Introduction

This course covers topics in time series/macroeconometrics. The topics are: Stochastic processes, with a focus on autoregressive (AR) and moving average models. We will discuss how to work with these processes, how to estimate the parameters using MLE, and how to forecast with these processes. We then discuss multivariate autoregressions (VARs). We discuss stationarity, Granger-causality, impulse response function analysis, "structural" VARs, and pseudo-Bayesian VARs for forecasting. Following VARs, we discuss GMM estimation of Euler equations that arise in dynamic economic models. We then present state-space modeling and the Kalman filter. These are very useful tools for dynamic economic models with unobserved shocks. We then discuss hidden Markov (Markov switching models). Following the Markov-switching model, we discuss issues associate with estimating complete equilibria of dynamic stochastic models. This includes computing linearized equilibria to nonlinear, dynamic, stochastic, rational expections models, and FIML estimation of linearized models using the Kalman filter. We then discuss estimation by simulation, including simulated method of moments and simulated MLE. We close the course with analysis of nonstationary time series, including cointegration.

Stochastic Processes

Preliminaries

A stochastic process generates a sequence of random variables, indexed by time. If $\{y_i\}$ is a stochastic process, its sample path, or realization, is an assignment to each i of a possible value for y_i. Thus, a realization of $\{y_i\}$ is a sequence of real numbers, indexed by time.

Note that we only observe one particular time series realization of the stochastic process $\{y_i\}$. Ideally, we would like to observe many different realizations of the stochastic process. If we did get a chance to see many realizations (in particular, suppose that the number of realizations went to ∞), then we would form the expected value of the random variable y at date t as:

$$E(y_t) = \lim_{N\to\infty} 1/N \sum_{i=1}^{N} y_{it}$$

This is called the *ensemble mean* for the stochastic process at date t. But of course, we only see a single realization of U.S. GNP - we don't get a chance to see other realizations. In some cases, however, the time series average of a single realization is a consistent estimate of the ensemble mean. We will see this in the following section when we discuss ergodicity.

Autocovariance

Consider a mean zero sequence of random variables: $\{x_t\}$. The jth autocovariance is given by:

$$\gamma_{jt} = E(x_t x_{t-j})$$

Thus, autocovariance is just the covariance between a random variable at different points of time. Notice that if j = 0, then the "0th" autocovariance is just the variance: $E(x_t)^2$. The ensemble mean of this autocovariance is given by:

$$\gamma_{jt} = \lim_{N\to\infty} 1/N \sum x_{it} x_{it-j}$$

Why Do We Care About Autocovariances?

Autocovariance measures the covariance between a variable at two different points in time. This has implications for a number of issues, including economic forecasting and understanding behavioral economic mechanisms.

Forecasting: If we know that the statistical relationship between a variable at different points in time, this can help us forecast that variable in the future. For example, suppose output is higher than normal today, and that when output is higher than average today, it also tends to higher than average tomorrow. This will lead us to forecast output tomorrow to higher than average. The details of how we make forecasts will be considered later.

Economic Modeling: Our economic models summarize the behavior of economic variables. If variables have large autocovariances, then some mechanism is causing persistence in these variables, and our models should explain that persistence through preferences, technologies, policies, or shocks. On the other hand, if variables have zero autocovariances, then the variables have no persistence, and out models should explain the mechanisms behind this.

The autocovariance matrix is:

Stationarity

Covariance Stationarity (Weak Stationarity)

If the mean of a random variable nor the autocovariances of the random variable depend on the date t, then the stochastic process is covariance stationary (weakly stationary):

$$E(y_t) = \mu$$

$$E(y_t - \mu)(y_{t-j} - \mu) = \gamma_j$$

Example 1: Suppose $\{y_t\}$ is a mean zero process with variance σ^2. Verify that this process is covariance stationary.

Example 2: Suppose $y_t = \alpha t + \varepsilon_t$, where t = {1,2,3,...} and ε is a normal random variable with mean 0 and variance σ^2. Show that this process is not covariance stationary.

Note that for autocovariances, a weakly stationary process implies that:

$$\gamma_j = \gamma_{-j}$$

Strict Stationarity

A process is strictly stationary if the joint distribution for the stochastic process does not depend on time. Note that a process that is strictly stationary with finite second moments must be covariance stationary. Since many issues we are interested in don't requie strict stationary, we will hereafter refer to a stationary time series as one that is covariance stationary.

Ergodicity

A process is ergodic for its mean if a sample average calculated from a single long realization converges in probability to its expected value. In this case, the ensemble average one would calculate from many realizations can be replaced by an average from a single (long) realization:

$$(1/T)\sum y_t \rightarrow E(y_t)$$

Ergodicity requires that the autocovariances tend to zero as j becomes large. In particular,it can be shown that a process is ergodic for its mean if it is "square summable":

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty$$

A process is ergodic for second moments if the sample autocovariances from a single realization converge in probability to the expected value:

$$1/(T-j)\sum_{t=j+1}^{T}(y_t - \mu)(y_{t-j} - \mu) \rightarrow \gamma_j$$

For Gaussian process, square summability is sufficient for ergodicity for the second moments.

Autocorrelation

Just as it is useful to normalize covariances by dividing by the respective variables' standard deviations, it is also useful to normalize autocovariances. The jth autocorrelation is denoted as $\rho_j = \frac{\gamma_j}{\gamma_0}$, which and is given by:

$$\frac{E(y_t y_{t-j})}{\sqrt{E(y_t)^2}\sqrt{E(y_{t-j})^2}}$$

Note that for ρ_0, it is equal to 1. Thus, autocorrelation tells us the correlation between a variable at two different points in time.

The White Noise Process

White noise is a serially uncorrelated process. Consider the following realization, $\{\varepsilon_t\}$ with zero-mean. It's first and second moments are given by:

$$E(\varepsilon_t) = 0$$

$$E(\varepsilon_t^2) = \sigma^2$$

$$E(\varepsilon_t \varepsilon_\tau) = 0, \tau \neq t$$

Note that this latter feature implies that the autocovariances are zero.

The white noise process is key because it is the foundation for most of the other stochastic processes we are interested in.

Why do we care about white noise?

In studying rational expectations models, we will be interested in understanding how the variables of interest respond to an unanticipated change in a random variable today, and in the future. This is called *impulse response function analysis.* We will discuss this later in the course.

Moving Average Processes

Recall the white noise process, $\{\varepsilon_t\}$. We now use this process to construct the moving average (MA) process. We first construct the MA(1) process:

$$y_t = \mu + \varepsilon_t + \theta \varepsilon_{t-1}$$

This is called a moving average process because the process is a weighted average of a white noise process. The term ε_t is often called an **innovation**.

The unconditional expectation of this process is:

$$E(y_t) = \mu + E(\varepsilon_t) + \theta E(\varepsilon_{t-1}) = \mu$$

The variance is:

$$E(y_t - \mu)^2 = E(\varepsilon_t)^2 + \theta^2 E(\varepsilon_{t-1})^2 = (1 + \theta^2)\sigma^2$$

The first autocovariance is:

$$E((y_t - \mu)(y_{t-1} - \mu)) = \theta\sigma^2$$

To see this, note:

$$E((y_t - \mu)(y_{t-1} - \mu)) = E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-1} + \theta\varepsilon_{t-2}))$$

Combining terms, we see that the only non-zero term is $E((\varepsilon_{t-1})(\theta\varepsilon_{t-1}))$.

Verify that all other autocovariances are 0, and verify that this is a covariance stationary process. .

The MA (1) process has non-zero autocorrelation at lag 1. It is given by:

$$\rho_1 = \frac{\theta\sigma^2}{(1 + \theta^2)\sigma^2} = \frac{\theta}{1 + \theta^2}$$

The magnitude of this coefficient depends on the value of the parameter θ. But note that it's maximum value is 0.5.

How do shocks today affect this random variable today and into the future? By construction, a one-unit shock to ε_t today changes the random variable y_t by the same amount. Tomorrow, this shock affects y_t by factor θ. But after that, the shock today has no affect on the random variable of interest.

The qth order MA process is denoted by MA(q), and is given by:

$$y_t = \mu + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Note that its variance is given by:

$$\gamma_0 = (1 + \sum \theta_i^2)\sigma^2$$

The autocorrelations are given by:

$$\rho_1 = \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}$$

Solve for $\rho_2, \rho_3,$ and ρ_4 for the MA (4) process.

Note that in these higher order MA processes, a shock today affects *y* for more periods. In particular, the number of periods into the future that a shock today has an affect is given by the order of the process.

We will need one more assumption to talk about well-defined MA processes when q is ∞. In this case, we will assume square-summability:

$$\sum_{j=0}^{\infty} \theta_j^2 < \infty$$

If the process is square summable, then it is covariance stationary.

Why do we care about MA processes?

The MA process is fundamental in analysing dynamic economic models, and is used to construct impulse response functions (IRF). The IRF measures the dynaic responses of variables we care about to shocks.

The MA process also plays a role in the *Wold Decomposition Theorem:*

The Wold Theorem states that any mean-zero covariance stationary process can be written as an infinite order moving average process plus a deterministic term:

$$y_t = \sum_{i=0}^{\infty} \varepsilon_{t-i} + \kappa_t$$

Autoregressive (Markov) Processes

Autoregressive, or Markov processes, are stochastic processes in which the random variable is related to lagged values of itself. The first-order process, or AR (1) process is given by:

$$y_t = \mu + \phi y_{t-1} + \varepsilon_t$$

Assume that ε is a white noise process.

Recall that the finite process MA was stationary. To gaurantee stationarity for the AR process, we need some more restrictions. If $|\phi| < 1,$ then a covariance stationary process exists. To see this, solve the difference equation backwards, which yields

$$y_t = \frac{\mu}{1-\phi} + \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}$$

Note that by solving this difference equation backwards, we have re-written it as a MA(∞).

This is called the moving average representation of the AR(1). This process is covariance stationary provided it is square summable. Note that it is square summable:

$$\sum_{0}^{\infty} \phi^i = \frac{1}{1-\phi} < \infty$$

The mean of this process is:

$$E(y_t) = \frac{\mu}{1-\phi}$$

The variance is:

$$\gamma_0 = \sigma^2 \frac{1}{1-\phi^2}$$

the jth autocovariance is:

$$\gamma_j = \sigma^2 \frac{\phi^j}{1-\phi^2}$$

The jth autocorrelation is thus:

$$\rho_j = \frac{\gamma_j}{\gamma_0} = \phi^j$$

The second order autoregressive process is:

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

Recall from the study of basic difference equations that this equation is stable provided that the roots of:

$$(1 - \phi_1 z - \phi_2 z^2) = 0$$

lie outside the unit circle. If this is satisfied, then the process is covariance stationary.

Why do we care about AR processes?

Statistically, almost ALL economic time series are well approximated by low order AR processes .

Behaviorally, almost ALL the dynamic economic models we write down can be represented (after linearization) as AR processes.

Lag Operators

It is sometimes convenient to write stochastic processes using lag operators, L. The lag operator shifts variables across their time indexes. For example,

$$Lx_t = x_{t-1}$$

$$L^2 x_t = x_{t-2}$$

Note that the lag operator, when applied to a constant, just returns that constant:

$$L^i \alpha = \alpha,$$

We can write the AR(2) process using the lag operator as follows:

$$y_t(1 - \phi_1 L - \phi_2 L^2) = \mu + \varepsilon_t$$

Note that the moving average representation is given by:

$$y_t(1 - \phi_1 L - \phi_2 L^2) = \frac{\mu}{(1 - \phi_1 - \phi_2)} + \frac{\varepsilon_t}{(1 - \phi_1 L - \phi_2 L^2)}$$

Thus, the infinite order moving average coefficients are given by:

$$\theta_1 L + \theta_2 L^2 + \ldots = (1 - \phi_1 L - \phi_2 L^2)^{-1}$$

We can also write the MA representation using backwards substitution, as in the first order case.

The autocovariances become a bit more complicated now. To calculate these, subtract off the mean for the process:

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \varepsilon_t$$

Multiply both sides by $(y_{t-j} - \mu)$, and then take expectations:

$$\gamma_j = \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2}$$

Note that for the AR(2) process, the autocovariances also follow a second order difference equation.

Similarly, the autocorrelations are given by:

$$\rho_j = \phi_1\rho_{j-1} + \phi_2\rho_{j-2}$$

Note that for j = 1, we have:

$$\rho_1 = \frac{\phi_1}{1 - \phi_2}$$

For j = 2, we have:

$$\rho_2 = \frac{\phi_1^2}{1 - \phi_2} + \phi_2$$

Given these expressions, the remaining autocorrelations can be solved recursively.

High Order AR Processes

The pth order AR process is given by:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-1} + \varepsilon_t$$

It is stationary provided that the roots of

$$1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = 0$$

all lie outside the unit circle.

The autocovariances and autocorrelations are solved for analagously to the second order case.

ARMA Processes

ARMA processes contain both autoregressive and moving average components. The ARMA (p,q) process is:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

Using lag operators, we can write this process as:

$$(1 - \phi_1 L - \phi_2 L^2 - \ldots \phi_{p^p}) y_t = \mu + (1 + \theta_1 L + \theta_2 L^2 \ldots \theta_q L^q)\varepsilon_t$$

Stationarity requires the usual assumption on the roots of the pth order AR polynomial (that is, they all lie outside the unit circle):

$$1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = 0$$

If the process is stationary, then it posesses an MA representation:

$$y_t = \mu^* + \psi(L)\varepsilon_t$$

where the MA parameters are given as:

$$\psi(L) = \frac{1 + \theta_1 L + \theta_2 L^2 + \ldots}{1 - \phi_1 L - \phi_2 L^2 - \ldots}$$

Be Careful not to Overparameterize!

With ARMA processes, there is a possibility for overparameterizing. To see this, consider the ARMA (1,1) process:

$$y_t = \phi y_{t-1} + \varepsilon_t + \theta_{\varepsilon t-1}$$

Now, suppose that $\phi = -\theta$. In this case, the ARMA (1,1) process is just white noise! This can be easily deduced from writing the process using lag operators,

$$y_t(1 - \phi L) = \varepsilon_t(1 - \phi L)$$

and multiply both sides of the equation by $(1 - \phi L)$. Thus, with ARMA processes, it is important to look out for parameter redundancies.

Invertibility

An MA process of the form:

$$y_t = (1 + \theta_1 L + \theta_2 L^2 + \ldots \theta_q L^q)\varepsilon_t$$

is invertible if the roots of the following polynomial lie outside the unit circle:

$$(1 + \theta_1 z + \theta_2 z^2 + \ldots \theta_q z^q) = 0$$

Invertibility means that the MA process can be written as an AR process. Thus, invertibility for MA processes basically works like stationarity works for AR processes.

Note for the MA (1) process:

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1}$$

, this means that the MA coefficient $\theta < |1|$. Note that this requirement is not really an issue, because for every non-invertible MA process, we can always find an invertible process that is observationlly equivalent. To see this, choose any value for θ greater than 1, and call that θ^*. Now, solve for the first order autocorrelation coefficient, which is given by

$$\frac{\theta^*}{1 + \theta^{2*}} = \rho^*$$

Note that this value lies between -0.5 and 0.5. Next, choose a value for $\theta < 1$ such that:

$$\frac{\theta}{1 + \theta^2} = \rho^*$$

This follows directly, given that the mapping between θ and ρ^* is continuous. Note that the invertible values that is implied by ρ^* is just the inverse:

$$\theta = \frac{1}{\theta^*}$$

Principles of Forecasting

We now discuss using current and past values of variables or their innovations. Define the collection of this information to be X.

First we define the forecast error:

$$\eta_{t+1} = y_{t+1} - y^*_{t+1} \mid_t$$

Mean square forecast error is:

$$E(y_{t+1} - y^*_{t+1} \mid_t)^2$$

One possibility in forecasting is to minimize this loss function. If we restrict ourselves to linear forecasts of y based on X, that is:

$$y^*_{t+1} \mid_t = \alpha' X_t$$

then the forecast that minimizes mean square error is the linear projection of y on X, which satisfies the following

$$E(y_{t+1} - \alpha' X_t) X_t' = 0$$

Note that is similar to least squares - the difference is that the linear projection involves the population moments, while least squares involves the sample moments.

Forecasting an AR Process

We now consider forecasting a stationary AR(1) process. Before we begin, we know the following. Since in a stationary process the effects of shocks die out, then long run forecasts will converge to the unconditional mean of the process, while short run forecasts will be related to the processe's most recent realization.

First, let's consider forecasting the process using lagged values of the innovations:

We know that:

$$y_{t+s} = \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \ldots + \psi_s \varepsilon_t + \psi_{s+1} \varepsilon_{t-1} + \ldots$$

Thus, the optimal linear forecast is:

$$E_t y_{t+s} = \psi_s \varepsilon_t + \psi_{s+1} \varepsilon_{t-1} + \ldots$$

and the forecast error is:

$$\varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \ldots + \psi_{s-1} \varepsilon_{t+1}$$

It is useful to use lag operators to construct our forecasts:

Note that

$$\frac{\psi(L)}{L^s} = L^{-s} + \psi_1 L^{1-s} + \ldots + \psi_s L^0 + \ldots + \psi_{s+n} L^n$$

where

$$L^{-s} x_t = x_{t+s}$$

Now define the annihilation operator - this replaces negative powers of L by 0, which is useful for forming forecasts:

$$\left[\frac{\psi(L)}{L^s} \right]_+ = \psi_s + \psi_{s+1} L + \psi_{s+2} L^2 + \ldots$$

Note that we can now write our forecasting equation as:

$$E_t y_{t+s} = \left[\frac{\psi(L)}{L^s} \right]_+ \varepsilon_t$$

We will return to this shortly.

Now consider forecasts of y based on lagged y's. Consider the process:

$$\eta(L) y_t = \varepsilon_t$$

Suppose that

$$\eta(L) = \psi(L)^{-1}$$

Then we have:

$$E_t y_{t+s} = \left[\frac{\psi(L)}{L^s} \right]_+ \eta(L) y_t$$

or susbtituting out the AR coefficients, we have:

$$E_t y_{t+s} = \left[\frac{\psi(L)}{L^s} \right]_+ \psi(L)^{-1} y_t$$

This is known as the *Wiener-Kolmogorov* prediction formula.

Example 1: Forecasting an AR(1): $y_t = \phi y_{t-1} + \varepsilon_t$

Note that

$$\psi(L) = 1 + \phi L + \phi^2 L^2 + \ldots$$

Also, note that:

$$\left[\frac{\psi(L)}{L^s} \right]_+ = \phi^s + \phi^{s+1} L^1 + \ldots = \frac{\phi^s}{1 - \phi L}$$

This implies:

$$E_t y_{t+s} = \left[\frac{\psi(L)}{L^s} \right]_+ \psi(L)^{-1} y_t = \frac{\phi^s}{1 - \phi L} (1 - \phi L) y_t = \phi^s y_t$$

Example 2: Forecasting an MA (1)

$$E_t y_{t+s} = \left[\frac{\psi(L)}{L^s} \right]_+ \psi(L)^{-1} y_t = E\left[\frac{1 + \theta L}{L^s} \right]_+ \frac{1}{1 + \theta L} y_t$$

Suppose that s = 1:

$$E_t y_{t+s} = E\left[\frac{1 + \theta L}{L} \right]_+ \frac{1}{1 + \theta L} y_t$$

or

$$E_t y_{t+s} = \frac{\theta}{1+\theta L} y_t$$

Maximum Likelihood Estimation

We will now discuss estimating the parameters of ARMA models.

MLE for AR Models

We begin with the simplest case, which is the AR(1) process.

Consider the first observation:

$$y_t = \mu + \phi y_{t-1} + \varepsilon_t$$

The first two moments of this first observation are:

$$E(y_0) = \mu/(1-\phi)$$

$$E(y_0 - \mu)^2 = \sigma^2/(1-\phi^2)$$

Now, let's define the pdf's for this first observation:

$$f_y(y_{1^*};\theta) = f_{y_1}(y_1;\mu,\phi,\sigma^2)$$

With Gaussian innovations, this becomes:

$$\frac{1}{\sqrt{2\pi}\sqrt{\frac{\sigma^2}{1-\phi^2}}} \exp \frac{(-\{y_1 - [\mu/(1-\phi)]\})^2}{2\sigma^2/(1-\phi^2)}$$

Next, consider the distribution of the second observation, condition on observing the first observation:

$$f_{y_{2^*}|y_{1^*}}(y_2 \mid y_1;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(-(y_2 - \mu - \phi y_1)^2)}{2\sigma^2}$$

Note that the joint density of the first 2 observations is:

$$f_{y_{2^*}} \mid y_{1^*}(y_2 \mid y_1;\theta) \cdot f_{y_1}(y_1;\theta)$$

We now are in a position to form the likelihood:

Note that for any observation, the density is given by:

$$f_{y_{t^*}|y_{t-1^*}}(y_t \mid y_{t-1};\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(-(y_t - \mu - \phi y_{t-1})^2)}{2\sigma^2}$$

Since the density of any observation at date t only depends on variables at date t-1, the joint density is given by the product of the individual densities:

$$f_{y_{1^*}}(y_1;\theta) \cdot \Pi_{t=2}^{T} f_{y_t|y_{t-1}}(y_t \mid y_{t-1};\theta)$$

Taking logs, we get:

$$L = \log f_{y_1}(y_1;\theta) + \sum_{t=2}^{\infty} \log f_{y_t|y_{t-1}}(y_t \mid y_{t-1};\theta)$$

Substituting, (and omitting constants), we obtain the log-likelohood as:

$$l = -\log(\sigma^2/(1-\phi^2) - \frac{(y_1 - (\frac{\mu}{1-\phi}))^2}{2\sigma^2/(1-\phi^2)} - \frac{T-1}{2}\log(\sigma^2) - \sum_{t=2}^{T}\left(\frac{(y_t - \mu - \phi y_{t-1})^2}{2\sigma^2}\right)$$

Exact MLE is tricky for this model, as it involves a nonlinear system of equations. Conditional MLE is easy to compute, and is asymptotically equivalent to exact MLE.

The idea behind conditional MLE is to treat the first observation deterministic and maximize the likelihood conditional on the first observation.

