stationary, n-dimensional, mean-zero vector process $\{\varepsilon_t\}$ is said to be *white noise* if $\mathrm{E}(\varepsilon_t \varepsilon'_s) = 0$ when $t \neq s$ and $\Sigma = E(\varepsilon_t \varepsilon'_t)$ is positive definite. An infinite sequence of n-dimensional vectors $\{z_t\}$ is said to be a *time invariant linear vector process* if they can be expressed in the form

$$z_t = \sum_{j=-\infty}^{\infty} C_j \varepsilon_{t-j}$$

for some n-dimensional white-noise process $\{\varepsilon_t\}$ and some infinite sequence of real $n \times n$ matrices $\{C_j\}$ whose elements are square summable. Usually we will make the stronger assumption that the elements of C are absolutely summable. Using the lag operator $C(L) = \sum_j C_j L^j$, we can write more succinctly $y_t = C(L)\varepsilon_t$. As in the scalar case, $C(L)$ is sometimes called a *linear filter*; 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 and y_t is called a *moving average process.* If $C_q(L)$ is a polynomial of finite order q, the moving average process $z_t = C_q(L)\varepsilon_t$ is said to be an MA(q) process. Its autocovariances are zero after lag q. Moving average processes with absolutely summable coefficients are always weakly stationary with autocovariances given by

$$\Gamma_r = \sum_{j=0}^{\infty} C_{r+j} \Sigma C'_j \qquad\qquad r \geq 0 \ .$$

These autocovariances will also be absolutely summable. For any nonsingular matrix D, the MA process with lag polynomial $D^{-1}C(L)$ and variance matrix $D\Sigma D'$ has the same autocovariances as the process with lag polynomial $C(L)$ and variance matrix Σ. Thus, without restricting the covariances, we will use the parsimonious representation where $C_0 = I_n$. (Alternatively, one could set $\Sigma = I_n$, but that usually turns out to be less convenient.)

2 Multivariate Spectrum

For an n-dimensional vector process $\{z_t\}$ with absolutely summable autocovariances, we define the matrix *autocovariance generating function* as the lag operator $G(L) = \sum_r \Gamma_r L^r$. The spectrum is defined as the $n \times n$ complex matrix function (for $-\pi < \lambda < \pi$)

$$\mathbf{s}(\lambda) = \frac{1}{2\pi} G(e^{-i\lambda}) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} \Gamma_r e^{-i\lambda r} \ .$$

Since $\Gamma_r = \Gamma'_{-r}$, the diagonal elements of $S(\omega)$ are real and the ij element of $S(\omega)$ is the complex conjugate of the ji element. The inverse relation is

$$\Gamma_r = \int_{-\pi}^{\pi} \mathbf{s}(\omega) e^{i\lambda r} d\lambda.$$

As in the scalar case, the spectral theory can be extended to the case where the autocovariances are not absolutely summable, but we omit the details. If $w_t = C(L)z_t$ for some lag polynomial matrix C with square summable coefficients, then

$$\mathbf{s}_w(\lambda) = C(e^{-i\lambda})\mathbf{s}_z(\lambda)C(e^{i\lambda})'.$$

The spectral representation theorem states that every discrete time vector stationary process can be written as

$$z_t = \int_{-\pi}^{\pi} e^{i\lambda t} dH(\lambda)$$

where $H(\lambda)$ is a continuous *vector* process with uncorrelated increments. When the autocovariances are absolutely summable, $dH(\lambda)$ has variance matrix $\mathbf{s}(\lambda)$. Again, we have the interpretation that z_t can be represented as a sum of sine and cosine functions with random weights; $\mathbf{s}(\lambda)$ measures the size of the weight given to the sinusoidal functions having frequencies in a neighborhood of λ. We shall work out the details only for the case $n = 2$.