Note that MLE for the parameters μ and ϕ is equivalent to OLS:

$$\min_{\mu,\phi} \sum_{t=2}^{T} (y_t - \mu - \phi y_{t-1})^2$$

This implies:

$$\begin{bmatrix} \mu_{ols} \\ \phi_{ols} \end{bmatrix} = \begin{bmatrix} T-1 & \sum y_{t-1} \\ \sum y_{t-1} & \sum y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_t \\ \sum y_{t-1} y_t \end{bmatrix}$$

The estimate for the innovation variance, σ^2_{ols} is given by:

$$\sigma^2_{ols} = \sum \frac{(y_t - \mu_{ols} - \phi_{ols} y_{t-1})^2}{T-1}$$

We have just covered the simplest case, which is the AR(1). Using conditional MLE the AR(p) process is done analagously, and also have the same asymptotic distribution as exact MLE.

MLE for MA Models

We now consider the MLE for the MA(1) model.

$$y_t = \mu + \varepsilon_t + \theta \varepsilon_{t-1}, \varepsilon_t \sim N(0, \sigma^2)$$

The tricky part is that we need to estimate a parameter using the past sequence of a latent variable: $\{\varepsilon_t\}$. If the value of ε_{t-1} was known with certainty, then we have a well-defined density:

$$f_{y_t^* \mid \varepsilon_{t-1}^*}(y_t \mid \varepsilon_t; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(-(y_t - \mu - \theta \varepsilon_{t-1})^2)}{2\sigma^2}$$

More specifically, suppose we knew $\varepsilon_0 = 0$. Then we can recursively solve for the entire sequence of $\{\varepsilon_t\}$ as:

$$\varepsilon_t = y_t - \mu - \theta \varepsilon_{t-1}$$

The conditional density of any observation is then given by:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(-\varepsilon_t)^2)}{2\sigma^2}$$

The log-likelihood can be formed using methods described earlier, and we get:

$$l = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum \frac{\varepsilon_t}{2\sigma^2}$$

However, because the $\varepsilon's$ depend on the parameters, we need to evaluate the likelihood numerically. We now describe some numerical techniques for doing this.

Grid Search Methods

The grid search method is a very simple and robust approach to evaluating this likelihood. The only drawback is that it becomes impractical for models with many parameters (that is, the "curse of dimensionality" applies).

But for a one-dimensional problem, it is easy and feasible.

Step 1: Define the feasible parameter space. For the MA (1) model, this means:

$$-1 < \theta < 1$$

(2) Define a finite grid over the parameter space. For example:

$$\{-0.995, -.990, \ldots, 0.995\}$$

Note that sometimes particular aspects of the problem suggest that nature and coarseness of the grid. For example, if you have some idea of where the optimum is, you can put many more points in that neighborhood than in other neighborhoods.

(3) Evaluate the likelihood at each point on the grid. Choose the grid point with the highest likelihood

(4) Optional - if the grid was coarse, then one can construct a new grid centered around the optimal point that was previously identified.

Newton Methods for Optimization

Newton methods use iterative, linearization techniques to solve for optima. This is useful in the MA model case, since the likelihood is characterized by a fairly complicated set of nonlinear equations.

There are a number of different Newton-type methods. We will focus on the Newton-Raphson method. This requires that the log-likelihood is concave and that the matrix of second derivatives exist.

Let θ be an (a x 1) vector. Let $g(\theta^0)$ be the gradient vector of the log-likelihood at the parameter vector θ^0 :

$$g(\theta^0) = \frac{\partial l(\theta)}{\partial \theta} \mid_{\theta=\theta^0}$$

Next, define H to be the matrix of second derivatives (multiplied by -1):

$$H(\theta^0) = -\frac{\partial l(\theta)}{\partial \theta \partial \theta'} \mid_{\theta=\theta^0}$$

Now, take a Taylor series expansion of the log-likelihood around the θ^0:

$$l(\theta) \cong l(\theta^0) + g(\theta^0)'[\theta - \theta^0] - \frac{1}{2}[\theta - \theta^0]'H(\theta^0)[\theta - \theta^0]$$

The Key idea: maximize the approximate likelihood, which requires differentiating with respect to θ and setting the derivative to 0. This yields:

$$g(\theta^0) = H(\theta^0)[\theta - \theta^0]$$

Now, suppose that θ^0 is an initial guess. The result above shows that an improved guess can be obtained by inverting H to get:

$$\theta - \theta^0 = H(\theta^0)^{-1} g(\theta^0)$$

The Newton-Raphson iterative algorithm therefore becomes:

$$\theta^{m+1} - \theta^m = H(\theta^m)^{-1} g(\theta^m)$$

Note that if the likelihood is quadratic, then this approach will yield the MLE in one step.

In the case where the second order expansion is not the exact likelihood, a modification using “step-size” is sometimes adopted. The idea is to form:

$$\theta^{m+1} - \theta^m = sH(\theta^m)^{-1} g(\theta^m)$$

where s is the step size, and is a scalar. The step size takes in a particular direction. What we do is to calculate θ^{m+1} for different values of s and then choose θ^{m+1} that yields the highest likelihood. Standard choices are values between -0.8 and 0.8.

We continue this iterative procedure until:

$$| \ \theta^{m+1} - \theta^{m} \ | < c$$

where c is some small number and is known as the convergence criterion.

Statistical Inference

We may be interested in testing hypotheses about the parameter vector θ. One way is to calculate standard errors for the elements of the vector. This can be done by calculating the asymptotic covariance matrix of the parameter vector.

The Asymptotic distribution of MLE is given by:

$$\hat{\theta} \sim N(\theta_0, T^{-1} I^{-1})$$

where I is the information matrix - the second derivative of the log-likelihood with respect to the paramter vector:

$$I = -T^{-1} \frac{\partial^2 l}{\partial \theta \partial \theta'} \ |_{\theta=\theta'}$$

Substituting, we get:

$$\hat{\theta} \sim N(\theta_0, E(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)')$$

and where the variance term is approximately given by the second derivatives of the likelihood:

$$E(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \cong \sum \frac{\partial^2 l}{\partial \theta \partial \theta'}$$

Alternatively, the variance term can be calculated as:

$$E(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \cong T^{-1} \sum h(\hat{\theta}, Y) h(\hat{\theta}, Y)'$$

where we have:

$$h(\hat{\theta}, Y) = \frac{\partial \log f(y_t \mid y_{t-1}, y_{t-2}, \ldots}{\partial \theta} \mid_{\theta=\hat{\theta}}$$

We can use these statistics to calculate standard errors of parameters.

Likelihood Ratio Tests

A general method of testing restrictions is using LR (likelihood ratio) tests. Asymptotically, it is often the case that:

$$2(l(\hat{\theta}) - l(\theta)) \approx \chi^2(m)$$

where $\hat{\theta}$ is the optimized paramter vector, θ is the restricted parameter vector, and m is the number of restrictions that are being tested.

Diagnostic Statistics

Typically, we when we fit ARMA models, we will not know the data generating process. Therefore we will have to make an initial guess regarding the type of model, and then test this guess. This can be boiled down as follows:

(1) Guess the model
(2) Estimate the parameters
(3) Assess the adequancy of the model

Guessing the type of model

An important principle to keep in mind is simplicity: it is typically better to consider simple models over complicated models. We can make an initial guess from the autocorrelation and partial autocorrelation functions.

Autocorrelation

Recall that MA processes have autocorrelation functions that are zero beyond their lag order. Thus, an MA(q) process has non-zero autocorrelation for the first q lags. In contrast, AR

processes have non-zero autocorrelation (in principle) at all lags.

Partial Autocorrelation

The partial autocorrelation function is related to the autocorrelation function. Consider fitting a regression of y_t on $y_{t-\tau}$, for $\tau = 1, 2, 3, \ldots$. It turns out that for the AR model, this function looks like the autocorrelation function for an MA model, and that for an MA model, this function looks like an autocorrelation function for an AR model.

Example 1:

The sample partial autocorrelation at lag 1 for a stochastic process is given by:

$$\frac{\sum u_t u_{t-1}}{\sqrt{\sum u_t^2 \sqrt{\sum u_{t-1}^2}}}$$

where $\{u_t\}$ is obtained from the following OLS regression:

$$u_t = y_t - \hat{\mu}_t - \hat{\rho} y_{t-1}$$

The sample autocorrelation at lag τ is obtained in an analgous way by adding in extra lagged terms in the following regression:

$$u_t = y_t - \hat{\mu}_t - \sum_{\tau} \hat{\rho}_\tau y_{t-\tau}$$

We then take the residuals from this OLS regression, and calculate their autocorrelations at different lags.

Testing for residual autocorrelation

If you have estimated a decent model for the process, then the residuals from the process should be white noise. In other words, there should be no autocorrelation in those residuals. A simple approach is to graph the autocorrelations of the residuals, and visually inspect them to see if there is substantial autocorrelation. A formal statistical test for white noise is the Box-Ljyung test. This is given as:

$$Q = T(T+2)\sum_{\tau}^{P}\frac{r_{\tau}^{2}}{T-1}$$

where T is the number of observations and P is the number of autocorrelations being tested, and r is the autocorrelation. Under the null hypothesis of white noise, then the test statistic Q is distributed as a χ^2 random variable with P-p-q degrees of freedom, where p is the order of the AR component of the model, and q is the order of the MA component of the model.

Information Criteria

One shortcoming of the previous approach to diagnostic checking is that it is possible to come up with a number of possible models that pass the residual autocorrelation test. An alternative approach to diagnostic checking is to use information criteria that discriminate between different types of models. The two most populare are the AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion).

Both criteria maximize the likelihood, but penalize for the number of parameters used. The AIC is chosen by minimizing

$$AIC = -2l(\psi) + 2n,$$

where $l(\psi)$ is the maximized value of the likelihood and n is the number of parameter used. The BIC is given by

$$BIC = -2l(\psi) + n\log(T),$$

where T is the sample size. The BIC tends to have better asymptotic properties. Note that you should choose T so that it is a common number that can be used over different models. For example, if you consider 3 models - AR(1), AR(2), and AR(3), choose T so that it is the same number for each of these test statistics. Thus, if you had 100 observations, the maximum usable sample for the AR(3) would 97, as you lose the first 3 observations. Thus, T would be a maximum of 97 for each of the models.

Achieving Stationarity: Part 1

Economic Time Series often violate our assumption of covariance stationarity. In particular, their mean is typically changing over time. Thus, the average value of GDP in the U.S. in the 1990s is much higher than the average value of U.S. GDP 100 years ago.

For the time being, we will deal with this type of nonstationarity simply by using stationary-inducing transformations of the data. We will now consider 2 of these transformations. But before we develop these transformations, a preliminary transformation to use is to take logs of the time series, unless they already are in logged form (e.g. interest rates). This is useful, since the time series typically are growing, and also is a useful way of dealing with certain types of heteroskedasticity.

First-differencing

The first approach we will consider is to take first differences. Thus, after taking logs, simply define a new variable, Δy_t, where it is defined as:

$$\Delta y_t = y_t - y_{t-1}$$

Given that we have logged the variable, note that this transformation measures the growth rate in the variable. This type of transformation almost always induces stationarity for processes that have means (in log levels) that change over time in a systematic way (e.g. trends).

To understand this, note that the log-difference transformation of a variable represents that variable in terms of its growth rates - that is, log-differencing real GNP yields the growth rate of GNP. Most growth rates of economic variables are stationary.

Removing Deterministic Trend Components

An alternative approach to inducing stationarity for processes that grow over time is to remove deterministic trend from their logged values. Removing a linear trend means taking the residuals from the following regression:

$$u_t = y_t - \hat{\mu} - \hat{\alpha} t$$

In addition to removing linear trends, one may also add a quadratic, cubic, etc. terms to this regression. In practice, removing these higher order terms is not commonly done.

Hodrick-Prescott Filtering

A third approach is to use the Hodrick-Prescott filter, or some *band-pass filter* to take out the nonstationarities. A thorough discussion of these methods requires getting into the frequency doman - rather than the time domain - which we will not do at this point. From a practical point of view, however, we would take the residuals as follows:

$$u_t = y_t - y_t^{HP}$$

where y_t^{HP} is the trend component of the time series as identifed by the HP procedure. For a complete discussion of the HP filter see Hodrick and Prescott's article in the *Journal of Money,*

Credit, and Banking.

All of these procedures induce stationarity by removing growth components from the time series. The first-differencing approach does it by forming growth rates. The deterministic trend approach does it by taking out a linear trend. The HP approach does it by taking out a non-linear trend. Often, the detrending approach used will be dictated by the specific question that you are addressing. For example, if you are interested in business cycle fluctuations - that is, deviations from the long-run trend - then it is probably better to use the linear or HP approach. There are some issues that arise in choosing a trend specification for the time series, but we will defer that to the end of class when we revisit nonstationarity.

Vector Autoregressions (VARs)

So far we have focused on modelling univariate time series. We will now turn to modelling multivariate time series using a reduced form technique known as vector autoregressions, or VARs. The idea is to fit autoregressions to a vector-valued time series process. Alternatively, we could have fit ARMa processes to the vector valued time series - these are called VARMA models. However, the VAR is much easier to estimate than the VARMA, so the profession uses the VARMA model very little, if at all.

To begin, let's begin with some notation:

y_t is an N x 1 vector of economic time series, μ is an N x 1 vector of constant terms, and ε_t is an N x 1 vector of white noise variables, and Φ is an N x N matrix of coefficients.

ε_t has the following properties:

$$E(\varepsilon_t) = 0, E(\varepsilon_t^2) = \Omega, E(\varepsilon_t \varepsilon_\tau) = 0, t \neq \tau$$

Note that Ω is a covariance matrix.

Using lag operator notation, the VAR(p) model can be written as:

$$(I_N - \Phi_1 L - \ldots - \Phi_p L^p) y_t = \mu + \varepsilon_t$$

or

$$\Phi(L) y_t = \mu + \varepsilon_t$$

Stationarity for vector-valued processes

A vector valued process is covariance stationary if its first and second moments are independent of calandar time. To establish the necessary conditions for stationarity, it is useful to write the VAR(p) process as a first order process. This can be done by "stacking" variables.

Let ξ_t be an Np x 1 vector, stacked as follows:

$$\xi_t = \begin{bmatrix} y_t - \mu \\ y_{t-1} - \mu \\ \vdots \\ y_{t-p+1} - \mu \end{bmatrix} = F\xi_{t-1} + v_t, v_t = \begin{matrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{matrix}$$

Note that F is given by:

$$\begin{bmatrix} \Phi_1 & \Phi_2 & \Phi_3 & \dots & \Phi_p \\ I_n & 0 & 0 & \dots & 0 \\ 0 & I_n & & & \\ \vdots & 0 & I_n & & \\ 0 & & & & \end{bmatrix}$$

Similarly, we have:

$$E(v_t v_t') = Q$$

where Ω is given by:

$$Q = \begin{bmatrix} \Omega & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 \\ \vdots & 0 & 0 & \\ 0 & 0 & & 0 \end{bmatrix}$$

The eigenvalues of the matrix satisfy:

$$| \; I_n\lambda^p - \Phi_1\lambda^{p-1} - \Phi_2\lambda^{p-2} - \dots - \Phi_p \; | = 0$$

This means that the VAR(p) is stationary as long as all λ that satisfy this determinental equation are less than one in absolute value.

This also implies that for

$$| \; I_n - \Phi_1 z - \Phi_2 z^2 - \dots - \Phi_p z^p \; | = 0,$$

all values of z lie outside the unit circle.

Example:

Suppose we have a two-variable, one-lag model. Then we have:

$$| \ I_n - \Phi_1 z \ | = 0$$

Suppose further that Φ_1 is:

$$\Phi_1 = \begin{bmatrix} .9 & .00 \\ .00 & .9 \end{bmatrix}$$

Then we need to find the values for z such that:

$$\det(I_2 - z * \Phi_1) = 0$$

It turns out that there is one value of z that satisfies this equation, which is $.9^{-1}$. Thus, the process is stationary.

We can also write the MA representation for this vector-valued process:

$$y_t = \mu + \varepsilon_t + \Psi_1 \varepsilon_{t-1} + \Psi_2 \varepsilon_{t-2} + \ldots$$

where

$$\Psi(L) = \Phi(L)^{-1}$$

This requires

$$[I_n - \Phi_1 L - \Phi_2 L^2 - \ldots - \Phi_p L^p][I_n + \Psi_1 L + \Psi_2 L^2 + \ldots] \ = I_n$$

Note that one can solve for the moving average coefficients as follows. First consider the case for the first lag. Since this must hold for all possible coefficient values, then we must have:

$$\Psi_1 = \Phi_1$$

Similarly, for the second lag, we must have:

$$\Psi_2 = \Phi_1\Psi_1 + \Phi_2$$

One can solve out for the other lags in an analagous fashion.

There is a key difference between the univariate case and the vector-valued case regarding autocovariances. Recall that in the stationary case with a univariate process, we had:

$$\gamma_j = \gamma_{-j}$$

This no longer need hold in the vector case. Instead, we have:

$$\Gamma_j' = \Gamma_{-j}$$

where Γ is the autocovariance matrix for the vector process. It can be shown that in the stacked, first order form, the autocovariances are given by:

$$\Gamma_j = \Phi_1\Gamma_{j-1} + \Phi_2\Gamma j - 2 + \ldots \Phi_p\Gamma_{j-p}$$

MLE for VARs

As in the case for univariate autoregressions, it is simple to conduct MLE for VARs using the conditional likelihood. The conditional likelihood for the VAR is very similar to that of the univariate case, with the exception that we no longer have a single innovation variance, but rather have a covariance matrix that we need to include as part of the likelihood. Rather than re-deriving the likelihood as we did previously, I will jump to the main result. The log likelihood for the VAR is given by:

$$l = \frac{-Tn}{2}\log(2\pi) + \frac{T}{2}\log \mid \Omega^{-1} \mid -\frac{1}{2}\sum[(y_t - \Pi' x_t)'\Omega^{-1}(y_t - \Pi' x_t)]$$

where x is given by:

$$x_t = [1\ y_{t-1}\ y_{t-2}\ \ldots]'$$

and where Π is an N x (Np+1) coefficient matrix, and where $|\ \Omega^{-1}\ |$ is the determinant of the inverse of the covariance matrix.

Maximizing the log of the likelihood with respect to Π yields the standard OLS formula:

$$\Pi_{ols} = \sum y_t x_t' [\sum x_t x_t']^{-1}$$

The MLE of Ω is also given by the usual formula. In particular, maximizing the likelihood and setting the result to 0 yields:

$$\Omega_{ols} = \frac{1}{T} \sum u_t u_t'$$

where u_t is the residual vector from the OLS estimation.

Hypothesis Testing

As in other models, we can use the maximized likelihood to conduct Likelihood Ration (LR) tests. The empirical likelihood of the model boils down to:

$$l = -(Tn/2)\log(2\pi) + T/2 \log\ |\ \Omega^{-1}\ | - Tn/2$$

For any test, we can form two log-likelihoods - one for a model of interest, and one for an alternative model. For example, suppose we wanted to test between a VAR(1) and a VAR(2). To do this, we form the following test statistic:

$$2(l_2 - l_1) = T\{\log\ |\ \Omega_2^{-1}\ | - \log\ |\ \Omega_1^{-1}\ |\}$$

Under the null hypothesis, this statistic is asymptotically distributed as a χ^2 random variable with degrees of freedom equal to the number of restrictions imposed under the null hypothesis. In this case, the number of restrictions is equal to N - that is, we are testing a VAR(2) vs. a VAR(1), with a VAR that includes N variables.

Chris Sims advocates making a modification to this statistic for use in finite samples:

$$(T - (Np + 1))\{\log\ |\ \Omega_2^{-1}\ | - \log\ |\ \Omega_1^{-1}\ |\}$$

The Uses of VARs

We will now discuss some of the uses of VARs. We start with the uncontroversial uses that related to VARs as **reduced form models.**

Data summary:

We often are interested in summarizing the dynamic relationships between data. One way to do this is by fitting a VAR to the data. This can also be useful, because most dynamic equilibrium models - when log-linearized - posses a **restricted** VAR representation. Thus, if one wants to "test" an equilibrium model to an unrestricted VAR, it can be done using the LR test. (Of course, it may not be very informative to test an equilibrium model this way - this is because all theoretical modes are abstractions, and we thus would be suprised if the economic model was not rejected in favor of an unrestricted reduced form model).

A particularl type of data summary is called **Granger-Sims Causality.**

This is a fairly simple issue that took a long time for the profession to sort out. First, it has nothing to do with causality in the literal sense. Instead, this is a statistical test for determining whether changes in one time series variable tend to help forecast another variable. In addition to its use in forecasting, it also has implications for certain types of rational-expectations models. We will first describe the issue, how to test for Granger-causality, and then describe some of its uses.

We will begin with a bivariate VAR, with two variables, y and x. The variable x is said to Granger-cause y if the mean squared forecast error of y conditioning on past y and past x is significantly smaller than if we only condition on past y. This amounts to testing whether lagged values of the variable x are statistically significant in a regression of y on past y and past x.

In other words, x fails to Granger-cause y if in the following linear projection:

$$y_t = \sum_{i=1} \alpha_i y_{t-i} + \sum_{i=1} \beta_i x_{t-i} + \varepsilon_t,$$

the coefficients β_i are all zero.

Testing for Bivariate Granger-Causality

We now describe how to test for Bivariate Granger-causality.

First, choose a VAR between x and y, and choose the lag length - that is, the number of lagged variables entering each equation. In the liteature, the lag order is typicallythe same for

each variable, though there is no specific reason why this need be the case. (One of the pitfalls of reduced form modelling is that no theory is being used to restrict the model....such as in terms of lag length...)

The model is given by:

$$y_t = \mu + \sum_{i=1}^{p} \alpha_I y_{t-i} + \sum_{i=1}^{p} \beta_i x_{t-i} + \varepsilon_t$$

We wish to test the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

We can test this either using an F-test.

To form the test statistic, we calculate the residual sum of squares from the model, which we denote as RSS_1. We then form the residual sum of squares from the model with the restrictions imposed - that is, with the β_i coefficients restricted to be zero. We denote this residual sum of squares as RSS_2.

We next form the test statistic:

$$\frac{(RSS_2 - RSS_1)/p}{RSS_1/(T - 2p - 1)}$$

We then check whether the value for this test statistic exceeds the critical value for the size of the test we chose - that is, we chose a 5% test, etc. If the value of the statistic exceeds the critical value, then we reject the null hypothesis that the variable x fails to Granger-cause y. This is because the improvement in the residual sum of squares with lagged values of x is statistically significant.