Suppose $z_t = (x_t, y_t)'$ is two-dimensional stationary process with spectrum

$$\mathbf{s}(\lambda) = \begin{bmatrix} s_{xx}(\lambda) & s_{xy}(\lambda) \\ s_{yx}(\lambda) & s_{yy}(\lambda) \end{bmatrix}$$

The diagonal elements are just the univariate spectral density functions. The cross-spectrum between y and x is defined as the 21 element

$$s_{yx}(\lambda) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} \gamma_{yx}(r) e^{-i\lambda r}.$$

It is generally complex and can be written as $R(\lambda)e^{i\theta(\lambda)}$, where $R(\lambda) = |s_{yx}(\lambda)|$ and $\tan\theta(\lambda) = \mathrm{Im}(s_{yx}(\lambda))/\mathrm{Re}(s_{yx}(\lambda))$. Three new concepts are often employed in studying the relationship between y and x:

1. The *gain* $|s_{yx}(\lambda)|/s_{xx}(\lambda)$ which measures how the weight $dH_x(\lambda)$ given to the cyclical component of frequency λ in the spectral decomposition of x_t is amplified to obtain the corresponding weight $dH_y(\lambda)$.

2. The *phase* -$\theta(\lambda)$ which measures how the cyclical component in y_t of frequency λ lags the corresponding component in x_t.

3. The *coherence*

$$\frac{|s_{yx}(\lambda)|^2}{s_{xx}(\lambda)s_{yy}(\lambda)}$$

   which measures the correlation between the weights $dH_x(\lambda)$ and $dH_y(\lambda)$.

The interpretation of the gain, phase and coherence is clearest in the special case where y_t is a filtered version of x_t plus error. If $y_t = C(L)x_t + u_t$ where u_t is uncorrelated with x_s for all t and s, then a little algebra shows

$$s_{yy}(\lambda) = s_{xx}(\lambda)|C(e^{-i\lambda})|^2 + s_{uu}(\lambda) \quad \text{and} \quad s_{yx}(\lambda) = C(e^{-i\lambda})s_{xx}(\lambda).$$

If $C(e^{-i\lambda})$ is written as $r(\lambda)e^{i\varphi(\lambda)}$, the gain is just $r(\lambda)$ and the phase is just $-\varphi(\lambda)$. They measure the effect of the filter on the amplitude and phase of the cyclical components of the time-series. Suppose, for example, $x_t = A\cos(\theta t) + B\cos(\theta t)$ where A and B are uncorrelated

random variables with zeri mean and unit variance. Let $y_t = 10x_{t-s} + \varepsilon_t$ where $\{\varepsilon_t\}$ is a unit-variance white noise process, independent of $\{x_t\}$. Then

$$y_t = 10\cos(\theta t - \theta s) + 10\sin(\theta t - \theta s) + \varepsilon_t \ .$$

The spectrum for $\{x_t\}$ is zero except for spikes at $\lambda = \pm\theta$. The spectrum for $\{y_t\}$ is everywhere equal to $(2\pi)^{-1}$ except for additional spikes at $\lambda = \pm\theta$. Since $C(e^{-i\lambda}) = 10e^{-i\lambda s}$, $s_{yx}(\lambda)$ is $10e^{-i\theta s}$ times a spike at $\lambda = \pm\theta$. Thus, at frequency θ, the phase is θs, the amount the filter causes the cycle to be lagged, and the gain of 10 measures the effect of the filter on the amplitutde of the cycle. The coherence at frequency $\pm\lambda$ is one if u_t has zero variance; it drops to zero as σ_ε^2 tends to infinity. At all other positive frequencies, the phase, gain and coherence are undefined because $s_{xx}(\lambda)$ is zero.

For arbitrary stationary time series y_t and x_t, we can always write $y_t = C(L)x_t + u_t$, where $C(L)x_t$ is the best linear predictor of y_t given all past, present, and future values of x_t and the prediction error u_t is by definition uncorrelated with all x_s.Γ Thus the interpretation of the cross spectrum given above works also for the general case.

3 Best Linear Predictors

In the first set of notes, we looked at a simple prediction problem known as singal extraction. Here we present a generealization. Let $\{x_t\}$ be a q-dimensional process and let $\{y_t\}$ be a p-dimensional process. We assume that the joint process $\{x_t, y_t\}$ is stationary. We partition the joint autocovariances and the joint autocovariance generating function as

$$\begin{aligned}
\Gamma_r &= E\begin{pmatrix} x_t \\ y_t \end{pmatrix}\begin{pmatrix} x_{t-r} & y_{t-r} \end{pmatrix} = \begin{pmatrix} \Gamma_{xx}(r) & \Gamma_{xy}(r) \\ \Gamma_{yx}(r) & \Gamma_{yy}(r) \end{pmatrix} \\
G(L) &= \sum_{r=-\infty}^{\infty} \begin{pmatrix} \Gamma_{xx}(r) & \Gamma_{xy}(r) \\ \Gamma_{yx}(r) & \Gamma_{yy}(r) \end{pmatrix} L^r = \begin{pmatrix} G_{xx}(L) & G_{xy}(L) \\ G_{yx}(L) & G_{yy}(L) \end{pmatrix}
\end{aligned}$$

For some $p \times q$ matrix lag operator $D(L) = \sum_{j=-\infty}^{\infty} D_j L^j$, let $D(L)x_t$ be the best linear predictor of y_t given all past, present, and future values of x_t. Since the prediction error $y_t - D(L)x_t$ must be uncorrelated with x_s for all s and t, it follows that $\Gamma_{yx}(r) \equiv E(y_t x'_{t-r}) = \sum_j D_j \Gamma_{xx}(r-j)$ and, hence,

$$G_{yx}(L) = \sum_{r=-\infty}^{\infty} \Gamma_{yx}(r)L^r = \sum_{j=-\infty}^{\infty}\sum_{r=-\infty}^{\infty} D_j L^r \Gamma_{xx}(r-j)L^{r-j} = D(L)G_{xx}(L)$$

so

$$D(L) = G_{yx}(L)G_{xx}^{-1}(L)$$

as long as the inverse exists.

This last result allows us to find a simple expression for the best linear predictor $D(L)x_t$ in terms of the moving average representation of $\{x_t, y_t\}$. Let $n = p + q$. Suppose

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{bmatrix} A(L) \\ B(L) \end{bmatrix} \varepsilon_t \ ,$$

where A is $q \times n$, B is $p \times n$, and $\{\varepsilon_t\}$ is an n-dimensional white noise process with covariance matrix $\mathrm{E}\epsilon_t\epsilon'_t = \Sigma$. Then, a little algebra yields

$$G_{xx}(L) = A(L)\Sigma A(L^{-1})' \quad \text{and} \quad G_{yx}(L) = B(L)\Sigma A(L^{-1})' \ .$$

It follows that

$$D(L) = B(L)\Sigma A(L^{-1})'[A(L)\Sigma A(L^{-1})']^{-1}.$$

In general, $D(L)$ will be a two-sided filter.

4 Vector ARMA Models

4.1 General Case

Let $A_p(L) = I - A_1L - ... - A_pL^p$ and $B_q(L) = I + B_1L + ... + B_qL^q$ be $n \times n$ lag polynomial matrices. Then, if ε_t is a vector white noise with covariance matrix Ω, any stationary solution $\{z_t\}$ of the difference equation

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

is called a vector ARMA(p,q) process. A complex number ξ satisfying the determinantal equation $|A(\xi)| = 0$ is called a root of $A(L)$. If none of the np roots lie on the unit circle, it can be shown that the difference equation has a unique stationary solution. If all the roots lie outside the unit circle, $A(L)$ has an inverse and z_t can be written as a backwards filter of ε_t; that is, $z_t = A_p^{-1}(L)B_q(L)\varepsilon_t$.The spectrum of this invertible ARMA process is given by

$$\mathbf{s}_z(\omega) = \frac{1}{2\pi}A_p^{-1}(e^{-i\omega})B_q(e^{-i\omega})\Omega B_q'(e^{i\omega})A_p^{-1}(e^{i\omega})' \ .$$

Since $[A_p(L)]^{-1}$ can be written as $C(L)/|A_p(L)|$, where C is the matrix of cofactors and $|A|$ is the determinant of A, we have

$$|A_p(L)|z_t = D(L)\varepsilon_t$$

where $D(L) = C(L)B_q(L)$ is a matrix of lag polynomials of maximum lag order $s = (n-1)p+q$. Thus, if there are no canceling common factors, each component of z_t has a univariate ARMA(np, s) representation with *identical* AR polynomial.