Clearly, Granger-causality has a role in forecasting. In particular, it tells us whether adding lagged values of a variable significantly improve prediction performance. Granger-Causality also has implications for certain forms of rational expectations models.

Example: Efficent Markets and Risk Neutrality

Consider the Lucas (1982) intertemporal asset pricing model, with risk-neutral households. Households have shares (claims) to an endowment stream, and can trade these shares at a

competitive price, p.

The representative household's preferences and budget constraint are given by:

$$\max E \sum \beta^t u(c_t)$$

subject to:

$$s_t(p_t + d_t) \geq c_t + p_t s_{t+1}$$

where s is the number of shares, p is the share price, and d is the dividend per share. Assume that the dividend is generated by a stationary AR(1) process. The equilibrium condition for this economy is given by:

$$u_{ct} = \beta E_t(u_{ct+1}(p_{t+1} + d_{t+1}))/p_t$$

Assuming that u(c) = c, and solving this equation forward, we obtain:

$$p_t = E_t \sum_{j=t+1} \{\beta^{j-t} d_j\}$$

This model implies that no variable should Granger-cause stock prices. Clearly, this is a testable implication of the model. Note, however, that this test makes a very strong maintained assumption - that is, households are risk neutral. Thus, any test of efficient markets models using Granger-causality is a joint test of the maintained assumption about the structure of the model, as well as the model's implication.

Note that since stock prices in this model embody information about future variables, they will tend to be good predictors of economic fundamentals. This is one reason why economists sometimes use stock prices or other asset prices to predict changes in variables like GNP.

Forecasting:

It seems plausible that forecasts for a variable like GNP can be improved by using information beyond that contained in the past values of GNP. For example, one might be able to improve on a univariate forecast by using a variable like consumption, interest rates, or productivity. VARs can be such a tool. We will later learn a Bayesian procedure for forecasting with VARs that produces excellent forecasting results.

Bayesian VARs

The work of Charles Nelson (AER, 1972) showed that univariate ARMA models for series like GNP and other macro variables produce superior forecasts to those from large scale, Keynesian macroeconomic models. It seems sensible to assume that as we bring in more information into a model, we should get better forecasts than if we just consider univariate models.

VARs are the obvious approach to forecasting. But it turns out that these models tend to produce **worse** forecasts than univariate ARMA models. The reason is because they are unrestricted models, and many of the variables turn out to be significantly related - which is called multicollinearity. This is the reason that unrestricted VARs often produce bad forecasts.

So we need to place some restrictions on the system. The most popular way is to use the pseudo-Bayesian procedure of Litterman. Bayesian analysis combines prior information with sample information.

He chooses a prior that each variable in the VAR is a random walk with drift:

$$y_t - y_{t-1} = \mu + \varepsilon_t$$

This is useful since we know that many economic time series are well approximated by this process. This means that the mean of the prior distribution of the coefficients for the first lag of each own variable is 1, and the mean of the prior distribution of all other coefficients is 0.

Litterman assumes that the covariance matrix for the prior distribution is diagonal, with standard deviation of γ for the coefficient on the first lag. For the first coefficient on each own lag, this means the prior is:

$$\phi_{ii}^{1} \sim N(1, \gamma^2)$$

For the other coefficients on own lags, he argued that we should have a tighter prior. He suggested :

$$\phi_{ii}^{s} \sim N(0, (\gamma/s)^2)$$

For the coefficients on lagged values of other variables, he suggested a prior standard deviation of:

$$\frac{w\gamma\tau_i}{s\tau_j}$$

Here, the term τ_i/τ_j corrects for differences in the scale of variables, and s is the lagged order of the variable. Litterman uses the standard deviations of the residuals from an AR(p) fit to the individual variables for the τ terms. The parameter w provides an alternative weighting relative to own lags. The idea is that own lagged values will be of more help in forecasting than lagged values of other variables. This means that the parameter w should be less than 1.

Litterman's Bayesian procedure is fully automated in some computer packages, including

RATS (which was written by Litterman). Forecasts produced for GNP and other macro variables from this procedure tend to outperform those from large-scale econometric model and univariate ARMA forecasts.

You can find a detailed discussion of this procedure in:
http://www.minneapolisfed.org/research/qr/qr843.pdf

Using VARs for other Purposes

The uses of VARs so far are quite uncontroversial. That is, using them for summarizing data and for forecasting. We now describe some of the controversial uses of VARs.

Impulse Response Analysis

One question we often ask is: "What happens to variables like GNP, employment, exchange rates, etc., when some shock hits the economy?" Some economists argue that VARs can be used to answer these questions. Let's look at how the VAR analysts proceed.

Keep in the back of your mind 2 requirements for this exercise:

We need to economically define what the shock is.

We need to econometrically identify the impact of the shock on the variable of interest.

We first write the model as a Vector Moving Average model:

$$y_t = \varepsilon_t + \Psi_1 \varepsilon_{t-1} + \ldots.$$

Note that we have:

$$\frac{\partial y_{t+s}}{\partial \varepsilon_t'} = \Psi_s$$

Note also the following. Suppose that all N innovations increase by some specific amount - call it δ_i- at the same time. Then we would have:

$$\Delta y_{t+s} = \frac{\partial y_{t+s}}{\partial \varepsilon_{1t}} \delta_1 + \frac{\partial y_{t+s}}{\partial \varepsilon_{2t}} \delta_2 + \ldots$$

This just says that the change in y at date t+s as a consequence of innovations at date t of a particular size is generated by the moving average coefficients multiplied by the innovations.

Now, a plot of the row i, column j element of Ψ_s as a function of s is called the impulse response function. It is the response of the variable of interest to a one-time impulse to another variable, with all other variables held constant.

The tricky part is how to interpret this impulse response. It is tempting to say that it measures the effect of one variable on another. There is only one case in which it does (at least, unambiguously). This is the case when Ω is a diagonal matrix.

To see this, consider slightly different question. First, suppose that we know that today's value of the y_{1t} is higher than expected. How does this information cause us to revise our forecast of $y_{i,t+s}$?

In particular, is this revision given by

$$\frac{\partial E(y_i, t+s \mid y_{1t}, x_{t-1})}{\partial y_{1t}} = \frac{\partial y_{i,t+s}}{\partial \varepsilon_{1t}}?$$

This will be the case if the covariance matrix is diagonal. But if it is not diagonal, then our revision of the forecast is not given by this MA coefficient alone. This is because if the innovations are correlated, then a change in the innovation in equation 1 will tend to be associated with changes in the innovations in other equations, given the nature of the covariance matrix. Thus, we would want to take into account this additional information in changing the forecast in response to the new information.

Given the fact that the innovations will likely be correlated, let's consider an alternative exercise. Suppose we want to ask how much the variable of interest changes in response to a change in the second variable in the system, **conditional on knowing how much the first variable changed:**

$$\frac{\partial E(y_i, t+s \mid y_{2t}, y_{1t}, x_{t-1})}{\partial y_{2t}}$$

Similarly, one can construct:

$$\frac{\partial E(y_i, t+s \mid y_{3t}, y_{2t}, y_{1t}, x_{t-1})}{\partial y_{3t}}$$

Calculating these recursive formulae require the moving average representation and also the covariance matrix. This can be done by factorizing the covariance matrix as follows. This will involve turning the correlated innovations into orthogonal innovations. We will later see

how this is useful for answering the question we initially posed.

Consider our covariance matrix, Ω. It turns out that there exists a unique factorization of this real symmetric (and positive definite) matrix that diagonalizes the matrix:

$$\Omega = ADA'$$

where A is a lower triangular matrix with the value 1 along the principal diagonal, and D is a diagonal matrix with positive entries along the principal diagonal. Given this factorization, we can define a new object:

$$u_t = A^{-1}\varepsilon_t$$

The variables in the vector u_t are uncorrelated with each other. To see this, note the following:

$$E(\mathrm{u}u') = E(A^{-1}\varepsilon_t\varepsilon_{t'}A^{-1\prime}) = A^{-1}\Omega A^{-1\prime}$$

We now need to show that this is a diagonal matrix. To see this, note:

$$A^{-1}\Omega A^{-1\prime} = A^{-1}ADA'A^{-1\prime} = D$$

Recall that D is a diagonal matrix, so these variables are indeed uncorrelated. This means we can answer our earlier forecasting question in a simple way.

Note first that:

$$\varepsilon_t = Au_t$$

For example, in a 3-variable system, this means:

$$\begin{bmatrix} 1 & 0 & 0 \\ a_{21} & 1 & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix}\begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{bmatrix} = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix}$$

Thus, $u_{1t} = \varepsilon_{1t}$, while $u_{2t} = \varepsilon_{2t} - a_{21}u_{1t}$. More generally, we will have the following

recursive orthogonalized system:

$$u_{jt} = \varepsilon_{jt} - \sum_i a_{ij} u_{it}$$

Note that since the u_{jt} are uncorrelated, then the projection of ε_{jt} on u_{1t} is equal to a_{j1}. We are now in a position to see how new information about ε_{1t} leads us to revise our forecast of ε_{jt} :

$$\frac{\partial E(\varepsilon_{jt} \mid \varepsilon_{1t})}{\partial \varepsilon_{1t}} = \frac{\partial E(\varepsilon_{jt} \mid u_{1t})}{\partial u_{1t}} = a_{j1}$$

Now, we can combine these into a vector to get:

$$\frac{\partial E(\varepsilon_t \mid y_{1t}, x_{t-1})}{\partial y_{1t}} = A_1 = \begin{bmatrix} 1 \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix}$$

Now we are in a position to determine how a change in ε_{1t} changes our forecast of y :

$$\frac{\partial E(y_{t+s} \mid y_{1t}, x_{t-1})}{\partial y_{1t}} = \Psi_s A_1$$

Similarly, we have:

$$\frac{\partial E(y_{t+s} \mid y_{2t}, y_{1t}, x_{t-1})}{\partial y_{2t}} = \Psi_s A_2$$

and generally, we have:

$$\frac{\partial E(y_{t+s} \mid y_{jt}, y_{j-1,t}, \ldots, x_{t-1})}{\partial y_{jt}} = \Psi_s A_j$$

Thus, the way in which we revise a forecast for a vector-valued time series involves not only the moving average coefficients, but also the correlation between the innovations. Note that the Ψ matrix captures the MA terms, while the A vector captures the correlation between the innovations.

Exercise: Show that if the ε_t are uncorrelated, then the A vectors have zeros in all elements

but one.

A plot of $\Psi_s A_j$ as a function of s is called the **impulse response function.** It desribes how a one-time orthogonalized innovation affects our forecast of the evolution of the vector of economic variables over time.

Keep in mind that these impulse response functions determined in part by how we recursively orthogonalized the innovations. We will return to this issue shortly.

A common approach to recursively orthogonalized these innovations is through a factorization called the *Cholesky Decomposition.*

The Cholesky decomposition factors the covariance matrix as follows:

$$\Omega = PP' = A\sqrt{D}\,\sqrt{D}\,A'$$

where $P = A\sqrt{D}$, and P is lower triangular and has the standard deviations of the u's on its prinicpal diagonal.

We threfore have as orthogonalized innovations:

$$v_t = P^{-1}\varepsilon_t = D^{-1/2}u_t$$

We can now obtain how our forecast revision works in response to new information:

$$\frac{\partial E(y_{t+s} \mid y_{jt}, y_{j-1,t}, \ldots, x_{t-1})}{\partial y_{jt}} = \Psi_s p_j,$$

where p_j is the jth column of P.

Fundamental and Nonfundamental Moving Average Representations

One issue to keep in mind is that the MA representation we have derived for the VAR is know as the *Fundamental Represenation.* This representation has the property that the coefficient matrix on the contemporaneous term is the identity matrix:

$$y_t = \mu + \varepsilon_t + \Psi_1\varepsilon_{t-1} + \ldots$$

There are nonfundamental representations as well. To see this, consider the premultiplying

ε_t by an n x n non-singular matrix, H:

$$u_t = H\varepsilon_t$$

Note that u is also white noise. Now we can write the MA representation as:

$$y_t = \mu + H^{-1}H\varepsilon_t + \Psi_1 H^{-1}H\varepsilon_{t-1} + \ldots$$

or:

$$y_t = \mu + J_0 u_t + J_1 u_{t-1} + \ldots$$

Thus it is possible to represent the vector in terms of the non-fundametnal representation, where this nonfundamental representation has a coefficient matrix on the contemporaneous term that is not the identity matrix.

Thus, to obtain the fundamental representation, we impose the nomralization that $\Psi_0 = I_n$.

Variance Decomposition

Another use of the VAR is to decompose forecast error variance into the orthogonalized innovations.

We begin with the forecast errors:

$$y_{t+s} - \hat{y}_{t+s} = \varepsilon_{t+s} + \Psi_1\varepsilon_{t+s-1} + \ldots + \Psi_{s-1}\varepsilon_{t+1}$$

the mean squared error of this s-period forecast is:

$$MSE = \Omega + \Psi_1\Omega\Psi_1' + \ldots$$

Now, let's see how we can decompose this forecast error variance using the orthogonalized innovations.

Recall:

$$\varepsilon_t = Au_t$$

This implies we have:

$$\Omega = A_1 A_1' var(u_{1t}) + A_2 A_2' var(u_{2t}) + \ldots$$

this lets us write the mean square forecast error as:

$$MSE = \sum_j var(u_{jt}) \{A_j A_j' + \Psi_1 A_j A_j' \Psi_1' + \Psi_2 A_j A_j' \Psi_2' + \ldots + \Psi_{s-1} A_j A_j' \Psi_{s-1}'\}$$

It follows that the contribution of the jth orthogonalized innovation to the MSE is:

$$var(u_{jt}) \{A_j A_j' + \Psi_1 A_j A_j' \Psi_1' + \Psi_2 A_j A_j' \Psi_2' + \ldots + \Psi_{s-1} A_j A_j' \Psi_{s-1}'\}$$

Moreover, it is useful to calculate the percent of the MSE, which requires normalizing the above expression by the MSE:

$$\frac{var(u_{jt}) \{A_j A_j' + \Psi_1 A_j A_j' \Psi_1' + \Psi_2 A_j A_j' \Psi_2' + \ldots + \Psi_{s-1} A_j A_j' \Psi_{s-1}'\}}{\sum_j var(u_{it}) \{A_j A_j' + \Psi_1 A_j A_j' \Psi_1' + \Psi_2 A_j A_j' \Psi_2' + \ldots + \Psi_{s-1} A_j A_j' \Psi_{s-1}'\}}$$

For the commonly used Cholesky decomposition, we obtain the following expression for the MSE in terms of orthogonalized innovations:

$$MSE = \sum_j \{p_j p_j' + \Psi_1 p_j p_j' \Psi_1' + \Psi_2 p_j p_j' \Psi_2' + \ldots + \Psi_{s-1} p_j p_j' \Psi_{s-1}'\}$$

Using VARs to structurally interpret economic time series

So far, we have considered *atheoretic* VARs as simply a way of summarizing information between the current values of economic variables and their lagged values. We write the reduced form VAR model as:

$$y_t = \mu + \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \varepsilon_t$$

Recall that reduced form models are in principal consistent with many different behavioral models. Picking out any specific structural model requires **identifying restrictions**. We shall now see how some VAR analysts come up with such restrictions.

Many dynamic behavioral economic models (after being log-linearized) can be written as follows:

$$B_0 y_t = \eta + B_1 y_{t-1} + \ldots + B_p y_{t-p} + u_t$$

where y is an N x 1 vector of variables, the B's are coefficient matrices, and the u's are white noise innovations with a diagonal covariance matrix.

Now, premultiply the behavioral model by B_0^{-1} :

$$y_t = B_0^{-1}\eta + B_0^{-1}B_1 y_{t-1} + \ldots + B_0^{-1}B_p y_{t-p} + B_0^{-1}u_t$$

Clearly, there is a mapping between the behavioral model and the reduced form VAR,

where:

$$\Phi_j = B_0^{-1}B_j$$
$$\mu = B_0^{-1}\eta$$
$$\varepsilon_t = B_0^{-1}u_t$$

This mapping has led a number of researchers to use VARs to structurally interpret economic time series.

Intepreting the Impulse Responses Recursively

In the economic model, we may wish to understand how structural shocks to the economy - that is, shocks to preferences, technologies, endowments, government policies, etc. - affect economic variables like output, consumption, investment, and employment.

Recall that these structural shocks are the u's in the behavioral economic model.

When we estimate the plain VAR, however, we don't recover the u's, but rather, we obtain estimates of the $\varepsilon's$.

Also recall that the mapping between the **reduced form innovations and the structural innovations** was given by:

$$u_t = B_0\varepsilon_t$$

Generally, how do we map between the reduced form and the behavioral model? This requires **identifying restrictions**

In this case, the mapping coincides with a **just-identified model:**

Now, recall how we orthogonalized the innovations previously, with the lower triangular matrix A:

$$u_t = A^{-1}\varepsilon_t$$

Note that the behavioral model coincides with the orthogonalized VAR **if**:

$$B_0 = A^{-1}$$

This requires:

(1) B_0 is lower triangular
(2) The triangularized ordering chosen for orthogonalization is identical to the economic structure of B_0
(3) The coefficients along the principal diagonal of B_0 are all 1

To see that this is a just-identified model, note that any positive definite matrix (Ω) has a unique lower triangular decomposition with 1's along the principal diagonal and a diagonal matrix D such that $\Omega = ADA'$. Thus unique values of B_0 and D can be found that satisfy the conditions.

Note that there are no testable restrictions imposed by this identification: that is, there are no overidentifying restrictions that can be tested.

How many ways are there to achieve this recursive identification?

There are n! ways of ordering the variables for orthogonalizing the VAR, which implies that there are n! possible identifications of the model.

A number of economists have used this recursive approach to identifying VARs to use them to structurally interpret data.

This requires the imposition of an economic theory that yields the triangular form for B_0 and that tells you how the variables are ordered.

Nonrecursive identification and estimation of VARs

There are other identifications for VARs, in addition to recursive identifications.

The basic idea is to continue to postulate the behavioral form discussed above, with innovations that have a diagonal covariance matrix, and then find a set of restrictions such that:

$$B_0^{-1} D B_0^{-1\prime} = \Omega$$

Note that the recursive orderings described above achieve this - but now we will consider other types of restrictions that achieve this as well. This amounts to finding a nonsingular matrix B_0 that diagonalizes the covariances matrix.

We now discuss how these models are estimated and identified.

First, consider estimation conditional on identification. The log-likelihood of the identified model (ignoring constants) is:

$$l = -(T/2)\log \mid B_0^{-1} D B_0^{-1\prime} \mid -1/2 \sum (y_t - \Pi' x_t)' (B_0^{-1} D B_0^{-1\prime})(y_t - \Pi' x_t)$$

This becomes:

$$l = -(T/2)\log \mid B_0^{-1} D B_0^{-1\prime} \mid -1/2 \sum \varepsilon_t' (B_0^{-1} D B_0^{-1\prime}) \varepsilon_t$$

Note that the latter term in the likelihood is:

$$\sum \varepsilon_t' (B_0^{-1} D B_0^{-1\prime}) \varepsilon_t = \sum trace(\varepsilon_t' (B_0^{-1} D B_0^{-1\prime}) \varepsilon_t)$$

Simplifying, it turns out that the likelihood becomesm (omitting constants):

$$l = T/2 \log \mid B_0 \mid^2 - T/2 \log \mid D \mid -(T/2) trace\{(B_0' D^{-1} B_0)\Omega)$$

Assuming that there are unique matrices B_0 and D satisfying the covariance matrix factorization of Ω, then it turns out that maximizing the likelihood yields estimates of B_0 and D that satisfy:

$$B_0^{-1} D B_0 = \Omega$$

In general, this is a nonlinear system of equations that can be solved using numerical methods.

Identification of non-recursive VARs

Identification requires both an order and rank condition.

The order condition is straightforward, and boils down to the number of informative elements of the covariance matrix Ω. Since this matrix is symmetric, it has N(N+1)/2 distinct elements. Note that the information in Ω must be able sufficient to identify both B_0 and D.

Since D is diagonal, it requires N parameters. Thus, we can in principal identify N(N+1)/2 parameters in B_0.

The rank condition is presented in detail in Hamilton, pp. 333-334.

Using Long-Run Restrictions to identify non-recursive VARs

We now consider one specific approach to identifying VARs, which is the long-run restriction approach of Blanchard and Quah (American Economic Review, 1989, vol. 79, no. 4, p. 655.)

They consider a 2-variable model

(unemployment and output) in which there are two types of shocks - 1 shock has only transitory effects on both variables, and the other shock has only transitroy effects on unemployment, but may have permanent effects on output. In addition, the 2 shocks are uncorrelated.

Their behavioral model is:

$$X_t = A_0 \varepsilon_t + A_1 \varepsilon_{t-1} + \ldots, var(\varepsilon) = I_2$$

where X is a vector that includes the unemployment rate and the growth rate of output.

They assume that the first term in the ε vector is the transitory shock. Since this shock has no effect on the level of output in the long run, it must be the case that the sum of the moving average coefficients on this shock in Δy_t must be zero:

$$\sum_{j=0}^{\infty} a_{11}(j) = 0$$

Now, consider the reduced form of this model:

$$X_t = v_t + C_1 v_{t-1} + \ldots, var(v) = \Omega$$

Note that:

$$v_t = A_0 \varepsilon_t, A_j = C_j A_0$$

and:

$$\Omega = A_0 A_0'$$

Note that all we need to do is figure out the elements of A_0, and we can recover everything else of interest.

Is A_0 identified? Based on the order conditinon, it is. To see this, note that the Ω identifies 3 distinct elements, which map into $A_0 A_0'$.. This gives us 3 restrictions on the 4 elements of A_0. The fourth restriction is given by:

$$\sum_{j=0}^{\infty} C_j A_0 = 0$$

Note that the one tricky issue is that identification requires estimating the infinite sum of the lag coefficients, C_j.

To do this in practice, we follow these steps:

(1) Estimate the VAR
(2) Obtain the MA representation

(3) Use the restrictions and the MA coefficients to obtain A_0

At this point, you have all the information.

Some Issues Associated with Structural VARs

VAR analysts use structural VARs for a number of purposes. One is to test economic hypotheses, another is to conduct policy analysis (See Sims, 1986 Federal Reserve Bank of Minneapolis Quarterly Review "Are Forecasting Models useable for Policy Analysis" - you can download this paper from www.minneapolisfed.org - go to the link for research, then go to the link for quarterly reviews.)

The main drawback to structural VARs is that there typically is not an explicit mappying between the VAR model and a fully articulated economic model. Thus, in the absence of a structural economic model, some economists, such as Robert Lucas and Edward Prescott, do not find VAR evidence very informative about testing economic hypotheses or about guiding economic policy. There view is a simple one: one cannot test economic hypotheses or usefully discuss economic policy advice without an explicit economic model. Other economists, such as Christopher Sims or Olivier Blanchard do find these models informative, despite the fact that there typically is no structural economic model that can be mapped into the VAR. (Note that this does not mean that it is not possible to find an explicit mapping between a fully parameterized economic model and the VAR, just that this is very rarely done in practice.)

Computing Confidence Intervals for Impulse Response Functions

Recall that the impulse response function is a complication non-linear function of the autoregressive parameters, and also involves correlations that arise from orthogonalization. Asymptotic formulae for standard errors for these functions are presented in Hamilton on pages 336-337. However, in small samples it turns out that bootstrapping works better. We will focus our discussion on Lutz Kilian, "Small Sample Confidence Intervals for Impulse Response Functions", Review of Economics and Statistics, pp.218-230, May 1998.

The problem that we face here is that the small sample distribution of impulse response functions may be significantly biased and skewed (even though asymptotically they are normal - see Hamilton). It is hard to correct the impulse responses directly for bias and skewness, because of their nonlinear nature. Kilian's idea is to remove the bias from the VAR coefficients prior to bootstrapping the estimate.

Signfiicant bias may be due to significant small sample bias from the VAR coefficients, or it may be due to nonlinearity, since the impulse responses are non-linear functions of the VAR coefficients. This latter point suggests that even small bias in the VAR coefficients can generate potentially large bias in the impulse response coefficients. So Kilian will first try to bias-correct the VAR coefficients.

Implementing Kilian's bootstrap

(1) Estimate the mean OLS bias by resampling.

First, estimate the VAR(p) model and generate 1000 bootstrap replications $\hat{\beta}^*$ from:

$$y_t^* = \hat{\mu} + \sum_{i=1}^{p} \hat{\beta}_i y_{t-i}^* + u_t^*$$

using non-parametric bootstrap techniques.

Recall how the bootstrap works. There are parameteric and non-parametric bootstraps. Parameteric bootstraps require distributional assumptions about a random variable that you are using in the bootstrap. Non-parametric do not. The non-parametric bootstrap in this case would be to take the sequence of the residuals: $\{u_t^*\}_{t=1}^{T}$. Note that we are assuming that these residuals are white noise. Note that if they were not, then we could add additional lags to the VAR until the residuals became white noise. Next, we reshuffle those residuals, using a random number generator. We denote these shuffled residuals as $\{u_t^i\}$Then we form:

$$y_t^i = \hat{\mu} + \sum_{i=1}^{p} \hat{\beta}_i y_{t-i}^i + u_t^i$$

Then we estimate a new set of VAR coefficients. We repeat this procedure the desired

number of times. Now, approximate the bias term:

$$\Psi = E(\hat{\beta} - \beta)$$

as:

$$\Psi^* = E^*(\hat{\beta}^* - \hat{\beta})$$

this yields a bias estimate of mean($\hat{\beta}^*$) – $\hat{\beta}$.

Step 2 Calculate the modulus of the largest root of the companion matrix of $\hat{\beta}$, and denote this as m($\hat{\beta}$). If m($\hat{\beta}$) ≥ 1, then set $\tilde{\beta} = \hat{\beta}$. If not, then construct the bias-corrected coefficient estimate $\tilde{\beta} = \hat{\beta} - \hat{\Psi}$.

Now, if the root is greater than or equal to 1, let $\hat{\Psi}_1 = \hat{\Psi}$ and set $\delta_1 = 1$. Next, define:

$$\hat{\Psi}_{i+1} = \delta_i \hat{\Psi}_i$$

$$\delta_{i+1} = \delta_i - .01$$

Set $\tilde{\beta} = \tilde{\beta}_i$ after iterating on $\hat{\beta} - \hat{\Psi}_i$ until m(β_i) < 1.

This stationarity correction is used to avoid pushing stationary impulse response functions into the nonstationary region. This correction does not have any asymptotic effects

Step 3 Substitute $\tilde{\beta}$ for $\hat{\beta}$ in the VAR equation and generate 2000 new bootstrap replications which we again denote as $\hat{\beta}^*$. To estimate the mean bias ($\hat{\Psi}^*$), we need to nest a separate bootstrap loop inside each of the 2000 bootstrap loops. This is computationally intensive, so a "short-cut" is used as follows. Use the first stage bias estimate $\hat{\Psi}$ as a proxy for $\hat{\Psi}^*$.

Step 4: Calculate $\tilde{\beta}^*$ from $\hat{\beta}^*$ and $\hat{\Psi}^*$, following the steps above.

Step 5: Calculate the perntile cut-offs - that is, choose α and $1 - \alpha$, picking a value (say .10) for α. Thus, we report the values for the impulse responses such that α and $1 - \alpha$ fraction fall within the confidence bands.

How well does Kilian's procedure work, both in absolute terms and in relative terms compared to asymptotics and other bootstrap procedures that don't correct for bias? His Monte Carlo work shows that it dominates. See his paper - we will go through some of this in class.

State Space Models and the Kalman Filter

The Kalman Filter is a very useful tool in time series analysis. For linear models, it permits the estimation of models with latent variables, MLE of ARMA models, models with time varying coefficients, etc. We begin with **state space models**. There are 2 key equations: the **Measurement Equation and the Transition Equation.**

The measurement equation is:

$$y_t = Z_t\alpha_t + d_t + \varepsilon_t, var(\varepsilon_t) = H_t$$

where y is an observable N x 1 vector, Z is an observable N x m matrix, d is an observalbe N x 1 vector, α is an m x 1 unobserved vector, and ε is a white noise vector with covariance matrix H. While the α vector elements are unobserved, we assume they are generated by a first order Markov process:

$$\alpha_t = \Gamma_t\alpha_{t-1} + c_t + R_t\eta_t, var(\eta_t) = Q_t$$

This is called the transition equation.

The state vector might be the unobserved state of the business cycle (boom or depression), it might be the unmeasured quality of an investment project or a worker, etc.

Γ is an m x m matrix, c is an m x 1 vector, R is an m x g martrix, and η is an g x 1 vector.

We assume that the initial state vector has a mean of a_0 and a covariance matrix P_0.

,

We also assume that the disturbance terms in the two equations are uncorrelated with each for all leads and lags and are uncorrelated with the initial state, a_0.

The matrices Z, d, H, Γ, c, R, Q are called the **system matrices**, and unless otherwise stated are non-stochastic. (Note that the matrix R is somewhat arbitrary, since we could always re-normalized the covariance matrix of η.)

If these matrices do not change over time but rather are fixed, then the system is called **time invariant or time homogeneous**.

Note that the transition equation in a time invariant model is a first order VAR. Let's look at some examples to see how we can cast models in state space form:

Example 1 AR(1) model with noise

$$y_t = \mu_t + \varepsilon_t$$
$$\mu_t = \phi\mu_{t-1} + \eta_t$$

This is a time invariant state space model, with the state being μ. Note that Z_t is 1, and R_t is one, and d_t and c_t are zero.

Example 2: AR(2)

$$y_t = \begin{bmatrix} 1 & 0 \end{bmatrix}\alpha_t$$

$$\alpha_t = \begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix}\alpha_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}\varepsilon_t$$

Example 3: The MA(1) model

$$y_t = \begin{bmatrix} 1 & 0 \end{bmatrix}\alpha_t$$

$$\alpha_t = \begin{bmatrix} \alpha_{1t} \\ \alpha_{2t} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\alpha_{t-1} + \begin{bmatrix} 1 \\ \theta \end{bmatrix}\varepsilon_t$$

Note that this follows, as $\alpha_{2t} = \theta\varepsilon_t$, and $\alpha_{1t} = \alpha_{2t-1} + \varepsilon_t = \varepsilon_t + \theta\varepsilon_{t-1}$.

The Kalman Filter

The Kalman Filter can now be used to estimate the parameters of the state space model, as well as to filter and smooth.

Define a_t as the optimal estimate of α_t. Define P_t as the m x m covariance matrix of the forecast errors:

$$P_t = E(\alpha_t - a_t)(\alpha_t - a_t)'$$

Now, note that the optimal estimator of α_t is given by:

$$a_{t|t-1} = \Gamma_t a_{t-1} + c_t$$

and we have:

$$P_{t|t-1} = \Gamma_t P_{t-1} \Gamma_t' + R_t Q_t R_t'$$

Note that the corresponding estimator of y is given by:

$$\hat{y}_{t|t-1} = Z_t a_{t|t-1} + d_t$$

The innovation vector is:

$$v_t = Z_t(\alpha_t - a_{t|t-1}) + \varepsilon_t$$

and the mean square error of the innovation vector is:

$$F_t = Z_t P_{t|t-1} Z_t' + H_t$$

Once new observations become available, we have the **updating equations are:**

$$a_t = a_{t|t-1} + P_{t|t-1} Z_t' F_t^{-1} (y_t - Z_t a_{t|t-1} - d_t)$$

$$P_t = P_{t|t-1} - P_{t|t-1} Z_t' F_t^{-1} Z_t P_{t|t-1}$$

Note that the prediction error plays a key role. In particular, the bigger the prediction error, the bigger the change made to the estimator of the state.

Prediction

Prediction just follows from the previous equations. In particular, the optimal estimator of the state vector at date t+l, conditional on date t information, is given by:

$$a_{t+l|t} = \Gamma_{t+l} a_{t+l-1} + c_{t+l}$$

and the associated MSE matrix is obtained from:

$$P_{t+l|t} = \Gamma_{t+l} P_{t+l-1|t} \Gamma_{t+l}' + R_{t+l} Q_{t+l} R_{t+l}'$$

The predictor of y_{t+l} is:

$$\hat{y}_{t+l|t} = Z_{t+l} a_{t+l|t} + d_{t+l}$$

and the prediction MSE is:

$$Z_{t+l} P_{t+l|t} Z_{t+l}' + H_{t+l}$$

Example: AR(1)

Recall the AR(1) with noise model. The forecast and MSE is given by:

$$\hat{y}_{t+l} = \phi^l a_t$$

$$MSE = \phi^{2l} P_t + (1 + \phi^2 + \ldots + \phi^{2(l-1)}) \sigma_n^2 + \sigma_\varepsilon^2$$

How To Do This:

Given your problem, cast it in state space form, being careful to define the state vector. measurement equation, transition equation, and the system matrices. Then it is just a matter of substituting, using the formulae.

Initializing the Kalman Filter

The first step is to provide initial conditions to get the Kalman Filter rolling. If the process is stationary, then the initial conditions for the filter are given by the unconditional mean and variance. Generally, the mean is given by:

$$a_0 = (I - \Gamma)^{-1} c$$

and the covariance matrix is given by:

$$vec(P_0) = [I - \Gamma(kron)\Gamma]^{-1} vec(RQR')$$

Note for the AR(1) with noise,we would have:

We would have:

$$a_0 = 0$$

and

$$P_0 = \sigma_\eta^2/(1 - \phi^2)$$

What if the model is nonstationary? In this case we need to estimate the initial conditions. There are two approaches. The first is to assume that α_0 is fixed, which means that P_0 is 0. We then estimate α_0 as a parameter. The second approach is to assume that α_0 is random and has a diffuse distribution, in which $P_0 = \kappa I$, with $\kappa \to \infty$. **This just means that we don't know anything about the initial state, and the starting values are constructed from the initial observations.**

Example: AR(1) with noise:

$$y_t = \mu_t + \varepsilon_t$$
$$\mu_t = \phi \mu_{t-1} + \eta_t$$

Suppose that $\phi = 1$. Then, the Kalman Filter implies:

$$a_1 = a_0 + \frac{P_0 + \sigma_\eta^2}{P_0 + \sigma_\eta^2 + \sigma_\varepsilon^2}(y_1 - a_0)$$

The MSE is given by:

$$P_1 = P_0 + \sigma_\eta^2 - \frac{(P_0 + \sigma_\eta^2)^2}{P_0 + \sigma_\eta^2 + \sigma_\varepsilon^2}$$

Note that in the limit as $P_0 \to \infty$, we have $a_1 = y_1$ and $P_1 = \sigma_\eta^2$

Derivation of the Kalman Filter

In a Gaussian state space model, the disturbances and the initial state are distributed normally. Given these assumption, we can derive the Kalman filter and construct the likelihood from the prediction errors.

The initial state is normally distributed at date 0. At date 1, the vector is given by:

$$\alpha_1 = \Gamma_1 \alpha_0 + c_1 + R_1 \eta_1$$

Note that the conditional mean is given by:

$$a_{1|0} = \Gamma_1 a_0 + c_1$$

and the covariance matrix is:

$$P_{1|0} = \Gamma_1 P_0 \Gamma_1' + R_1 Q_1 R_1'$$

To obtain the distribution of α_1 conditional on y_1, we have:

$$\alpha_1 = a_{1|0} + (\alpha_1 - a_{1|0})$$
$$y_1 = Z a_{1|0} + d_1 + Z_1(\alpha_1 - a_{1|0}) + \varepsilon_t$$

Note that the mean and covariance matrix of the vector (α_1, y_1) is given by:

$$\begin{bmatrix} a_{1|0} \\ Z_1 a_{1|0} + d_1 \end{bmatrix}, \begin{bmatrix} P_{1|0} & P_{1|0} Z_1' \\ Z_1 P_{1|0} & Z_1 P_{1|0} Z_1' + H_1 \end{bmatrix}$$

The multivariate normal distribution yields:

$$a_1 = a_{1|0} + P_{1|0} Z_1' F_1^{-1} (y_1 - Z_1 a_{1|0} - d_1)$$

and covariance matrix

$$P_1 = P_{1|0} - P_{1|0} Z_1' F^{-1} Z_1 P_{1|0}$$

where we have:

$$F = Z_1 P_{1|0} Z_1' + H_1$$

Repeating these steps for t = 2....T yields the Kalman filter.

We are now in a position to discuss ML estimation of the parameters of the system.

ML Estimation of State Space Models

Estimation of these models is fairly straightforward. First, write the measurement equation as:

$$y_t = Z_t a_{t|t-1} + Z_t(\alpha_t - a_{t|t-1}) + d_t + \varepsilon_t$$

Note that the conditional distribution of y is normal with:

$$\tilde{y}_{t|t-1} = Z_t a_{t|t-1} + d_t$$

and with covariance matrix given by:

$$F_t = Z_t P_{t|t-1} Z_t' + H_t$$

For a Gaussian model, the likelihood is (omitting constants) is:

$$l = -1/2 \sum \log \mid F_t \mid -1/2 \sum v_t' F_t^{-1} v_t$$

where v is the vector of prediction errors.

A univariate model can often be re-parameterized so that the variance of one of the disturbance terms is a scale factor, which we will call σ_*^2. Using lower case letters for the univariate case, with this practice, the measurement equation becomes:

$$y_t = z_t' \alpha_t + d_t + \varepsilon_t, Var(\varepsilon_t) = \sigma_*^2 h_t$$

The transition equation is unchanged, except the covariance matrix of η_t is redefined as $\sigma_*^2 Q$. If the initial covariance matrix is also specified up to the scale factor - that is, $\text{Var}(\alpha_0) = \sigma_*^2 P_0$ then the Kalman filter can be run independently of the scale factor. In this case, the variance of the prediction errors becomes:

$$var(v_t) = \sigma_*^2 f_t$$

Thus, the log likelihood becomes:

$$l = -T/2 \log(\sigma_*^2) - 1/2 \sum \log(f_t) - \frac{1}{2\sigma_*^2} \sum \frac{v_t^2}{f_t}$$

Now, differentiating with respect to σ_*^2 and setting to 0 yields:

$$\tilde{\sigma}_*^2 = (1/T) \sum \frac{v_t^2}{f_t}$$

Substituting, this lets us obtain the conentrated likelihood (again ignoring constants):

$$l = -1/2 \sum \log(f_t) - \frac{T}{2} \log(\sigma_*^2)$$

where σ_*^2 is a function of the parameters to be chosen.

Example 1: AR(1) with noise

$$y_t = \mu_t + \varepsilon_t$$
$$\mu_t = \phi\mu_{t-1} + \eta_t$$

Let $\sigma_*^2 = \sigma_\varepsilon^2$, and $\sigma_\eta^2 = q\sigma_\varepsilon^2$

Since we are concentrating out σ_ε^2, we have:

$$P_0 = q/(1-\phi^2)$$

Using the Kalman Filter for ARMA Models

The Kalman filter can be used to construct the exact likelihood for ARMA models. Let m =max(p,q+1). Then the ARMA(p,q) model can be expressed as

$$\alpha_t = \begin{bmatrix} \phi_1 & \\ \phi_2 & \\ \vdots & I_{m-1} \\ \phi_p & 0' \end{bmatrix} \alpha_{t-1} + \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{m-1} \end{bmatrix} \varepsilon_t$$

This may be regarded as a transition equation in which Γ_t and R_t are constant, and $Q_t = \sigma^2$.

Note that the measurment equation becomes:

$$y_t = z_t'\alpha_t$$

Varying Parameter Regression

Usually we consider fixed parameter models of the form:

$$y_t = x_t'\beta + \varepsilon_t$$

We now discuss how to estimate the parameters of models in which the parameters varying over time.

We first discuss recursive least squares. This allows us to track the stability of coefficients over time.

Suppose that an estimator of β has been constructed using the first t-1 observations. The next observation may be used to construct a new estimator without inverting a cross-product matrix. This is done by recursively updating via the Kalman filter. Let X_t be the vector of all the x observations, let x be a single observation of x, and let b_t be the estimator of β. Then:

$$b_t = b_{t-1} + (X_{t-1}'X_{t-1})^{-1}x_t(y_t - x_t'b_{t-1})/f_t$$

and

$$(X_t'X_t)^{-1} = (X_{t-1}'X_{t-1})^{-1} - (X_{t-1}'X_{t-1})^{-1}x_tx_t'(X_{t-1}'X_{t-1})^{-1}/f_t$$

where we have:

$$f_t = 1 + x_t'(X_{t-1}'X_{t-1})^{-1}x_t$$

This regression model can be cast in state space form with the system vectors given by:

$$z_t = x_t, h_t = 1, \alpha_t = \beta, \alpha_t = \alpha_{t-1}, \Gamma = I, Q = 0$$

Thus, the prediction equations are quite easy:

$$a_{t|t-1} = a_{t-1}, P_{t|t-1} = P_{t-1}, P_t = (X_t'X_t)^{-1}$$

We need to initialize the Kalman filter, but this can be done with a diffuse prior, which

means that the first k observations are used to construct starting values for the k coefficients:

$$b_k = (X_k' X_k)^{-1} X_k' y_k$$

Given these starting values, then all other values can be computed directly without any further matrix inversions. Note that the final estimator - b_T will be identical to the least squares estimator run over the entire sample.

The prediction errors are given by:

$$v_t = y_t - x_t' b_{t-1}$$

By construction, these errors have zero mean, variance $\sigma^2 f_t$. The standardized residuals are given as:

$$\tilde{v}_t = \frac{v_t}{\sqrt{f_t}}$$

These are known as recursive residuals.

Note that we also have:

$$SSE_t = SSE_{t-1} + v_t^2$$

$$SSE_t = (y_t - X_t b_t)'(y_t - X_t b_t)$$

Random Walk Parameters

Suppose we have a model with random walk parameters, and k regressors:

$$y_t = x_t' \beta_t + \varepsilon_t$$
$$\beta_t = \beta_{t-1} + \eta_t, \eta \sim N(0, \sigma^2 Q)$$

Note that Q tells us how much the parameter vector can vary. If Q = 0, then we have the standard linear regression model. If Q is positive definite, then the parameters will vary. Thus, this can be seen as an extension of the recursive least squares procedure.

Again, we use a diffuse prior so that we just use the first k observations to construct starting values. This can be done explicitly by expressing the first k-1 coefficient vectors in

terms of the kth vector:

$$\beta_k = \beta_{k-1} + \eta_t = \beta_{k-2} + \eta_t + \eta_{t-1}$$

Thus, the first k equations may be written as:

$$y_t = x_t'\beta_k + \zeta_t, t = 1, \ldots, k$$
$$\zeta_k = \varepsilon_k, \zeta_t = \varepsilon_t - x_t'(\eta_{t+1} + \ldots + \eta_t), t = 1, \ldots, k-1$$

The covariance matrix of the disturbance vector ζ_k is $\sigma^2 V$, where the ijth element of V is:

$$v_{ij} = \delta_{ij} + [\min(k-i, k-j)]x_i' Q x_j$$

where δ_{ij} is one for i=j and zero otherwise.

This yields:

$$b_k = (X_k' X_k)^{-1} X_k' y_k$$
$$P_k = (X_k V_k^{-1} X_k')^{-1}$$

where the covariance matrix is $\sigma^2 P_k$.

We can then run the Kalman filter using the above expressions for starting values, and then using the log-likelihood we developed previously, for the summation running from k+1 onwards.

GMM Estimation of Dynamic Models

Generalized Method of Moments Estimation has become standard in the last 20 years since Lars Hansen's 1982 *Econometrica* paper.

The original idea was to choose parameters such that population moments are equated to sample moments. This idea was known as "Method of Moments." Hansen generalized this such that it may not be possible to choose enough parameters such that all the moments of interest were match to their sample counterparts.

Hansen's generalization was to define a deviation vector, which we denote as g, whose elements are the difference between population moments and sample moments. He then defined a criterion function, using that deviation vector, and the idea is to optimize that criterion function:

$$g' S' g$$

where S is a weighting matrix that weights the individual elements of the deviation vector.

The specific formulation is as follows:

w is an (h x 1) vector of random variables that are stationary and ergodic, θ is an (a x 1) parameter vector, and $h(\theta, w)$ is a vector-valued function that is random Let θ_0 be the true parameter vector, which satisfies the following:

$$E(h(\theta_0, w)) \ = 0$$

The rows of the h function are called the orthogonality conditions. Let Y be a (th x 1) vector of all the variables in w, and let g be the sampe average of h:

$$g(\theta; Y) \ =) 1/T) \Sigma_t h(.)$$

What GMM does is to choose θ so that the sample moment comes as close as possible to the analagous population moment.

$$\hat{\theta} \ = \ \arg\min\{g_t' S_t g_t\}$$

where S is a sequence of positive definite matrices.