4.2 Vector Autoregression

In practice, it seems hard to get reasonable estimates of vector ARMA(p,q) models with limited data unless n, p and q are very small. As a result, most empirical work sets $q = 0$. The resulting vector autoregressive model (often called a VAR) is widely used in applied macroeconomics. Allowing for r additional exogenous variables $x_{1t}, ..., x_{rt}$ such as intercepts and time trends, the model takes the form

$$z_t = A_1z_{t-1} + ... + A_pz_{t-p} + Bx_t + \varepsilon_t. \tag{1}$$

It is assumed that p is large enough so the lagged z's capture all the autocorrelation in the series. In addition, it is usually assumed that the matrices $A_1, ..., A_p$ and B are unrestricted so every lagged variable and exogenous variable appears in every equation with a nonzero coefficient.

If we multiply both sides of equation (1) on the left by the nonsingular matrix A_0 we obtain another equation that generates the same solution $\{z_t\}$. The two models are equivalent and there is no loss in using the parsimonious parameterization. Hence, in defining a VAR model, we set $A_0 = I_p$.

Using the Beveridge-Nelson decomposition, equation (1) can be rewritten as

$$\triangle z_t = -\Pi z_{t-1} + A^*(L)\triangle z_{t-1} + Bx_t + \varepsilon_t \tag{2}$$

where $\Pi = I - A_1 - ... - A_p$ and $A^*(L)$ is a lag polynomial of degree $p-2$. This version is often called the "error correction" form of the model. The name comes from the following interpretation. Suppose the only exogenous variable is an intercept term so Bx_t is just the vector b. One might say that the system is in long-run equilibrium if $\triangle z_t$ has mean zero for all t. This can occur only if $\Pi z_t = b$. If Π is invertible, this means that the model can be rewritten as

$$\triangle z_t = -\Pi(z_{t-1} - \Pi^{-1}b) + A^*(L)\triangle z_{t-1} + \varepsilon_t.$$

The vector $\Pi^{-1}b$ can be interpreted as the long-run equilibrium value for z_t and $\triangle z_t$ is interpreted as responding to the disequilibrium "error" $z_{t-1} - \Pi^{-1}b$. A similar interpretation holds even if Π is not invertible as we shall see when we consider cointegration models.

4.3 Estimation of VAR Parameters

As long as the lag polynomial $I - A_1L - ... - A_pL^p$ is invertible, ε_t is uncorrelated with the right-hand variables. Each of the n equations in (1) can be estimated by an OLS regression with $m = np + r$ regressors and $T^* = T - p$ observations. These will be the conditional Gaussian maximum likelihood estimates, where we condition on the first p observations.

To see this, let Y be the $T^* \times n$ matrix of observations for the "dependent" variables z_t in (1), let X be the $T^* \times m$ matrix of observations for the "explanatory" variables (x's and lagged z's), let V be the $T^* \times n$ matrix of errors, and let Π be a $m \times n$ matrix of regression coefficients. The regression model (1) can be written compactly as

$$Y = X\Pi + V.$$

Under normality, twice the conditional log likelihood function is (except for an additive constant)

$$L = T^* \log |\Omega^{-1}| - tr\Omega^{-1}V'V.$$

Differentiating with respect to Ω^{-1} and using the fact that $\partial \log |S|/\partial S = S^{-1}$, we find $\widehat{\Omega} = V'V/T^*$. Thus the MLE for Π is found by maximizing the concentrated likelihood function

$$L^* = -T^* \log |\frac{V'V}{T^*}| - T^*n$$

or, equivalently, minimizing $|V'V| = |(Y - X\Pi)'(Y - X\Pi)|$. But defining the OLS estimator $P = (X'X)^{-1}X'Y$ and noting that $I - X(X'X)^{-1}X'$ spans the space orthogonal to X, we see that

$$|(Y - X\Pi)'(Y - X\Pi)| = |(Y - XP)'(Y - XP) + (P - \Pi)'X'X(P - \Pi)|$$

is minimized at $\Pi = P$.