Note that if the number of orthogonality conditions is equal to the number of parameters, we should be able to set the criterion function to 0. If there are more moment conditions than parameters, then we will typically not be able to set this to 0. Instead, we may have to trade off improvement on one moment condition at the expense of another. The extent to which this happens is governed through the weighting matrix, S.

One of the appealing aspects of GMM estimation is that it does not require you to fully specify an economic model. Instead, these orthogonality conditions can just be a subset of a model. We will see how this works later.

The Optimal Weighting Matrix

It turns out that the optimal weighting matrix for GMM is the inverse of the asymptotic variance matrix of the moment conditions: (see Hansen's 1982 paper for a derivation). Thus we can obtain an estimate of this asymptotic variance matrix as:

$$S \ = \ (1/T) \Sigma h(\theta_0) h(\theta_0)'$$

provided that h vector is serially uncorrelated. Note one complication, which is that we need to know θ_0 if we are to calculate the weighting matrix, but of course, we need to know the weighting matrix if we are to calculate θ. To get around this problem, what we do is start out with an initial weighting matrix (typically the identity matrix), and estimate the parameters off of that initial criterion function. Then we can form an estimate of the asymptotic variance matrix, and recalculate the parameter vector using this new estimate of the weighting matrix. We can iterate until the weighting matrix is roughly unchanged.

If the vector process is serially correlated, then we can follow the Newey-West procedure:

$$S_T \ = \ \Gamma_{0,T} + (1/T) \Sigma_{t=v+1} \{1 - (v/(q+1))\} (\Gamma_{v,\,T} + \Gamma_{v,T}')$$

where Γ_j is the autocovariance matrix at lag j.

Asymptotic Distribution of GMM Estimators

See Hamilton, page 414-415. He shows that the GMM estimator is asymptotically normal, distributed:

$$\sqrt{T}(\theta_T - \theta_0) \rightarrow N(0, V)$$
$$V = \{DS^{-1}D'\}^{-1}$$
$$D = p\lim\{\partial g/\partial\theta'\}$$

Testing Overidentifying Restrictions

When the number of orthogonality conditions exceeds the number of parameters, then the model is overidentified and we can test the overidentifying restrictions. Hansen's test that all the sample moments are zero is given by the following:

$$\sqrt{T}g(\theta_0, Y_T)'S^{-1}\sqrt{T}g(\theta_0, Y_T) \rightarrow \chi^2(r)$$

Note that in this expression, we are evaluating the moment conditions at the "true" parameter vector. What if we replace the true parameter vector with the estimated parameter vector?

This is a little tricky. What we need to do is note that when we have the (a x 1) parameter vector θ, and when we have the (r x 1) moment condition vector, then the test statistic for r > a becomes:

$$\sqrt{T}g(\theta_T, Y_T)'\hat{S}^{-1}\sqrt{T}g(\theta_T, Y_T) \rightarrow \chi^2(r-a)$$

Note that this is easy to calculate - it is the sample size multiplied by the optimized value of the objective function.

Examples of GMM Estimation

Example 1: OLS estimation

The standard linear regression model is:

$$y_t = x_t'\beta + u_t$$

Note that the key assumption in OLS is that the right-hand side variables are uncorrelated with the shock. This implies: condition:

$$E(x_t u_t) = 0$$

This implies:

$$E(x_t(y_t - x_t'\beta)) = 0$$

which delivers the moment condition:

$$h = x_t(y_t - x_t'\beta)$$

In the standard regression model, the number of moment conditions is equal to the number of

regressors. This means we can zero out all the moment conditions:

$$0 = g(.) = (1/T)\sum x_t(y_t - x_t'\beta)$$

which implies the usual OLS estimator:

$$\hat{\beta} = (\sum x_t x_t')^{-1} \sum x_t' y_t$$

Calculating the variance of $\hat{\beta}$, we have:

$$D' = \partial g/\partial \theta' \mid_{\theta=\theta_T} = (1/T)\sum \frac{\partial x_t(y_t - x_t'\beta)}{\partial \beta'} \mid_{\beta=\beta_T}$$
$$= -(1/T)\sum x_t' x_t$$

Calculating S under the assumption of i.i.d shocks, we get:

$$\lim(1/T)\sum u_t u_t' x_t x_t' = \sigma^2 E(x_t x_t')$$

It can be shown that this yields the usual variance matrix for the OLS estimator:

$$\sigma^2(\sum x_t x_t')^{-1}$$

Suppose instead that we had serially correlated and heteroskedastic disturbances. The we would estimate the weighting matrix as:

$$S_T = \Gamma_{0,T} + \sum_{v=1}^{q}\{1 - [v/(q+1)]\}(\Gamma_{v,T} + \Gamma_{v,T}')$$

where we have:

$$\Gamma_{v,T} = (1/T)\sum u_t u_{t-v} x_t x_{t-v}'$$

Example 2: Instrumental Variables

Consider the model:

$$y_t = z_t'\beta + u_t$$

where z is a (k x 1) vector. Suppose that z is correlated with the disturbance term u. In this case, OLS estimation will yield biased and inconsistent estimates of the coefficient vector. Suppose that there are some variables x such that z is correlated with x, but x is uncorrelated with u:

$$E(xu) = 0$$

The orthogonality conditions are:

$$E[x_t(y_t - z_t'\beta_0)] = 0$$

This can be written as a GMM problem, in which w = (y,z,x), $\theta = \beta$, and a = k, and:

$$h = x_t(y_t - z_t'\beta)$$

Suppose that a = k = r. Then we have an exactly identified model:

$$0 = (1/T)\sum x_t(y_t - z_t'\hat{\beta})$$

$$\hat{\beta} = \sum(x_t z_t')^{-1}\sum x_t' y_t$$

Note that :

$$D = (1/T)\sum \frac{\partial x_t(y_t - z_t'\beta)}{\partial \beta'} \mid_{\beta=\hat{\beta}} = -(1/T)\sum x_t z_t'$$

We therfore have:

$$V = \{[(1/T)\sum z_t x_t']\hat{S}^{-1}[(1/T)\sum x_t z_t']\}^{-1}$$

If the disturbance term is i.i.d., then we have:

$$S = \sigma^2(1/T)\sum x_t x_t'$$

Substituting, we get:

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = \sigma^2[\sum x_t z_t']^{-1}[\sum x_t x_t'][\sum z_t x_t']^{-1}$$

Example 3: Dynamic Rational Expectations Models

Rational expectations models are very popular to estimate using GMM, because they naturally imply sets of orthogonality conditions. In particular, they imply that forecast errors and variables available at the time the forecast was made should be uncorrelated. For example, a firm's error in forecasting sales should be uncorrelated with the firm's sales at the time the forecast was made.

A benefit of this approach is that the parameters of nonlinear, as well as linear rational expectations models can be estimated. Moreover, the estimation does not require that the full equilibrium of the model be solved, which would be the case if we wanted to estimate the model using Full Information Maximum Likelihood (FIML). Instead, this is a "limited information" estimation strategy that is easy to implement.

A well known case of GMM estimation of RE models is in asset pricing. With rational expectations, prices will reflect all current information. **(See Hamilton, Hayashi, and Lars Hansen and Ken Singleton, Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models, Econometrica, 1982, pp. 1269-1286, available on www.jstor.org)**

Consider the following representative agent, endowment economy, in which there is a single productive asset (sometimes this is called a "Lucas Tree", because it produces consumption each period) and a single physical, non-storable good. Household can competitively trade shares in the tree. Hereafter, we will consider these shares to be analogous to shares of stock in the actual economy.

Each share yields a per-share "dividend", which is paid in the form of output. We can also introduce other assets into the model, which we will call...The household's problem is given by:

$$\max E_0 \sum \beta^t u(c_t)$$

$$s.t.$$

$$s_t(p_t + d_t) + \sum b_{it}(x_{it} + q_{it}) \geq c_t + s_{t+1}p_t + \sum q_{it}b_{it+1}$$

where d is the dividend (per share), z is the number of shares, p is the real price, and c is consumption. The dividend is drawn from a stationary first-order Markov process. We assume that u(c) is concave and twice continuously differentiable.

Let's define an equilibrium for this economy. This consists of a set of Value functions and policy functions for the household for share choices, choices for other assets, and consumption, given pricing functioins for the shares and the other assets, such that the household choices maximize expected utility, given prices. The equilibrium also requires market clearing - that is, we need pricing functions such that $z_t = z$ for all t, and that $b_{it} = 0$ for all i and all t. By Walras law, if these markets clear, then so does the goods market, and we have $c_t = d_t$.

We can write this model as a dynamic programming problem:

$$v(s,d) = \max\{u(c) + \beta E v(s',d')\}$$
$$s.t.$$
$$s_t(p_t + d_t) + \sum b_{it}(x_{it} + q_{it}) \geq c_t + s_{t+1}p_t + \sum q_{it}b_{it+1}$$

Using the budget constraint to substitute out for consumption, we get the following Euler equation for the Household's optimal choice of shares in the tree:

$$-pu_c + \beta E v_{s'} = 0$$

The envelope condition implies:

$$v_s = u_c(p + d)$$

This implies the stochastic Euler equation for the household's first order condition for shares:

$$p_t = \beta E_t[u_{ct+1}(p_{t+1} + d_{t+1})]/u_{ct}$$

This is a first order (potentially non-linear) difference equation. This can be solved forward to obtain the fundamental solution to the price as:

$$p_t = E_t \sum_{j=t+1}^{\infty} \beta^j u_{cj} d_j / u_{ct}$$

The household's first order condition for the other assets is given by:

$$q_{it} = \beta E_t[u_{ct+1}(x_{it} + q_{it+1})]/u_{ct}$$

This can also be solved forward to get:

$$q_{it} = E_t \sum_{j=t+1}^{\infty} \beta^j u_{cj} x_{ij} / u_{ct}$$

Suppose further that we are interested in estimating the preference parameters of the model. This requires that we take a stand on the utility function. For the time being, let's assume it is given by:

$$u(c_t) = \frac{c_t^{1-\sigma} - 1}{1-\sigma}, \sigma \neq 1$$
$$\ln(c_t), \sigma = 1$$

Thus, we wish to estimate the preference parameters β and σ. Note that β is the household's discount factor, and σ governs the household's risk aversion, and also their intertemporal elasticity of substituion. In particular, the IES is given by $1/(1-\sigma)$ and the coefficient of relative risk aversion is $= \sigma$.

The basic idea is that the equilibrium of this model, as well as other rational expectations models, involves stochastc Euler equations, or stochastic first order conditions that must be satisfied. These Euler equations imply a set of population orthogonality conditions that in general depend nonlinearly on the parameters. This estimation procedure boils down to a non-linear IV estimator.

Note that to solve out for the full equilibrium of the model, we would need to evaluate the expectational nonlinear difference equations. This can be done analytically for specific assumptions about the forcing process (e.g. consumption), but in general needs to be done numerically.

Usually, more orthogonality conditions are available than are parameters, so the model is overidentified, and the overidentifying restrictions can thus be tests. A benefit of this type of estimation is that one typically cannot find a closed form solution for the equilibrium of these models. However, one can almost always write down the stochastic Euler equations that characterize the equilibrium of the model.

Now let's consider how to use GMM to estimate the preference parameters in this model using the investor's first order conditions. First, rewrite the foc for stock share holdings as:

$$1 = \beta E_t[u_{ct+1}(p_{t+1} + d_{t+1})]/(p_t u_{ct})$$

Let's define the gross return on the ith asset between dates t and t+1 as:

$$R_{it+1} = \frac{p_{it+1} + d_{it+1}}{p_{it}}$$

This implies for all dates that:

$$E_t\{1 - \beta[\left(\frac{c_{t+1}^{-\sigma}}{c_t^{-\sigma}}\right)R_{it+1}]\} = 0$$

We will use this population orthogonality condition in the GMM estimation. Recall that we

noted often these models have more orthogonality conditions than parameters. This is true in this model, since the orthogonality condition above holds for *any* asset.

Note the economics behind this result. It implies that in expectation, the Expected product between the intertemporal marginal rate of substitution in consumption and the return on any asset is equated:

$$E_t\{\beta[\left(\frac{c_{t+1}^{-\sigma}}{c_t^{-\sigma}}\right)R_{it+1}]\} = E_t\{\beta[\left(\frac{c_{t+1}^{-\sigma}}{c_t^{-\sigma}}\right)R_{jt+1}]\}$$

for all i,j. There are two ways for assets returns to satisfy this expectational equation. Through differences in mean returns, and/or differences in the covariance between asset returns and the inertemporal marginal rate of substitution.

Note that a "good" asset is one that has high returns, and also that tends to have high payoffs in states when consumption is low. Since the equilibrium pins down the return relationship as above, this means that assets which have very high returns on average will also tend to have relatively low payoffs in low consumption states, and that assets with relatively low returns on average will also tend to have relatively high payoffs in low consumption states. To see this note we can decompose the moment condition into:

$$1 = E_t\left(\beta\left(\frac{c_{t+1}^{-\sigma}}{c_t^{-\sigma}}\right)\right)E_t(R_{it+1}) + cov(\beta\left(\frac{c_{t+1}^{-\sigma}}{c_t^{-\sigma}}\right), R_{it+1})$$

Now, we can define M moment conditions, all of the form described above, indexed by the type of asset. For m different assets, define:

$$h(y_{t+1}, b_0) = \begin{bmatrix} \{1-\beta[\left(\frac{c_{t+1}^{-\sigma}}{c_t^{-\sigma}}\right)R_{1t+1}]\} = 0 \\ \{1-\beta[\left(\frac{c_{t+1}^{-\sigma}}{c_t^{-\sigma}}\right)R_{2t+1}]\} = 0 \\ \vdots \\ \{1-\beta[\left(\frac{c_{t+1}^{-\sigma}}{c_t^{-\sigma}}\right)R_{mt+1}]\} = 0 \end{bmatrix}$$

where y_{t+1} are observable variables at date t+1, and b_0 is a parameter vector. Finally, let $u_{t+1} = h(.)$.

Now define $f(y_{t+1}, z_t, b_0) = h(kron)z_t$. This implies that E(f) = 0.

Now, we can set this up as a GMM problem.

$$g_T(b) = \frac{1}{T}\sum f(.) \quad |_{b=b_0}$$

This should be close to 0 assympotitically. Define J_T as :

$$J_T(b) = g_T(b)' W_T g_T(b)$$

where W is our weighting matrix. Now define:

$$D_0 = E\left[\frac{\partial h}{\partial b}(kron)z \right]$$

and define:

$$S_0 = E(ff')$$

which implies that the weighting matrix is given by:

$$W = S_0^{-1}$$

Note that this is the optimal weighting matrix under the assumption of no serial correlation. Given this weighting matrix, the asymptotic covariance matrix is given by:

$$(D_0' W_0 D_0)^{-1}$$

We obtain consistent estimatates of the D and W matrices as:

$$D_T = \frac{1}{T} \sum \frac{\partial h_t}{\partial b}(kron)z_t$$

$$W_T = \frac{1}{T} \sum ff'$$

Recall that a suboptimal estimator of W needs to be used initially, and then we can iterate, sequentially updating our estimate of W_T.

Empirical Results

Hansen and Singleton use monthly data on nondurables and services consumption per capita, and three sets of stock returns: an equally weighted portfolio of all NYSE stocks, a value-weighted portfolio of the same stocks, and an equally weighted portfolio of chemical stocks, transportation, and retail trade.

The vector of instruments was lagged values of consumption growth and the returns, using lags ranging from 1,2,4 or 6. Note that the more lags you use, the more over-identifying restrictions there are to test.

If you check Hansen and Singleton's tables, you will see that they estimate economically plausible values of the discount factor (a litte less than one, which is reasonable given that it is annual data) and a value of the curvature parameter (around log utility) for the single return models. For the multiple return models, the curvature parmater is close to 0, and there is substantial evidence against the model in that the J-test has a high value. Thus, this leads to rejection of the null hypothesis.

Overview of Estimating and Testing Nonlinear Rational Expectations Models

Estimating nonlinear RATEX models is in principle simple using GMM. The steps are as follows:

(1) Specify the economic model - that is, write down the objective functions and constraints, and any other features of the economic environment that are required for the first order necessary conditions.

(2) Solve for the first order necessary conditions

(3) Write these first order conditions in the form of a population orthogonality condition.

(4) Set up the sample orthogonality conditions

(5) Specify the weighting matrix - note that this depends on whether there is serial correlation in the disturbance term

(6) Form the criterion function, and choose parameter values to minimize the criterion function

(7) Calculate standard errors of the parameters using the weighting matrix and the scores

(8) Evaluate Hansen's J-statistic if there are over-identifying restrictions. High values mean that you will tend to reject the restrictions.

Interpreting Rejections of the Restrictions

Let's think about this rejection of the intertemporal asset pricing model. Literally, it means that we reject the over-identifying restrictions. But there are other questions we are interested in. For example, How bad of a failure is this quantiatively? In other words, while the statistical rejection of the model is significant, should we reject the model economically, or is the model still usefull, despite this rejection? Another question is what caused the statistical failure of the model? Does this failure help us construct models that are less likely to be rejected by the data. These questions are unanswered by the test statistic.

First, let's try to understand the statistical failure of the model. A common interpretation of the failure is that rational expectations/no abritrage fails. This interpretation had led to much of the behavioral finance literature (e.g. Shiller's book "Irrational Exuberance".)

But the failure could also reflect measurement error in the data, or model misspecification. Regarding this latter possibility, we can conceive of many possibilities - failure of the representative agent construct, the wrong class of preferences, etc.

Given these different possible interpretations of rejecting the null, it is useful to step back and try to shed light on what specifically is causing the model to fail. When we reject the null hypothesis, we therefore would like to conduct additional analyses. We will now consider 2 such analyses - Mehra and Prescott (1985) and Hansen and Jaganathan (1991).

Mehra and Prescott chose a very different strategy than Hansen-Singleton. Instead of estimating paramer values off of the Euler equations, they solve for the full equilibrium of the model under an assumption about the forcing process (consumption growth).

In particular, they assume a two-state Markov chain model for consumption growth so they can get a simple expression for:

$$q_{it} = E_t \sum_{j=t+1}^{\infty} \beta^j u_{cj} x_{ij} / u_{ct}$$

They then ask whether the model can plausibly account for the difference between stock and bond returns.

They report the following basic statistcis:

average consumption growth = 2% with standard deviation = 3.5%

average return on gov. bonds = 1% with standard deviation = 5.7%

average return on stocks = 7% with standard deviation = 16.7%

They look at the difference between the two security returns, which means the only parameter that matters is σ.

They consider σ between 0 and 10.

They find a risk premium between stocks and bonds of 0.4%. Note that the true risk premium is 7%, which is 17 times bigger!

Thus, one needs to use very, very high risk aversion parameters to explain the risk premium on stocks - in the neighborhood of 50. This magnitude of the risk premium is much larger than any estimate in the literature.

Note at some level, the Mehra-Prescott paper and the Hansen-Singleton paper are quite similar. Both deal with the same representative agent, endowment economy model, and both are aimed at assessing the model's ability to account for asset prices. The Mehra-Prescott paper has been much more influential, because it more clearly demonstrated the major economic puzzle - that the representative agent power utility model cannot come close to explaining risk premia - than did the Hansen-Singleton paper, which demonstrated that the model was statistically rejected.

An even simpler and more transparent way to evaluate the implications of asset pricing models is in Hansen-Jaganathan, Journal of Political Economy, 1991, pp. 225- 262.

Consider the following implications of the model we have been using, and denote the following object:

$$m_{t+1} = \beta \frac{u_{ct+1}}{u_{ct}}$$

Given this definition, we have:

$$1 = E_t(m_{t+1} R^i_{t+1})$$

This implies

$$0 = E_t(m_{t+1}(R^i_{t+1} - R^j_{t+1}))$$

By the law of iterated expectations, we also have:

$$0 = E(m_{t+1}(R^i_{t+1} - R^j_{t+1}))$$

We can write this as:

$$0 = E_t(m_{t+1}) E_t(R^i_{t+1} - R^j_{t+1}) + cov(m_{t+1}, (R^i_{t+1} - R^j_{t+1}))$$

This implies:

$$0 = E_t(m_{t+1})E_t(R^i_{t+1} - R^j_{t+1}) + corr(m_{t+1}, (R^i_{t+1} - R^j_{t+1}))std(m_{t+1})std(R^i_{t+1} - R^j_{t+1})$$

Rearranging, we get:

$$-\frac{E_t(R^i_{t+1} - R^j_{t+1})}{std(R^i_{t+1} - R^j_{t+1})} = corr(m_{t+1}, (R^i_{t+1} - R^j_{t+1}))\frac{std(m_{t+1})}{E_t(m_{t+1})}$$

Since the correlation ranges between -1 and 1, we can establish a bound on the standard deviation m relative to the expectation of m:

$$\frac{std(m_{t+1})}{E_t(m_{t+1})} \geq \frac{E_t(R^i_{t+1} - R^j_{t+1})}{std(R^i_{t+1} - R^j_{t+1})}$$

The right hand side of the equation boils down to a number - the mean of the excess return of security "i" relative to security "j", divided by the standard deviation of the excess return. We can then trace out the frontier of this relationship, considering alternative values of the mean of m_{t+1}, which then imply different values for the standard deviation of m. The bottom line from Hansen and Jaganathann's paper is that we require a very volatile intertemporal marginal rate of substitution to reconcile the model with the data.

Why do we say this? The right hand side of the inequality is around 0.5, which is the average excess return is in the neighborhood of .07, and the standard deviation is about .14. ratio Now, consider the left-hand side. The mean of the intertemporal marginal rate of substitution is around 1. Given this number, we now consider the standard deviation of the intertemporal marginal rate of substituion. Note that if this is the standard model, then this standard deviation is equal to consumption growth raised to a power, which is equal to the curvature parameter, σ. The standard deviation of consumption growth is about .01 to .02.. This implies a curvature parameter between 25 to 50, which is much larger than the conventional wisdom about this parameter, and coincides with the values implied by Mehra and Prescott's analysis.