The same argument applies to equation (2). Gaussian maximum likelihood estimates of the parameters of the error-correction form of the model are obtained by equation-by-equation OLS.

5 Granger Causality

Following the terminology of earlier time-series scholars, Granger suggested that, for a bivariate stationary process $\{x_t, y_t\}$, a causal relationship between the two series could be defined in terms of best linear predictors. In particular, if $\mathbf{P}(a|b)$ represents the best linear predictor of a given b, $X_t = (x_t, x_{t-1}, x_{t-2}, ...)$, and $Y_t = (y_t, y_{t-1}, y_{t-2}, ...)$, then y is said to cause x if and only if $\mathbf{P}(x_{t+1}|X_t, Y_t)$ actually depends on Y_t. If, in addition, $\mathbf{P}(y_{t+1}|Y_t, X_t)$ does not depend on X_t, a one-way causal relationship from y to x is said to exist. (The restriction to *linear* predictors is made for convenience and could be relaxed.) Although one can easily give examples where this definition does not seem appropriate, these examples typically involve the introduction of a third variable that causes both x and y. If applied in the appropriate conditional setting, Granger's definition seems to be consistent with the way the word "causality" is used in everyday speech. Of course, in practice, it may not be possible to condition on the appropriate third variables thus making the Granger approach infeasible.

Suppose the stationary process $\{x_t, y_t\}$ has the moving average representation

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} P(L) & Q(L) \\ R(L) & S(L) \end{bmatrix} \begin{bmatrix} u_t \\ v_t \end{bmatrix}$$

where $\{u_t, v_t\}$ is a bivariate white-noise process with covariance matrix Σ and $Q(0) = R(0) = 0$. In scalar notation, we have:

$$\begin{aligned} x_t &= P(L)u_t + Q(L)v_t \\ y_t &= R(L)u_t + S(L)v_t \quad . \end{aligned}$$

If the 2×2 lag polynomial matrix is invertible, we find the autoregressive representation

$$\begin{aligned} x_t &= A(L)x_{t-1} + B(L)y_{t-1} + u_t \\ y_t &= C(L)x_{t-1} + D(L)y_{t-1} + v_t \end{aligned}$$

where u_t and v_t are uncorrelated with lagged x's and y's. Since $\mathbf{P}(x_{t+1}|X_t, Y_t) = A(L)x_t + B(L)y_t$, by Granger's definition y causes x if and only if $B(L)$ is nonzero. Note that $B(L) = 0$ if and only if $Q(L) = 0$, so an alternative characterization in this invertible case is that y causes x if and only if $Q(L)$ is nonzero.

Sims pointed out a third way to characterize causation. He proved that, in the invertible case, y causes x (in the sense of Granger) if and only if the best linear predictor of y_t given all current, past and future values of x actually depends on the future values. That is, defining $X_{t+} = (x_{t+1}, x_{t+2}, ...)$, then y causes x if and only if $\mathbf{P}(y_t|X_t, X_{t+})$ actually depends on X_{t+}. The following proof of Sims' characterization is based on the one given in the textbook.