The importance of the Hansen-Jaganathann paper is that it clearly demonstrates the main issue associated with trying to reconcile the asset market data with the model, and it does it even more transparently that Mehra-Prescott. This tells us that understanding asset returns requires a theory that delivers a highly volatile intertemporal marginal rate of substitution. Recent research along these lines include Alvarez and Jermann (2000 Econometrica), Lustig (2002, Stanford, unpublished), Campbell and Cochrane (JPE, 2000?)

Hidden State Markov Models

We now discuss models with "changes in regime". That is, there are discrete changes in the parameter values of the stochastic process, and the regime that is governing the process is unobserved. We can think of lots of applications for this type of model. For example, the business cycle might be modelled as a stochastic process with depression and boom states, both of which are unobserved. The central bank may fluctuate between expansionary policy, or restrictive policy states. Investors may have pessimistic or optimistic states. Productivity may have a high state or a low state. Innovative discoveries may be a high discovery state and a low discovery state, etc.

Note that this model represents a particular class of non-linear time series models. It is worthwhile working through Hamilton for this material, as he is the developer of much of this literature. These notes follow Hamilton very closely.

Consider Figure 22.1 in Hamilton (p. 678) which shows the exchange rate for the Mexican Peso vs. the dollar. The simplest regime-change model that would reasonably approximate this time series is one with a change in mean:

$$y_t = \mu_1 + \phi(y_{t-1} - \mu_1) + \varepsilon_t, t \leq t_1$$
$$y_t = \mu_2 + \phi(y_{t-1} - \mu_2) + \varepsilon_t, t > t_2$$

The simplest model for a discrete valued random variable is a *Markov Chain Model.*

Markov Chains

Let s_t be a random variable that can assume integer values. Suppose that the probability that s equals a specific value depends on one lag of s.:

$$\Pr(s_t = j \mid s_{t-1} = i) = p_{ij}$$
$$\sum_j p_{ij} = 1$$

We denote P as the transition matrix:

$$P = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{N1} \\ p_{12} & & & \\ \vdots & & & \\ p1N & & & p_{NN} \end{bmatrix}$$

Representing a Markov Chain with a VAR

Let ξ_t be a N x 1 vector whose jth element =1 if $s_t = j$, and whose jth element is 0

otherwise.

Note that if $s_t = i$, then the jth element of ξ_{t+1} is a random variable that takes on the value of unity with probability p_{ij} and takes the value 0 otherwise. This implies that the random variable has expectation p_{ij}. Thus, the conditional expectaton given $s_t = i$ is given by:

$$E(\xi_{t+1} \mid s_t = i) = \begin{bmatrix} p_{i1} \\ p_{i12} \\ \vdots \\ p_{iN} \end{bmatrix}$$

Note that this implies:

$$E(\xi_{t+1} \mid \xi_t) = P\xi_t$$
$$\xi_{t+1} = \xi_t + v_{t+1}$$

where

$$v_{t+1} = \xi_{t+1} - E(\xi_{t+1} \mid \xi_t, \xi_{t-1}, \ldots)$$

Setting up the Markov chain this way implies that it has the form of a first order VAR, with v_t being a martingale difference sequence (white noise).

Forecasts for a Markov Chain

Note that we can generate forecasts from this process. In particular, note:

$$\xi_{t+m} = v_{t+m} + Pv_{t+m-1} + \ldots P^{m-1}v_{t+1} + P^m\xi_t$$

Thus, our optimal forecast for ξ_{t+m} is given by $P^m\xi_t$. Note that since the jth element of ξ_{t+m} will be one if s_{t+m} = j, and 0 otherwise, the jth element of the N x 1 vector $E_t(\xi_{t+m})$ is just the probability that s_{t+m} takes on the value j, conditional on the state at date t.

Reducible Markov Chains

The transition matrix for a two-state Markov chain is:

$$\begin{bmatrix} p_{11} & 1 - p_{22} \\ 1 - p_{11} & p_{22} \end{bmatrix}$$

Note that if $p_{11} = 1$, then the matrix is upper triangular, and state 1 is called an absorbing state - that is, once in state 1, you stay there forever. With this property, the Markov chain is

reducible. More generally, an N state Markov chain is reducible if the transition matrix can be written as:

$$\begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$$

in which B denotes a (K x K) matrix such that $1 \leq K < N$. (Note that this form assumes you can choose which state to call state 1, which to call state 2, etc.). Note that if P is upper block triangular, then so is P^m. Thus, once you enter that state, then you stay there.

A Markov chain is *irreducible* if the matrix cannot be written as upper block triangular.

Ergodic Markov Chains

Since the probabilites sum to 1, note that every column of P sums to 1. Thus, $P'1 = 1$, where 1 is an N x 1 vector of 1s.. Thus unity is an eigenvalue of P' and that the vecotr of 1s is the associated eigenvector. Since a matrix and its transpose share the same eigenvalues, then unity is an eigenvalue of the transition matrix P for any Markov chain.

Now consider an N-state irreducible Markov chain with transition matrix P. Suppose that one of the eigenvalues is unity and the others are inside the unit circle. Then the markov chain is said to be *ergodic*, and we denote the N x 1 vector of ergodic probabilities as π.

This vector is defined as the eigenvector of P associated with the unit eigenvalue - that is:

$$P\pi = \pi$$

Note that we have normalized the eigenvector so that its elements sum to unity ($1'\pi = 1$). It can be shown that:

$$\lim_{m \to \infty} P^m = \pi' 1$$

This just says the following: the long-run forecast for an ergodic Markov chain is independent of the current state. The long run forecast for ξ_{t+m} as m goes to infinity is governed by the ergodic probabilities, regardless of the current value of ξ. It follows that the vector of ergodic probabilites are also the unconditional probabilities of ξ.

Caculating Ergodic Probabilities

For a N-state ergodic process, the vector of unconditional probabilities is a vector π with the properties that $P\pi = \pi$ and $1'\pi = 1$, where $1'$ is a N-dimensional vector of 1's. We therefore are looking for a vector π such that

$$A\pi = e_{N+1}$$

where e_{N+1} denotes the (N + 1)th column of I_{N+1} and where A is of dimension (N +1) x N, and is given by:

$$A = \begin{bmatrix} I_N - P \\ 1' \end{bmatrix}$$

A solution can be found by premultiplying the above equation for π by $(A'A)^{-1}A'$:

$$\pi = (A'A)^{-1}A'e_{N+1}$$

This implies that π is the (N+1) column of $(A'A)^{-1}A'$.

I.I.D. Mixture Distributions

We now begin constructing some statistical regime-switching models. Let the regime be indexed by an unobserved markov chain. A special case of these processes is the an i.i.d. mixture distribution.

The random variable that indexes the regime is denoted as s, and there are N possible regimes: ($s_t = 1, \ldots, N$). When the process is in regime 1, then the observed random variable y_t is drawn from a $N(\mu_1, \sigma_1^2)$ distribution, and so forth. The density of y conditional on the random variable s taking on the value j is:

$$f(y_t \mid s_t = j; \theta) = \frac{1}{\sqrt{2\pi}\,\sigma_j} \exp\left\{ \frac{-(y_t - \mu_j)^2}{2\sigma_j^2} \right\}$$

Let the unconditional probability that s takes on the value j be given by:

$$P\{s_t = j; \theta\} = \pi_j$$

(Note that the parameter vector θ includes these probabilities).

Recall that the probability of event A given event B is given by:

$$P\{A \text{ and } B\}/P\{B\}$$

and this implies that

$$P\{A \text{ and } B\} = P\{A \mid B\}P\{B\}$$

We therefore have the joint-density function of y and s as:

$$p(y_t, s_t = j; \theta) = \frac{\pi_j}{\sqrt{2\pi}\,\sigma_j} \exp\left\{ \frac{-(y_t - \mu_j)^2}{2\sigma_j^2} \right\}$$

Note that the unconditional density of y can be recovered by summing over all j:

$$f(y_t; \theta) = \sum_j \frac{\pi_j}{\sqrt{2\pi}\,\sigma_j} \exp\left\{ \frac{-(y_t - \mu_j)^2}{2\sigma_j^2} \right\}$$

Since s is a latent variable, this unconditional distribution is the relevant one for analysis. Moreover, if s is i.i.d., then the log likelihood can be calculated as:

$$l = \sum_t \log(f(y_t;\theta)$$

Then, we can solve for θ_{MLE} as maximizing the above expression, subject to the constraint that the probabilities are non-negative and sum to 1.

Making Inferences about the Regime

Recall we have:

$$P\{s_t = j \mid y_t;\theta) = \frac{p(y_t, s_t = j;\theta)}{f(y_t;\theta)} = \frac{\pi_j f(y_t \mid s_j = j;\theta)}{f(y_t;\theta)}$$

Given an estimate of θ, we can use the log-likelihood and the expression for the conditional density to evaluate the probability for each observation of any specific regime. Intuitively, if the data are generated from a mixture of 2 normals, one which has a mean of 0 and variance of 1, and another that has a mean of 10 and a variance of 1, and we observe a value of 11, then we should be able to infer it is much more likely that the regime was the high mean process regime rather than the low mean process regime.

The EM Algorithm

We now describe briefly the EM algorithm. EM stands for "Estimation and Maximization". It turns out that θ_{MLE} yields (where all parameters are estimates)

$$\hat{\mu}_j = \frac{\sum_t y_t \bullet (P\{s_t = j \mid y_t;\theta\})}{\sum_t (P\{s_t = j \mid y_t;\theta\})}$$

$$\hat{\sigma}_j^2 = \frac{\sum_t (y_t - \mu_j)^2 \bullet (P\{s_t = j \mid y_t;\theta\})}{\sum_t (P\{s_t = j \mid y_t;\theta\})}$$

$$\hat{\pi}_j = \frac{1}{T} \sum P\{s_t = j \mid y_t;\theta\}$$

(See Hamilton for a derivation of these MLE estimates of the parameters).

Note that if we were certain of the regime, then the probabilites are either 0 or 1, and we just take sample averages. But when the probabilites are between 0 and 1, then we are taking weighted averages. But note that this is a system of nonlinear equations. We can make progress using the EM algorithm.

The process is:

Start with an initial guess for θ, and calculate $P\{s_t = j \mid y_t;\theta\}$ from the relevant

expression for this probability:

$$P\{s_t = j \mid y_t;\theta) = \frac{\pi_j f(y_t \mid s_j = j;\theta)}{f(y_t;\theta)}$$

We can then evaluate the 3 expressions for the parameters, which gives us a new estimate of θ. We then iterate until the θ estimates converge.

Regime Switching with Dependent Processes

Now that we have done the i.i.d. case, we turn to the more relevant case of serially correlated variables. Suppose that the process is:

$$y_t = c_{st} + \phi_s y_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim N(0, \sigma^2)$$

We now model s as an N-state Markov chain with s independent of ε.

Let y_t be an (n x 1) vector of observed endogenous variables and z is a (k x 1) vector of observed exogenous variables. Let Y include the history of all observed variables.

The conditional density of y is:

$$f(y_t \mid s_t = j, z_t, Y_t; \alpha)$$

Note that α is a vector of parameters. If there are N different regimes, then there are N different densities. We collect those conditional densities in an (N x 1) vector which we denote as η_t.

Example - suppose y is a scalar, there is only a constant for the exogenous variables, and the unknown parameters are the c's and the $\phi's$, and σ^2. Suppose there are 2 regimes. Then the 2 densities are:

$$\eta_t = \begin{bmatrix} f(y_t \mid s_t = 1, y_{t-1}; \alpha) \\ f(y_t \mid s_t = 2, y_{t-1}; \alpha) \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{\frac{-(y_t - c_1 - \phi_1 y_{t-1})^2}{2\sigma^2}\right\} \\ \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{\frac{-(y_t - c_2 - \phi_2 y_{t-1})^2}{2\sigma^2}\right\} \end{bmatrix}$$

Note that we have assumed that the conditional density depends only on the current regime, and not past regimes:

$$f(y_t \mid s_t = j, z_t, Y_t; \alpha) = f(y_t \mid s_t = j, s_{t-1} = i, s_{t-2} = k, \ldots, z_t, Y_t; \alpha)$$

This is not a big deal, since we can stack a model with regime dependence into this particular form (see Hamilton, page 691). The basic idea is to augment the state vector and define states that include information at date t and at date t-1. In particular for a 2 state model, we expand it to a four state model such that state 1 is if the current and past state were both state 1, state 2 is if the current state is 2 and the past state is 1, and so forth.

To see this more clearly, we have:

$$s_t = 1 \text{ if } s_t^* = 1 \text{ and } s_{t-1}^* = 1$$
$$s_t = 2 \text{ if } s_t^* = 2 \text{ and } s_{t-1}^* = 1$$
$$s_t = 3 \text{ if } s_t^* = 1 \text{ and } s_{t-1}^* = 2$$
$$s_t = 4 \text{ if } s_t^* = 2 \text{ and } s_{t-1}^* = 2$$

If we let p_{jt}^* denote $P\{s_t^* = j \mid s_{t-1}^* = i\}$, the we have a transition matrix given by:

$$\begin{bmatrix} p_{11}^* & p_{11}^* & \\ p_{12}^* & p_{12}^* & \\ & p_{21}^* & p_{21}^* \\ & p_{22}^* & p_{22}^* \end{bmatrix}$$

The we can write the 4 densities as:

$$f(y_t \mid y_{t-1}, s_t = 1; \alpha) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\{\frac{-[(y_t - \mu_1) - \phi(y_{t-1} - \mu_1)]^2}{2\sigma^2}$$
$$f(y_t \mid y_{t-1}, s_t = 2; \alpha) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\{\frac{-[(y_t - \mu_2) - \phi(y_{t-1} - \mu_1)]^2}{2\sigma^2}$$
$$f(y_t \mid y_{t-1}, s_t = 3; \alpha) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\{\frac{-[(y_t - \mu_1) - \phi(y_{t-1} - \mu_2)]^2}{2\sigma^2}$$
$$f(y_t \mid y_{t-1}, s_t = 4; \alpha) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\{\frac{-[(y_t - \mu_2) - \phi(y_{t-1} - \mu_2)]^2}{2\sigma^2}$$

Inference and Evaluating the Likelihood

The parameters of interest are α and the regime switching probabilities. Collect these parameters into a vector, which we denote as θ. We wish to estimate these parameters conditional on a history of data, Y_t. Now, suppose that we knew θ. Note that even if we did know the parameter vector, we don't know which regime was operative - we can only make probabilitic statements about that regime. In the i.i.d. case that we examined above, the evaluation of s depended only on y_t. Generally, it will involve all observations.

Denote $P\{s_t = j \mid Y_t; \theta\}$ as the probability based on data through time t and conditional on knowing θ. Collect these probabilites for all N states and stack them in an (N x 1) vector, which we denote as $\hat{\xi}_{t|t}$. We can also make a forecast of this vector for date t+1, conditional on date t information. Denote this as $\hat{\xi}_{t+1|t}$.

Then the optimal inference and forecast for each date t in the sample can be found by

iterating on these equations:

$$\hat{\xi}_{t|t} = \frac{(\hat{\xi}_{t|t-1} \bullet \eta_t)}{1'(\hat{\xi}_{t|t-1} \bullet \eta_t)}$$
$$\hat{\xi}_{t+1|t} = P\hat{\xi}_{t|t}$$

where the "dot" in the equations is element-by-element multiplication, and recall that η_t is the conditional density vector, and that P is the transition matrix. Given a starting value for $\hat{\xi}_{1|0}$ and knowledge of θ, one can iterate on these two equations. We can also evalute the log likelihood from this iteration, where the log-likelihood is:

$$l = \Sigma_t \log f(y_t \mid z_t, Y_{t-1}; \theta)$$
$$f(y_t \mid z_t, Y_{t-1}; \theta) = 1'((\hat{\xi}_{t|t-1} . \eta_t)$$

This may seem a bit mysterious. Let's see what this all means.

Deriving the likelihood

Note that z_t is exogenous, which means that it contains no information about s other than what is contained in the history vector Y_t. The jth element of $\hat{\xi}_{t|t-1}$ is:

$$P\{s_t = j \mid z_t, Y_{t-1}; \theta\}$$

The jth element of η_t is:

$$f(y_t \mid s_t = j, z_t, Y_{t-1}; \theta)$$

Then the product of these two objects is the joint density:

$$P\{s_t = j \mid z_t, Y_{t-1}; \theta\} \bullet f(y_t \mid s_t = j, z_t, Y_{t-1}; \theta) =$$
$$p(y_t, s_t = j \mid z_t, Y_{t-1}; \theta)$$

Now, the density of the observed vector y_t, conditioned on past observations is the sum of the N magnitudes in the above equation for j = 1,2,...,N. This sum can be written as a vector:

$$f(y_t \mid z_t, Y_{t-1}; \theta) = 1'((\hat{\xi}_{t|t-1} \bullet \eta_t)$$

If the joint density is divided by the density of y_t, then we obtain the conditional distribution of s_t :

$$\frac{p(y_t, s_t = j \mid z_t, Y_{t-1}; \theta)}{f(y_t \mid z_t, Y_{t-1}; \theta)} = P\{s_t = j \mid z_t, Y_{t-1}; \theta\} = P\{s_t = j \mid Y_t; \theta\}$$

It follows immediately that:

$$P\{s_t = j \mid Y_t;\theta\} = \frac{p(y_t, s_t = j \mid z_t, Y_{t-1};\theta)}{1'((\hat{\xi}_{t|t-1} \bullet \eta_t)}$$

Note that the numerator of the above equation is the jth element of $\hat{\xi}_{t|t-1} \bullet \eta_t$, while the left hand side of the equation is the jth element of $\hat{\xi}_{t|t}$. If we collect the equations for j = 1,2,...,N into a vector, we obtain:

$$\hat{\xi}_{t|t} = \frac{\hat{\xi}_{t|t-1} \bullet \eta_t}{1'((\hat{\xi}_{t|t-1} \bullet \eta_t)}$$

To understand the second equation from the previous sub-section, recall we had

$$\hat{\xi}_{t+1|t} = P\hat{\xi}_{t|t}$$

Taking expectations, we get

$$E(\xi_{t+1} \mid Y_t) = PE(\xi_t \mid Y_t) + E(v_{t+1} \mid Y_t)$$

Since v_t is white noise, we thus get

$$\hat{\xi}_{t+1|t} = P\hat{\xi}_{t|t}$$

Starting out the algorithm

Given a value for $\xi_{1|0}$ in conjunction with our equations for inference and forecasting to calculate $\hat{\xi}_{t|t}$ for any t. How do we choose this initial value? One approach is to just use the vector of unconditional probabilities (recall we denoted these as π). An alternative is to set $\xi_{1|0} = \rho$, where ρ is a fixed (N x 1) vector of nonnegative numbers that sum to unity. Alternatively, we could use MLE, subject to the constraint that $1'\rho = 1$ and the elements of ρ are non-negative.

Forecasts and Smoothed Inferences for the Regime

Suppose we would like to forecast the regime of the state in the future? Alternatively, suppose that we would like to know the regime of the state in the past? Both of these can be done.

Now, let $\hat{\xi}_{t|\tau}$ represent the (N x 1) vector whose jth element is $P\{s_t = j \mid Y_\tau;\theta\}$ Note that for $t > \tau$, this represents a forecast about the regime for a future period. For $t < \tau$, this represents the smoothed inference about the regime the process was in at date t based on data obtained through some later date τ.

The optimal m-period forecast of ξ_{t+m} is found by taking expectations of both sides conditional on date-t information:

$$\hat{\xi}_{t+m|t} = P^m \hat{\xi}_{t|t}$$

In general, smoothed inferences can be calculated as

$$\hat{\xi}_{t|T} = \hat{\xi}_{t|t} \bullet \{P' \bullet [\hat{\xi}_{t+1|T}(\div)\hat{\xi}_{t+1|t}]\}$$

where $(\div)$ denotes element-by-element division. The smoothed probabilities are found by iterating on the above equation backwards.

Forecasting the Observed Variables

We are also interested in forecasting the observed variables, in addition to the regime that state is in.

From the conditional density:

$$f(y_t \mid s_t = j, x_t, Y_{t-1}; \alpha)$$

we can forecast y_{t+1} conditional onknowing Y_t, x_{t+1}, and s_{t+1}.

Consider the AR(1) specification:

$$y_{t+1} = c_{st+1} + \phi_{st+1} y_{t-1} + \varepsilon_{t+1}$$

Then our forecast is:

$$E(y_{t+1} \mid s_{t+1} = j, Y_t; \theta) = c_j + \phi_j y_t$$

Note that there are N different conditional forecasts associated with the N possible values for the state. The relationship between the conditional and unconditional forecasts is given by:

$$E(y_{t+1} \mid z_{t+1}, Y_t; \theta) = \int y_{t+1} \bullet f(y_{t+1} \mid z_{t+1}, Y_t; \theta) dy_{t+1} =$$

$$\int y_{t+1} \{\sum_{j=1} p(y_{t+1}, s_{t+1} = j \mid z_{t+1}, Y_t; \theta)\} dy_{t+1} =$$

$$\int y_{t+1} \{\sum_j [f(y_{t+1} \mid s_{t+1} = j, z_{t+1}, Y_t; \theta) P\{s_{t+1} = j \mid z_{t+1}, Y_t; \theta\}]\} dy_{t+1} =$$

$$\sum_j P\{s_{t+1} = j \mid Y_t; \theta\} \int y_{t+1} \bullet f(y_{t+1} \mid s_{t+1} = j, z_{t+1}, Y_t, \theta) dy_{t+1} =$$

$$\sum_j P\{s_{t+1} = j \mid Y_t; \theta\} E(y_{t+1} \mid s_{t+1} = j, z_{t+1} Y_t, \theta)$$

Example:

Suppose we have N different forecasts for our N different states. Collect those N different forecasts into a (1 x N) vector, and denote that vector as h'. Then the expected value of y is:

$$E(y_{t+1} \mid Y_t; \theta) = h' \hat{\xi}_{t+1 \mid t}$$

Maximum Likelihood Estimation

In the previous discussion about iterating on:

$$\hat{\xi}_{t \mid t} = \frac{(\hat{\xi}_{t \mid t-1} \bullet \eta_t)}{1'(\hat{\xi}_{t \mid t-1} \bullet \eta_t)}$$

$$\hat{\xi}_{t+1 \mid t} = P \hat{\xi}_{t \mid t},$$

We took the parameter vector to be fixed and known. Once an iteration is completed through the sample for a given θ, then the value of the log likelihood can be solved out from:

$$l = \Sigma_t \log f(y_t \mid z_t, Y_{t-1}; \theta)$$

The value of θ that maximizes the log likelihood can be computed numerically using the same methods as we desribed above.

In the case that the only restrictions are that the transition probabilities sum to 1, and the probabilities are non-negative, then Hamilton shows that the optimal estiimates of the transition probabilities are give by:

$$\hat{p}_{ij} = \frac{\sum_{t=2} P\{s_t = j, s_{t-1} = i \mid Y_t; \hat{\theta}\}}{\sum_{t=2} P\{s_{t-1} = i \mid Y_t; \hat{\theta}\}}$$

This formula states that the transition probability is given by the number of times that state j followed state i, relative to the number of times the process was in state i.

If the initial probability vector is taken to be a separate vector of parameters that is constrained only by $1'\rho = 1$, and $\rho \geq 0$, then the MLE of ρ is given by the smoothed inference about the initial state:

$$\hat{\rho} = \hat{\xi}_{1 \mid T}$$

The MLE of α is governed by the following first order condition:

$$\sum \left(\frac{\partial \log(\eta_t)}{\partial \alpha'} \right) \hat{\xi}'_{t|T} = 0$$

Here η is the (N x 1) vector of the stacked densities ($f(y_t \mid s_t = j, z_t, Y_t; \alpha)$) for j = 1,2,...,N, and $\frac{\partial \log(\eta_t)}{\partial \alpha'}$ is the (N x k) matrix of derivatives of the log of these densities with respect to the k parameters.

Example

Suppose we have the model:

$$y_t = z_t \beta_{s_t} + \varepsilon_t$$

Note that the coefficients depend on the state we are in. So with regime 1 we have β_1, and so forth. In this example, the η vector is given by:

$$\begin{bmatrix} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\{\frac{-[(y_t - z_t'\beta_1)]^2}{2\sigma^2} \\ \vdots \\ \vdots \\ \frac{1}{\sqrt{2\pi}\,\sigma} \exp\{\frac{-[(y_t - z_t'\beta_N)]^2}{2\sigma^2} \end{bmatrix}$$

Given this structure, the first order condition for all j becomes:

$$\sum_{t=1} (y - z_t'\hat{\beta}_j) z_t \bullet P\{s_t = j \mid Y_T; \hat{\theta}\} = 0$$

Moreover, the ML estimate of the innovation variance is:

$$\frac{1}{T} \sum_t \sum_j (y_t - z_t'\hat{\beta}_j)^2 \bullet P\{s_t = j \mid Y_t; \hat{\theta}\}$$

Note that the ML estimates show that $\hat{\beta}_j$ satisfy a probability weighted OLS orthongonality condition, where the probabilities are given by the probability that the state was j. Note that the $\hat{\beta}_j$ can be solved from the following euqaitons:

$$\hat{\beta}_j = \{\sum_t (\tilde{z}_t(j)\tilde{z}_t'(j))\}^{-1} \sum_t (\tilde{z}_t(j)\tilde{y}_t(j))$$

where we define:

$$\tilde{z}_t(j) = z_t \sqrt{P\{s_t = j \mid Y_t; \theta\}}$$
$$\tilde{y}_t(j) = y_t \sqrt{P\{s_t = j \mid Y_t; \theta\}}$$

Note also that the ML estimate of σ^2 is the sum of squared residuals from the N regressions.

Dynamic Structural Models - Part II

Previously, we learned how to use GMM to estimate the parameters of structural linear or nonlinear models. This amounted to using a non-linear instrumental variable estimator. GMM is a limited information estimator that ignores certain information. That is, GMM let us estimate parameters in one or more equations without solving out for the equilibrium of the full model.

However, we might be interested in asking how we estimate parameters for an entire model economy, and what information should be used to estimate the parameters. To do these tasks, we need to solve out for the full equilibrium of the model, and use Full Information Maximum Likelihood to estimate the parameters.

Therefore, there are two general steps in quantiatively implementing dynamic models:

(1) **Computing the equilibrium of the economic model** (or computing the optimal allocations, if the economy can be written as a social planner's problem - that is, if the equilibrium is pareto optimal). For our purposes, the equilibrium (or the optimum in the case of a planning problem) consists of a set of equations that relate endogenous quantities and prices to state variables and that admit a closed form solution. Note that this generically may either be a linear or a non-linear mapping between the endogenous and exogenous variables.

(2) **Given the equilibrium laws of motion, choosing values for the parameters in the model.**

Some remarks:

Note that (2) does not require (1) - Recall GMM just needs Euler equations, not the full equilibrium of the model. Note that in general (1) will need (2), in the sense that many of the questions we will ask will depend on parameter values.

Computing Equilibria:

(1) There are two classes of models that we know have closed form solutions. Models with quadratic objective functions, and linear constraints is the first class. It is easy to see why they have closed forms - because the first-order conditions are linear, which combined with the linear constraints, yields linear equilibria.

In particular, they yield linear difference equations for the equilibria. These models can be solved by solving the linear difference equations using any suitable method, such as method of undetermined coefficients, solving it forward, etc. We will review later how this works.

The other class is models with log-linear objectives (i.e. Cobb-Douglas functional forms) with constraints that are log-linear. This class of models includes the one-sector growth model with Cobb-Douglas utility and complete depreciation on capital.

For other cases, generally we need to compute the equilibria approximately. There are many ways to compute equilibria. This falls into two broad categories: linearizing, or log-linearizing the models, and non-linear solution methods.

Linearizing (from here on out we will refer to linearizing and log-linearizing as

"linearizing") can be done two ways. Taking a linear-quadratic approximation of Bellman's equation around the deterministic steady state of the model, and then solving that approximated economy. Since this is the linear-quadratic framework we spoke of earlier, we can get closed form solutions that we can solve explicitly . The other approach is take to the first order conditions of the model, and then linearize those first order conditions around the deterministic steady state.

Since this can be as big (or a bigger) of an issue as choosing parameter values, let's briefly review linearization, which is well described by Harald Uhlig in "Computational Methods for the Study of Dynamic Economies", edited by Ramon Marimon and Andrew Scott, 1999, Oxford University Press.

The basics

Find the necessary conditions characterizing the equilibrium/optimium, including first-order conditions, constraints, etc. Let's call these the necessary equations.

Pick parameter values and find the steady state (we will return to parameter values later). For now, let's assume that we have values for all the parameters in the model.

Solve for the equilibrium laws of motion by solving the difference equations. We will use the method of undetermined coefficients.

Log-linearizing -

Take a Taylor series expansion to replace all equations by approximations around the steady state of the model.

Define:

$$x_t = \log(X_t) - \log(X)$$

For example, first order conditions often have the form:

$$1 = f(x_t, x_{t-1})$$
$$1 = E_t[g(x_{t+1}, x_t)]$$

where f(0,0) = 1, and g(0,0) = 1. The the first order taylor expansion yields:

$$0 \approx f_1 x_1 + f_2 x_2$$
$$0 \approx E_t[g_1 x_{t+1} + g_2 x_t]$$

Recall the first-order Taylor Series Expansion:

$$F(Y) \approx F(Y^*) + F_1(Y^*)[Y - Y^*]$$

One useful approach of what we will do now is that derivatives are often not required. Intuitively, this is because we will be taking linear approximations, and derivatives are nothing more than linear approximations.

To see this, note we can write for x_t near zero:

$$X_t \approx X\exp(x_t)$$

We also have:

$$\exp(x_t + ay_t) \approx 1 + x_t + ay_t$$

$$x_t y_t \approx 0$$

$$E_t[a\exp(x_{t+1})] \approx E_t[ax_{t+1}]$$

This latter approximation holds up to a constant.

These results imply:

$$\exp(x_t) \approx 1 + x_t$$
$$aX_t \approx aXx_t$$
$$(X_t + a)Y_t \approx XYx_t + (X + a)Yy_t$$

These latter 2 approximations hold up to a constant

Log-Linearizing the Growth Model

Let's log-linearize a simple version of the stochastic growth model:

$$\max E_0 \sum \beta^t \{\log(C_t) - \gamma N_t\}$$

subject to:

$$Z_t K_t^{\rho} N_t^{1-\rho} + (1-\delta)K_{t-1} - C_t - K_t = 0$$

$$\log(Z_t) = (1-\psi)\log(Z) + \psi\log(Z_{t-1}) + \varepsilon_t$$

The first order conditions are:

$$1/C_t = \Lambda$$

$$A = \Lambda(1-\rho)Y_t/N_t$$

$$\Lambda_t = \beta E_t\{\Lambda_{t+1}R_{t+1}\}$$

$$R_t = \rho Y_t/K_{t-1} + 1 - \delta$$

The Steady State

$$1 = \beta R$$

$$R = \rho Y/K + 1 - \delta$$

Now, let's write variables as the product of steady state values and deviations from the steady state. For example, we have:

$$C_t = C\exp(c_t)$$

and we can form the other variables in the exact same way.

The resource constraint becomes:

$$C\exp(c_t) + K\exp(k_t) = Y\exp(y_t) + (1-\delta)K\exp(k_{t-1})$$

Approximating, we obtain:

$$C(1+c_t) + K(1+k_t) \approx Y(1+y_t) + (1-\delta)K(1+k_{t-1})$$

Note that the constant terms will drop out, because C $+\delta K = Y$. So we get:

$$Cc_t + Kk_t = Yy_t + (1-\delta)Kk_{t-1}$$

So we now have the resource constraint in percent deviations. Note that we could have also written this explicitly in share form:

$$(C/Y)c_t + (K/Y)(k_t - (1-\delta)k_{t-1}) = y_t$$

Now let's get the other equations in deviation form. Let's do this by presenting the original equation, and then the deviations equation:

$$1/C_t = \Lambda_t$$
$$-c_t = \lambda_t$$

$$A = \Lambda_t(1-\rho)Y_t/N_t$$
$$n_t = y_t + \lambda_t$$

$$R_t = \rho Y_t/K_{t-1} + 1 - \delta$$
$$Rr_t = \rho Y/K(y_t - k_{t-1})$$

$$Y_t = Z\exp(z_t)K_{t-1}^{\rho}N_t^{1-\rho}$$
$$y_t = z_t + \rho k_{t-1} + (1-\rho)n_t$$

$$C_t + K_t = Y_t + (1-\delta)K_{t-1}$$
$$Cc_t + Kk_t = Yy_t + (1-\delta)Kk_{t-1}$$

$$\Lambda_t = \beta E_t[\Lambda_{t+1}R_{t+1}]$$
$$\lambda_t = E_t[\lambda_{t+1} + r_{t+1}]$$

$$\log(Z_t) = (1-\psi)\log(Z) + \psi\log(Z_{t-1}) + \varepsilon_t$$
$$z_t = \psi z_{t-1} + \varepsilon_t$$

Solving the system using the method of undetermined coefficients

The easiest way to solve this out is to write all the variables as linear functions of a vector of lagged endogenous variables and exogenous variables.

Approach 1 - Brute Force

Assume that the linearized equilibrium relationships take the following form:

$$0 = E_t[Fx_{t+1} + Gx_t + Hx_{t-1} + Lz_{t+1} + Mz_t]$$
$$z_{t+1} = Nz_t + \varepsilon_{t+1}$$

We assume that z_t is a stationary process.

Note that the stochastic growth model takes this form - to see this in terms of the variables and the timing of the variables, note that we can write the Euler equation generically as:

$$1/C_t(z_t, K_{t-1}) = \beta E_t(1/C_t(z_{t+1}, K_{t+1})[z_{t+1}F_{K_{t+1}} + 1 - \delta)]$$

We are looking for a solution of the endogenous variables that depend on the state variables:

$$x_t = Px_{t-1} + Qz_t$$

(Recall from dynamic programming, that the solution of a dynamic programming model is such that the control variables are a time invariant function of the state variables.)

It turns out that the matrices must satisfy the following conditions:

$$0 = FP^2 + GP + H$$

Also, given P, let V denote the matrix:

$$V = N'(kron)F + I_k(kron)(FP + G)$$

then we have:

$$VQ = -vec(LN + M)$$

Note that pre-multiplying by V^{-1} solves for Q.

Where did this come from?

Plugging in

$$x_t = Px_{t-1} + Qz_t$$

into

$$0 = E_t[Fx_{t+1} + Gx_t + Hx_{t-1} + Lz_{t+1} + Mz_t]$$

and using

$$z_{t+1} = Nz_t + \varepsilon_{t+1}$$

to form the expectation of z_{t+1} yields:

$$0 = ((FP + G)P + H)x_{t-1} + ((FQ + L)N + (FP + G)Q + M)z_t$$

Since this equation must hold for all possible x_{t-1} and all possible z_t, it follows that

$$((FP + G)P + H) = 0$$
$$((FQ + L)N + (FP + G)Q + M) = 0$$

Equating the coefficient of x_{t-1} to 0 yields the quadratic equation for P as above. Taking the columnwise vectorization fo the coefficient matrices of z and collecting terms in vec(Q) yields the equation for V.

Approach 2

We have a list of endogenous variables of size (m x 1) that we call x, a list of other endogenous variables (jump) variables of size (n x 1) that we call y, and a list of exogenous variables of size (k x 1) that we call z. Note that the distinction between what we are doing now and what we did in the brute force approach before is that before we treated all endogenous variables as being potential endogenous state variables. Now, we will recognize that some of the endogenous variables will not be endogenous state variables. For example, in the stochastic growth model, the only lagged endogenous variable that is a state variable is the capital stock.

The equilibrium relationships are:

$$0 = Ax_t + Bx_{t-1} + Cy_t + Dz_t$$

$$0 = E_t[Fx_{t+1} + Gx_t + Hx_{t-1} + Jy_{t+1} + Ky_t + Lz_{t+1} + Mz_t]$$

$$z_{t+1} = Nz_t + \varepsilon_{t+1}$$

Here we assume that C is (l x n), $l \geq n$ and of rank n, tha F is of size (m + n -1) x n, and that N has stable eigenvalues.

We now look for equations of the form:

$$x_t = Px_{t-1} + Qz_t$$
$$y_t = Rx_{t-1} + Sz_t$$

Solving this out requires solving a quadratic matrix equation.

It turns out that we have the following:

P satisfies the following matrix quadratic equations:

$$0 = C^0AP + C^0B$$

$$0 = (F - JC^+A)P^2 - (JC^+B - G + KC^+A)P - KC^+B + H$$

where we have C^0 is an (l-n) x l matrix whose rows form a basis for the null space of C', and C^+ is the pseudo-inverse of C.

This means:

$$C^0C = 0$$
$$C^+CC^+ = C^+$$
$$CC^+C = C$$

We also have:

$$R = -C^+(AP + B)$$

Given P and R, let V be:

$$V = \begin{bmatrix} I_k(kron)A & I_k(kron)C \\ N'(kron)F + I_k(kron)(FP + JR + G) & N'(kron)J + I_k(kron)K \end{bmatrix}$$

Also, we have:

$$V\begin{bmatrix} vec(Q) \\ vec(S) \end{bmatrix} = -\begin{bmatrix} vec(D) \\ vec(LN + M) \end{bmatrix}$$

Note that if we plug in the equilibrum law of motion from above, we get:

$$(AP + CR + B)x_{t-1} + (AQ + CS + D)z_t = 0$$

Again, this has to hold for all possible deviations in z and x. Thus, the coefficient matrices on x_{t-1} and z_t are equal to 0.

We also have:

$$0 = ((FP + JR + G)P + KR + H)x_{t-1} + ((FQ + JS + L)N + (FP + JR + G)Q + KS + M)z_t$$

Again, the coefficient matrices need to be 0.

Taking the columnwise vectorization of the coefficient matrices of z_t in the last two equations and collecting terms in vec(Q) and vec(S) yields formulae for Q and S.

To find P and R, rewrite the coefficient matrix on x_{t-1} in the second to the last equation as

$$R = -C^{+}(AP + B)$$
$$0 = C^{0}AP + C^{0}$$

Note that $[(C^{+})', (C^{0})']$ is non-singular and that $C^{0}C = 0$. Then, use this equation for R in the coefficient matrix on the large zero equation above, which yields the conjectured solution.

This procedure gave us the equilibrium laws of motion for the full economy, approximated around the deterministic steady state. So this procedure requires you to buy into two things: the first is that you are in the neighborhood of the steady state. Thus, it will be accurate for small deviations around the steady state, but the quality of the approximation will become worse as you move farther away from the steady state. Second, the first-order conditions must hold. Thus, this requires that the economic decision makers in the model are at an interior solution.

Non-linear Solution Methods

(To Be Added)

Choosing Parameter Values

We now can compute the full equilibrium for the economy of interest with equilibrium laws of motion that have been linearly approximated around the deterministic steady state. Now all we need is a set of values for parameters. There are two ways to go here: Statistical estimation using formal loss functions, or calibration. We will discuss both.

There are basically 4 approaches to estimation: (1) Full Information Maximum Likelihood (FIML), (2) GMM, which is a limited information estimator, (3) Simulated method of moments (SMM), (4) Simulated maximum likelihood and simulated pseudo-maximum

likelihood (SML, SMPL). The first approach is restricted basically to linear models with Gaussian innovations.

Basic Differences between Full and limited information estimation

(To be added)

Estimation

Given our linearized economy, we can estimate the parameters using FIML. Before we talk about the mechanics of that, let's be clear on what we are buying into if we do this.

Assumption - We have the right model

Whenever we use a parameteric model to estimate a parameter, we are assuming that the model is the true data generating process (DGP). If not, we do not get the correct value of the parameter in population. To see this, just recall the following. Suppose we are interested in estimating the parameter γ, and the true DGP is:

$$y_t = \gamma x_t + \beta z_t + \varepsilon_t$$

But we estimate:

$$y_t = \phi x_t + u_t$$

Then the relationship between ϕ and γ is given by:

$$\phi = \gamma + \beta cov(z,x)/var(x)$$

So if we estimate a parameter using a misspecified model, then we will typically get a value for the parameter that is biased and inconsistent.

But it is important to bear in mind that the structural models we often use are significant abstractions from reality. Thus, they may be far from the "true" data-generating process.

In any case, with that in mind, let's go ahead and blast off. Our discussion will closely follow "Mechanics of Forming and Estimating Dynamic Linear Economies", Research Department Staff Report 182, available off of the web at:
http://www.minneapolisfed.org/research/sr/sr182.html

FIML estimation of dynamic models is straightforward provided that the model is linear and the innovations are Gaussian. This is the case that we will consider.

Suppose that we have linearized the model and it is in the following form:

$$x_{t+1} = A_0 x_t + B w_{t+1}$$
$$E(ww') = I$$

(as in the case of Uhlig's presentation.)

Note that in the case of the simplest stochastic growth model, the vector x would include hours worked, investment, capital, consumption, output, and the technology shock. Note also that this takes the form of a vector autoregression. Note however that the VAR from this model economy will end up having a number of parameter restrictions imposed by the theory, as opposed to the unrestricted VARs we used previously.

In addition to the linearized VAR equation above, suppose we also have the following measurement error specification:

$$z_t = G x_t + v_t$$
$$v_t = D v_{t-1} + \eta_t$$

We assume that the process v_t is stationary and that η_t is a white noise process that satisfies:

$$E\eta\eta' = R$$
$$E(w_{t+1}\eta_s) = 0, all\ t,s$$

First let's talk about the measurement error. This will be key in the estimation process, since in general we will have too few shocks in the system.

Now define the following:

$$\bar{z}_t = z_{t+1} - D z_t$$

It follows that:

$$\bar{z}_t = (GA_0 - DG)x_t + GCw_{t+1} + \eta_{t+1}$$

This means we can write a state space model for $(x_t, \bar{z}_t)$ as follows:

$$x_{t+1} = A_0 x_t + C w_{t+1}$$
$$\bar{z}_t = \overline{G} x_t + gC w_{t+1} + \eta_{t+1}$$

where $\overline{G} = GA_0 - DG$

Following our previous discusson of state space models, the equation for x is our state equation, and our equation for $\bar{z}_t$ is our measurement equation. Thus, we have a noisy measure of x, which is $\bar{z}_t$.

Hansen, McGrattan, and Sargent proceed to use an "innovations" representation for estimation. Define the following:

$$\begin{aligned}\hat{x}_t &= E[x_t \mid \bar{z}_{t-1}, \ldots \bar{z}_0, \hat{x}_0] \\ u_t &= \bar{z}_t - E[\bar{z}_t \mid \bar{z}_{t-1} \ldots \bar{z}_0, \hat{x}_0] \\ \Omega_t &= E u_t u_t' = \overline{G} \Sigma_t \overline{G} + R + GCC'G'\end{aligned}$$

where $\sum_t$ is the covariance matrix of the state vector x_t

We also have:

$$\Sigma_{t+1} = A_0 \Sigma_t A_0' + CC' - (CC'G' + A_0 \Sigma_t \overline{G}')\Omega_t^{-1}(\overline{G}\Sigma_t A_0' + GCC')$$

We can now write the state space model in "innovations form" as:

$$\begin{aligned}\hat{x}_{t+1} &= A_0 \hat{x}_t + K_t u_t \\ \bar{z}_t &= \overline{G}\hat{x}_t + u_t \\ K_t &= (CC'G' + A_0 \Sigma_t \overline{G}')\Omega_t^{-1}\end{aligned}$$

Now, since $\bar{z}_t$ is a linear combination of z_{t+1} and z_t, it follows that the history of z_{t+1} and the history of $\bar{z}_t$ span the same space, so that

$$u_t = z_{t+1} - E(z_{t+1} \mid z_t, \ldots z_0, \hat{x}_0)$$

So u_t is the innovation in z_{t+1}.

We can now link up the model to a VAR. First, note that:

$$\begin{aligned}\hat{x}_{t+1} &= A_0 \hat{x}_t + K_t u_t \\ z_{t+1} - Dz_t &= \overline{G}\hat{x}_t + u_t\end{aligned}$$

Note that we can derive the following moving average representation:

$$z_{t+1} = [I - DL]^{-1}[I + \overline{G}(I - A_0 L)^{-1} KL]u_t$$

where L is the lag operator.