We first show that $Q(L) = 0$ implies that $\mathbf{P}(y_t|X_t, X_{t+})$ does not depend on X_{t+}. Define $w_t = v_t - u_t\sigma_{uv}/\sigma_{uu}$, so w_t is uncorrelated with u_s for all s and t. Then, $Q(L) = 0$ implies $u_t = P^{-1}(L)x_t$ and hence

$$y_t = R(L)u_t + S(L)[w_t + \frac{\sigma_{uv}}{\sigma_{uu}}u_t] = T(L)x_t + S(L)w_t$$

where $T(L) = [R(L) + S(L)\sigma_{uv}/\sigma_{uu}]P^{-1}(L)$ is a one-sided backward filter. But w_t is uncorrelated with x_s for all s and t. Thus, $Q(L) = 0$ implies

$$\mathbf{P}(y_t|\ X_t, X_{t+}) = \mathbf{P}[T(L)x_t|\ X_t, X_{t+}] + \mathbf{P}[S(L)w_t|\ X_t, X_{t+}] = T(L)x_t$$

which does not depend on future values of x_t.

Conversely, suppose $\mathbf{P}(y_t|X_t, X_{t+}) = \Psi(L)x_t$ where $\Psi(L)$ is a one-sided backwards filter. Then $\eta_t \equiv y_t - \Psi(L)x_t$ is uncorrelated with x_s for all s and t. Both η_t and x_t have univariate moving average representations, say

$$\begin{aligned} \eta_t &= F(L)\varepsilon_t \\ x_t &= G(L)\nu_t \end{aligned}$$

where $F(L)$ and $G(L)$ are one-sided backwards filters with $F(0) = G(0) = 1$; ε_t and ν_t are each white-noise series. Now define $e_t = \varepsilon_t + \psi\nu_t$, where $\psi = \Psi(0)$. Since η_t and x_s are uncorrelated for all s and t, it follows that ε_t and ν_s are uncorrelated for all s and t. Hence $\{e_t, \nu_t\}$ is a bivariate white-noise series. Substituting, we find that

$$\begin{aligned} y_t &= \Psi(L)x_t + \eta_t = \Psi(L)G(L)\nu_t + F(L)[e_t - \psi\nu_t] \\ &= [\Psi(L)G(L) - \psi F(L)]\nu_t + F(L)e_t \\ &= H(L)\nu_t + F(L)e_t \end{aligned}$$

where $H(L) = \Psi(L)G(L) - \psi F(L)$ is a one sided backwards filter and $H(0) = 0$. Thus the assumption that the best linear predictor of y_t given all values of x does not depend on future x implies that $\{x_t, y_t\}$ has a triangular moving average representation of the form

$$\begin{aligned} x_t &= G(L)\nu_t \\ y_t &= H(L)\nu_t + F(L)e_t \end{aligned}$$

and hence y does not cause x.

6 Dynamic Structural Models

6.1 The simultaneous equations model

The VAR model (1) expresses the conditional mean of z_t (given its past history) as a linear function of its most recent past history. The parameters have no particular meaning in terms of economic behavior or technology. The equations simple represent a convenient way to capture the autocorrelations in the data. They are useful for forecasting the future from past realizations under the assumption that the process that generates the data has not changed. Sometimes, however, economists wish to specify relationships among economic variables that *do* have economic meaning. The Marshallian supply and demand model for a nonstorable good under perfect competition is a standard example. The system consists of three equations: one relates quantity demanded to price, a second relates quantity supplied to price, and a third says that over the period in question price has adjusted to equate quantity supplied and quantity demanded. Let z_t be a two-dimensional vector consisting of quantity traded in period t and the equilibrium price during that period. Then, if x_t is a vector of exogenous variables (like weather condtions) that describe the conditions of the market during period t and if we permit supply and demand to depend on past history, we might write this model in linear form as

$$A(L)z_t + B(L)x_t = \varepsilon_t. \tag{3}$$

The vector ε_t is usually interpreted as representing omitted variables that affect supply and demand. If we assume that, conditional on all $x's$ and on *past* values of z_t, the ε_t are white noise, then we have a model that looks like the VAR model (1). But there are three important differences.

1. In the supply-demand model, one equation is interpreted as describing the behavior of consumers and the other equation is interpreted as describing the behavior of producers; in a standard VAR model the equations have no necessary economic interpretation.

2. There is no reason to believe every variable that appears in the demand equation necessarily appears in the supply equation. For example, we might postulate that weather conditions affect supply but not demand.

3. In the supply-demand model, there is no reason to impose the restriction $A_0 = I$. Indeed, the assumption that price is determined by the interaction of supply and demand precludes that assumption.

It is true that, given a supply equation and a demand equation, one can always solve for a "reduced form" that looks like a VAR model with $A_0 = I$. Indeed, if the goal is simply to forecast price and quantity from past history, this form is all that is necessary. Sometimes, however, our goal is not simply forecasting. We may want to estimate and test hypothesis about the parameters of the individual supply or demand equations.

More generally, economists often propose models that describe the behavior of different groups of people. The coefficients in a given structural equation are defined in terms of some economic theory. Although one will typically set units of measurement by assuming the diagonal elements of A_0 are equal to one, there may be compelling reasons to allow a given structural equation to contain more than one current endogenous variable. The resulting equilibrium model is often called the *dynamic simultaneous equations model.* Setting $A_0 = I$ is no longer just a convenient normalization but changes the meaning of the remaining parameters.

In such a structural interpretation, the individual equations are taken to be autonomous, each describing the behavior of a specific group of agents. It may be meaningful to change the coefficients of one equation while keeping the coefficients of the other equations constant. Thus the parameters in the matrices $A_0, ..., A_p$ may be of interest outside of their role in generating forecasts under unchanging structure. For example, they may be of use in predicting the effects of policy intervention. Of course, estimating the matrices $A_0, ..., A_p$ will generally not be possible unless some further restrictions are imposed. Although the parameters of a VAR model can be consistently estimated by least squares regressions, estimation of the structural parameters in a simultaneous equations model will need some alternative procedure such as the method of instrumental variables.

The dynamic simultaneous equations model was first proposed by researchers at the Cowles Foundation in the 1940's. Early applications were times-series studies of markets for individual agricultural goods and of the aggregate macro economy. In recent years, the model is often used with cross-section data to study the behavior of individual decision-making units. In both cases, the estimation methods used are the ones initially proposed by the Cowles researchers fifty years ago.

6.2 LIML and 2SLS

Suppose we are interested in one equation of a simultaneous system, say

$$y = Z\delta + u$$

where y is a $T \times 1$ vector of observations on an endogenous variable, Z is a $T \times q$ matrix of "explanatory" variables, and u is a $T \times 1$ vector of unobserved errors. The columns of Z represent (1) other endogenous variables, (2) lagged endogenous variables, and (3) exogenous variables. If enough lagged endogenous variables have been included, u_t will be a white noise series and uncorrelated with all endogenous variables dated earlier than t. However, u_t will be correlated with any contemporaneous endogenous variable. If we have not included enough lags, even lagged endogenous variables may be correlated with u_t.

Let X be the matrix of observations on m "predetermined" variables; that is, variables that at time t are uncorrelated with u_t. We assume that $m \geq q$; that is, the number of predetermined variables is at least as great as the number of parameters to be estimated. Using these as instruments, the two-stage least-squares (2SLS) estimator of δ is that vector d that minimizes

$$(y - Zd)'X(X'X)^{-1}X'(y - Zd).$$

The limited information maximum likelihood (LIML) estimator maximizes the conditional Gaussian likelihood function subject only to the restriction that some of the possible predetermined variables do not appear in the structural equation of interest. It can be shown that this estimator is obtained by minimizing

$$\frac{(y - Zd)'X(X'X)^{-1}X'(y - Zd)}{(y - Zd)'[I - X(X'X)^{-1}X'](y - Zd)}.$$

If $T^{-1/2}X'u$ has a limiting normal distribution and $T^{-1}X'Z$ converges in probability to a matrix having rank q, both estimators are consistent and have the same large-sample distribution. The LIML estimator seems to have better small-sample properties, particularly when $m - q$ is large.

6.3 Impulse response functions and structural VARs

See Hamilton, sections 11.4-11.6.