Doing lots of substitution yields the VAR:

$$z_{t+1} = \phi z_t + u_t$$
$$\phi = \{D + (I - DL)\overline{G}[I - (A_0 - K\overline{G})L]^{-1} KL\}$$

Note that theory imposes significant restrictions on the AR matrix, ϕ.

The Log-Likelihood of the Model

The basic idea is to use the innovations representation to form the standard Gaussian log-likelihood:

$$l(\theta) = \sum \{\log \quad | \ \Omega_t \ | \ + trace(\Omega_t^{-1} u_t u_t')\}$$

We can use either analytical or numerical derivatives. Recall the formula for numerical derivates:

$$\frac{\partial l}{\partial \theta} \approx \frac{l(\theta + \varepsilon e) - l(\theta - \varepsilon e)}{2\varepsilon}$$

where ε is a small, positive number and e is a vector of zeros, except with the value one for the relevant parameter that corresponds to that entry in that vector. In some cases, the optimization problem may involve constraints on the parameter values (for example, we rule out utility function parameters that would violate preference axioms). In this case, we would numerically optimize the likelihood. Programs such as MATLAB have routines that do constrained optimization.

Standard errors for the parameters can be found from the derivatives of the likelihood:

$$S_e = diag(\sqrt{(\sum \frac{\partial l}{\partial \theta} \frac{\partial l}{\partial \theta}')^{-1}}$$

Estimating Time Series Models by Simulation Methods

It is now routine to simulate dynamic, nonlinear rational expectations model to address a variety of questions (see Kydland and Prescott (1982)). It is also possible to estimate the parameters of these models using simulation methods. Some of the initial simulation estimators were advances by McFadden and Pakes and Pollard, who considered the estimation of discrete choice models that arise in cross-sectional applications. We will consider time series applications here. One complication is that we will need to deal with serially correlated disturbances, which McFadden and Pakes/Pollard's estimators don't allow.

We begin with Lee and Ingram's approach. It is described in "Simualtion Estimation of Time Series Models", J*ournal of Econometrics*, 47 (1991), pp. 197-205.

Consider a stochastic nonlinear model which has an m X 1 dimensional equilibrium. We will denote the vector-valued equilibrium stochastic process as $y_j(\beta) \geq 1,$ and we will denote this as $y_j(\beta),$ where j indexes the length of the vector y. Let β be an l X 1 parameter vector that include the primitive parameters in the model, such as parameters describing preferences, technologies, endowments, etc. This vector may include other parameters as well, which we discuss later.

Under the null hypothesis, the model is the true data generating process when evaluated at parameter vector $\beta_0,$ (recall that we have discussed some of the pitfalls associated with this assumption about stochastic, dynamic general equilibrium models.) We will refer to β_0 as the "true" parameter vector. Under the null, there will be an empirical counterpart in actual data to the equilibrium stochastic process from the model.

So we have data from the model $\{y_j(\beta_0)\}$ and actual data that corresponds to that equilibrium stochastic process. We will call the actual data $\{x_t\}.$ Note that in practice, we only have a finite realization of $x_t.$ We can always generate as long a sample as we want from our model.

How does the simulation estimator work? The basic idea is as follows:

Simulate the model, and generate $\{y_j(\beta)\}$. Choose values for the parameter vector β such that we equate the model simulated moments to the actual data moments. Note the following two definitioins:

$$H_T(x) = \frac{1}{T}\sum_{t=1}^{T} h(x_t)$$

$$H_N(y(\beta)) = \frac{1}{N}\sum_{j=1}^{N} h(x_t)$$

Note that $H_T(x)$ is an s X 1 vector of statistics formed as a time series average of some function of observed data, and $H_N(y(\beta))$ is a corresponding vector of statistics calculated from the economic model using simulated data. If the process x and y are ergodic, then we have as T and N go to infinity:

$$H_T(x) \rightarrow E[h(x_t)]$$

$$H_N(y(\beta)) \rightarrow E[h(y_j(\beta))]$$

Moreover, under the null that the economic model is the true DGP, we have:

$$E[h(x_t)] \;= E[h(y_j(\beta))]$$

We now will talk about estimating parameters in the model, exploiting this relationship.

To begin with, we need a weighting matrix, which is of dimension s x s, and we will call this weighting matrix W. Note that it's dimension depends on the the dimension of the actual data we are fitting the model to. We assume that the rank of the weighting matrix is l, which is the dimension of the parameter vector β. We now choose β to minimize the following quadratic form:

$$\beta_{TN} = \arg\min[H_T(x) - H_N(y(\beta))]' W_T [H_T(x) - H_N(y(\beta))]$$

Next, define an integer n = N/T > 1 and define the following functions:

$$g_T(\beta) = \frac{1}{T}\sum_{t=1}^{T} f_t(\beta) = \frac{1}{T}\sum_{t=1}^{T}[h(x_t) - \frac{1}{n}\sum_{k=1}^{n} h(y_{k,t}(\beta))]$$

Note that we are indexing model simulations by the pair (k,t), k = 1,...,n and t = 1,...,T. For example, $y_{k,t} = y_{n(t-1)+k}$. Note that we are assuming that the length of the simulated data series exceed the length of the actual data series. This makes sense, since we can generate the simulated data for free, and it reduces the variances of the estimators - at least that component arising from simulation error.

Asymptotics

This estimator is in the same class as Lar's Hansen's GMM estimator. Establishing consistency and asymptotic normality can therefore be proved showing that the problem satisfies the assumptions that Hansen's proofs require. One condition is that functions are continuous:

$$\lim_{\delta \to 0} E[\sup\{\mid h(y_j(\beta_0)) - h(y_j(\beta_1)) \mid : \beta_0, \beta_1 \in S, \mid \beta_0 - \beta_1 \mid < \delta\}] = 0,$$

where S is the parameter space. Deriving asymptotic distribution of β_{TN} requires some other assumptions. Define $w_t = f_t(\beta_0)$. Let:

$$v_i = E[w_t \mid w_{t-i}, w_{t-i-1}, \ldots] - EE[w_t \mid w_{t-i-1}, w_{t-i-2}, \ldots]$$

Now assume that $Ew_t w_t'$ is finite, that $E[w_t \mid w_{t-i}, w_{t-i-1}, \ldots]$ convergres in mean square to 0, and $\sum E[v_i v_i']^{1/2}$ is finite.

Define

$$R_x(i) = E\{[h(x_t) - E(h(x_t))][h(x_{t-i}) - E(h(x_{t-i}))]'$$
$$R_y(i) = E\{[h(y_t(\beta_0))\} - E\{h(y_t(\beta_0))\}]X[E\{h(y_{t-i}(\beta_0))\} - E\{h(y_{t-i}(\beta_0))\}]'$$

also, define:

$$\Omega = \sum_{i=-\infty}^{\infty} R_x(i)$$

and note that under the null, we also have:

$$\Omega = \sum_{i=-\infty}^{\infty} R_y(i)$$

From Hansen (1982), we have the following asymptotic results:

$$\sqrt{T}[H_T(x) - E(h(x_t))] \rightarrow N(0, \Omega)$$
$$\sqrt{N}[H_N(y(\beta_0)) - E(h(y(\beta_0)))] \rightarrow N(0, \Omega)$$

This implies that:

$$cov\{H_T(x) - H_N(y(\beta_0))\} = (1 + 1/n)\Omega$$

Now, since N/T goes to n as T, N go to infinity, we have

$$\sqrt{T}(\beta_{TN} - \beta_0) \rightarrow N(0, \Lambda)$$
$$\Lambda = | \ (B'WB)^{-1}B'W(1 + 1/n)\Omega WB(B'WB)^{-1'}$$
$$B = E\left[\frac{\partial h(y_j(\beta))}{\partial \beta}\right]$$

Note that the asymptotic covariance matrix for the estimator depends on the choice of the weighting matrix. Recall from Hansen that the optimal choice of the weighting matrix for this problem is:

$$W = [(1 + 1/n)\Omega]^{-1}$$

This implies:

$$\sqrt{T}(\beta_{TN} - \beta_0) \rightarrow N(0, [B'(1 + 1/n)^{-1}\Omega^{-1}B]^{-1})$$

One can obtain a consistent estimator of Ω using Newey and West's procedure. Note again that since simulations are basically free to produce, it is useful to set n to a large number. This lets us reduce the variance of the estimator by reducing simulation noise.

Practical Issues about Simulating

Suppose our model was:

$$y_j = \alpha y_{j-1} + e_j, \mid \alpha \mid < 1$$
$$e_j = \rho e_{j-1} + \omega_j, \mid \rho \mid < 1$$
$$\omega \sim NID(0,1)$$

Suppose also that y is a scalar process. To simulate the process, we need to choose values for the parameters, y_0, and e_0. Given these values and a realization of $\{\omega\}$, we can calculate the sequence of e's and y's. We will want to choose e_0 from its stationary distribution, which means drawing it from a normal $(0,1/(1-\rho^2))$ distribution.

Note that in the linear case, it is easy to construct the unconditional distribution for e_0. However, if y depends on its own past in a non-linear fashion, then it may be hard to insure that the initial realization is drawn from the stationary distribution.

In these cases, one could choose an arbtitrary value of y_0, then simulate the model for 3N periods, and discard the first 2N observations. If the process is stationary and N is large, then the process should be in its stationary distribution after 2N periods.

Note that for the simulations, you should keep the random numbers fixed, rather than drawing new random numbers each time.

Example

Suppose you had an actual sequence of data of y: $\{y_t\}_1^T$.

Consider a first order AR Gaussian process for that data:

$$y_t = \mu + \rho y_{t-1} + \varepsilon_t, \varepsilon \sim N(0, \sigma^2)$$

Now, you wish to estimate the paraterms (μ, ρ, σ^2) using this method. Here is how to proceed. Pick initial values of the parameters.

Draw y_0 from its stationary distribution. Draw $\{\varepsilon\}_2^N$ from its distribution. Keep these random numbers fixed.

Form the sequence $\{\hat{y}_t\}_1^N$

Now form the criterion function:

$$[H_T(x) - H_N(y(\beta))]' W_T [H_T(x) - H_N(y(\beta))]$$

We need to choose some moments to estimate the parameters.

Choose $H_T(X)$ as the mean, the first-order autocovariance, and the variance of the process

$$\frac{1}{T}\sum y_t$$
$$\frac{1}{T-1}\sum (y_t - \bar{y})(y_{t-1} - \bar{y})$$
$$\frac{1}{T-1}\sum (y_t - \bar{y})^2$$

Choose the model analogues, where recall that N is the simulation length:

$$\frac{1}{N}\sum \hat{y}_i$$
$$\frac{1}{N-1}\sum \{(\hat{y}_i - \hat{y})(\hat{y}_{i-1} - \hat{y})\}$$
$$\frac{1}{N-1}\sum (\hat{y}_i - \hat{y})^2$$

Note that the terms $\hat{y}$ in the second equation above refers to the mean of the simulated y series in the model.

Choose the weighting matrix as in Hansen 1982 as the inverse of the asymptotic covariance matrix of the moment conditions, modified along the lines indicated above (Recall that N > T).

Minimize the criterion function. A general approach that can be used in nonlinear as well as linear models is the Newton method. (See below).

Optimization and Testing

Recall we need to minimize:

$$[H_T(x) - H_N(y(\beta))]' W_T [H_T(x) - H_N(y(\beta))]$$

We will set (or approximately set) "l" linear combinations of the "s" statistics equal to 0. There are a variety of ways to do this. One could use grid-based methods, if the dimension of the parameter vector was small. The grid-based method means we search over the discretized parameter space to find the minimum.

Alternatively, we could use Newton-type methods.

Recall how the Newton method works. It is an iterative, linearization techniques to solve for optima.
For this case, let's call the objective function F(θ).

Now define the gradient vector:

$$f(\theta^0) = \frac{\partial F(\theta)}{\partial \theta} \mid_{\theta=\theta^0}$$

Next, define H to be the matrix of second derivatives (multiplied by -1):

$$H(\theta^0) = -\frac{\partial F(\theta)}{\partial \theta \partial \theta'} \mid_{\theta=\theta^0}$$

Now, take a Taylor series expansion of the objective around the θ^0:

$$F(\theta) \cong F(\theta^0) + f(\theta^0)'[\theta - \theta^0] - \frac{1}{2}[\theta - \theta^0]' H(\theta^0)[\theta - \theta^0]$$

The Key idea: minimize the criterion function, which requires differentiating with respect to θ and setting the derivative to 0. This yields:

$$f(\theta^0) = H(\theta^0)[\theta - \theta^0]$$

Now, suppose that θ^0 is an initial guess. The result above shows that an improved guess can be obtained by inverting H to get:

$$\theta - \theta^0 = H(\theta^0)^{-1} f(\theta^0)$$

The Newton-Raphson iterative algorithm therefore becomes:

$$\theta^{m+1} - \theta^m = H(\theta^m)^{-1} f(\theta^m)$$

We continue this iterative procedure until:

$$\mid \theta^{m+1} - \theta^m \mid < c$$

where c is some small number and is known as the convergence criterion. Alternatively, one can also use programmed optimizers. MATLAB and GAUSS both have optimization programs that you can use directly.

Simulated Maximum Likelihood Estimation

The Lee-Ingram paper documented how to estimate parameters of models using simulated method of moments (SMM). We will now discuss how to estimate models using simulated maximum likelihood (SMLE).

The class of models we will consider is broader than just SMLE, which require Gaussian innovations. We will consider models with potentially non-Gaussian innovations, and discuss how we can deal with them. This broader class of models is estimated using pseudo simulated maximum likelihood (PSMLE). The basic reference is Hal White's *Estimation, Inference, and Specification Analysis*, Cambridge University Press, 1994.

The basic idea

Draw shocks from their distributions.

Simulate the endogenous variables, given initial values for their parameters.

Form a log-likelihood based on the first two moments

Maximize the log-likelihood

There are two issues with this estimator. The first is that a closed form-expression for the log-likelihood is not available. In this case, we simulate the likelihood to optimize. The second is if the innovations aren't Gaussian. In this case, we just optimize over the first 2 moments, which of course are sufficient statistics for the Gaussian case. In the non-Gaussian case, it only provides a second-order approximation to the true likelihood. And the accuracy of the approximation depends on how much higher moments deviate from the Gaussian case.

We will focus the discussion around a particular example, which is Krusell, Ohanian, Violante, and Rios-Rull "Capital-Skill Complementarity and Inequality", *Econometrica*, September, 2000, which focused on different substitution elasticities between capital and highly skilled labor, and capital and less-skilled labor.

Facts:

In the last 25 years, big increase in relative wage of college educated workers:

$$\frac{w_{st}}{w_{ut}} \uparrow$$

At the same time, the relative supply of college educated workers has grown:

$$\frac{s_t}{u_t} \uparrow$$

If both the relative price and the relative quantity of skilled labor have increased, then the relative demand for skilled labor has gone WAY up

Economic Question: What shifted the relative demand for labor?

The basic idea: Technological change that helped skilled workers more than unskilled workers

Consider the production function:

$$y_t = f(u_t, s_t, k_{et}) = (k_{et} + u_t)^{\theta} s_t^{1-\theta}$$

Relative price of skilled labor is given by:

$$\frac{f_{st}}{f_{ut}} = \frac{(1-\theta)}{\theta} \frac{(k_{et} + u_t)}{s_t}$$

Note that increases in the stock of capital equipment raise the relative marginal product of skilled labor, which raises the wage premium of skilled to unskilled labor (assuming competition)

$$y_t = c_t + x_{st} + \frac{x_{et}}{q_t} = A_t G(k_{st}, k_{et}, u_t, s_t)$$

$$Y_t = k_{st}^{\alpha}[\mu u_t^{\sigma} + (1-\mu)(\lambda k_{et}^{\rho} + (1-\lambda)s_t^{\rho})^{\frac{\sigma}{\rho}}]^{(1-\alpha)/\sigma}$$

$$u_t = \psi_{ut} h_{ut}$$
$$s_t = \psi_{st} hlst$$

$$\ln(\psi_t) = a_0 + \gamma t + \omega_t$$

The percentage change in the skill premium is:

$$g_{\pi t} = (1-\sigma)(g_{hut} - g_{hst}) + \sigma(g_{\psi st} - g_{\psi ut}) + (\sigma - \rho)\lambda\left(\frac{k_{et}}{s_t}\right)^{\rho}(g_{ket} - g_{hst} - g\psi st)$$

The stochastic elements in the model

There are two sources of randomness in the model: innovations in labor productivity and innovations in te

$$\ln(\psi_t) = a_0 + \gamma t + \omega_t$$

where $\ln(\psi_t)$ is the 2 x 1 vector of log of skilled and less-skilled labor efficiencies.

$$\psi_t = \begin{bmatrix} \gamma_{ut} \\ \gamma_{st} \end{bmatrix}$$

The equations to be estimated are first order conditions for a firm hiring two types of labor, and an arbitrage condition for investors holding the two types of capital:

$$\begin{aligned} \frac{w_{st}h_{st} + w_{ut}h_{ut}}{y_t} &= lsh(\psi_t, X_t; \phi) \\ \frac{w_{st}h_{st}}{w_{ut}h_{ut}} &= wbt(\psi_t, X_t; \phi) \\ (1 - \delta_s) + A_{t+1}G_{kst+1} &= E_t\left(\frac{q_t}{q_{t+1}}\right)(1 - \delta_e) + q_t A_{t+1} G_{ket+1} \\ \varepsilon_t &= (1 - \delta_s) + A_{t+1}G_{kst+1} - \left(\frac{q_t}{q_{t+1}}\right)(1 - \delta_e) + q_t A_{t+1} G_{ket+1} \end{aligned}$$

Estimating the Model Parameters

The problem we face is that we have a nonlinear state space model. It takes the following form. The measurement equation is:

$$z_t = f(X_t, \psi_t, \varepsilon_t; \phi)$$

The state equation is:

$$\ln(\psi_t) = a_0 + \gamma t + \omega_t$$

Note that the function f contains 3 equations: the rate or return equality conditon, and the labor share equations. X is the vector of production inputs.

The nonlinearity of f and the latent nature of labor quality make this problem tricky to deal with. The nonlinearity prevents standard Kalman filtering techniques.

So we use simulated pseudo maximum likelihood. Since there are possible endogeneity issues, we use the two-step version of this estimator.

(1) Generate predicted values of the inputs using a constant, captial stock, and some lagged variables. Denote fitted values as $\tilde{X}_t$

(2) Assume distribution of $\omega_t,$ indexed by i = 1 to S

(3) Generate

$$\ln(\psi_t^i) = a_0 + \gamma t + \omega_t^i$$

(4) Generate

$$Z_t^i = f(\tilde{X}_t, \psi_t^i, \varepsilon_t^i; \phi)$$

Obtain the first and second moments:

$$m_s = \frac{1}{S}\sum f(\tilde{X}_t, \psi_t^i, \varepsilon_t^i; \phi)$$
$$V_s = \frac{1}{S}\sum (Z_t^i - f(\tilde{X}_t, \psi_t^i, \varepsilon_t^i; \phi))(Z_t^i - f(\tilde{X}_t, \psi_t^i, \varepsilon_t^i; \phi))'$$

Given these moments, we can write the criterion function as:

$$l = \frac{1}{2T}\sum_{t=1}^{T}\{(Z_t^i - m_s(\tilde{X}_t, \phi))'(V_s(\tilde{X}_t, \phi))^{-1}(Z_t^i - m_s(\tilde{X}_t, \phi)) + \ln \mid V_s(\tilde{X}_t, \phi)) \mid\}$$

we then maximize that criterion function

The standard error formulae are given in the appendix to the paper.

Calibration

Recall that a drawback to estimating parameters is that we assume the model is the true data generating process. When the model substantially abstracts from reality - as formal behavioral models usually do - then the estimated parameters may be presumed to be biased and inconsistent. An alternative is calibration. Just bring this word up, and you are bound to get a lively discussion going!

So what is up with calibration? Calibration is choosing parameter values for a model economy so that some subset of statistics produced by the model economy is the same as those in actual data.

Everything pretty much boils down to the question you are addressing. This determines the model you use, and ultimately, all the controversy over calibration is really about the model, and not its parameter values. So discussions about calibration are really discussions about what type of theoretical model should be used to address a question.

The basic procedure is as follows:

(1) Write down a fully specified parameteric model.

This means there are state variables that are defined, and there is a mapping between the state variables and the control variables. This mapping is a function of model parameters.

(2) Choose values for the parameters

Basically all the controversy about this practice boils down to the **theoretical model.** Once one writes down the model, there is not much disagreement over what parameter values you choose.

For example, suppose you are using a representative agent one-sector growth model to study

(2) For example, suppose you take a one-sector growth model. It is given by:

$$\max \sum \beta^t u(c_t, 1 - l_t)$$

$$f(k, l) + (1 - \delta)k_t = c_t + k_{t+1}$$

Choosing parameter values boils down to:

$$\beta, \delta, f(k, l), u(c, 1 - l)$$

Note that there is not much disagreement over what values should be assigned to β or δ.

That leaves the production function and the utility function.

For production, this just involves share and substituion elasticty parameters. Again, in this particular model, there is not much disagreement over what values you would assing. Labor's share of output is about .7. the substitution elasticity is between .5 and 1.5.

What about the utility function? Again, the parameters govern shares and subsitution elasticities. There is more uncertainty over these values.

One restriction on the utility function regards balanced growth. Recall that utility functions are consitent with balanced growth for the two following functions:

$$u(c, 1 - l) = \log(c) + g(1 - l)$$

where g is concave. Alternatively, we have:

$$u(c, 1 - l) = \frac{(c_t^{\gamma}(1 - l)^{1-\gamma})^{1-\sigma} - 1}{1 - \sigma}$$

The intertemporal elasticity of substiution in consumption is estimated between close to zero and 1. The intertemporal elasticity of labor is more debatable. Estimates based on micro data for males is close to 0. Estimates for the macroeconomy including women yield a high number.

So at the end of the day, if you are using a one-sector growth model, most people will use parameters that are quite similar, with the possible exception for labor supply elasticities. So

regarding this parameter, one could conduct a sensitivity analysis which includes high and low elasticity numbers.

What is more debatable is whether you would use the one-sector neoclassical growth model, or some alternative theoretical framework.

Some issues to keep in mind when calibrating

(1) **Understand what parameter values matter for the question you address.** For example, in business cycle research, the risk aversion parameter is unimportant. However, labor supply elasticities and the amount of persistence in the technology shock are important. Understanding what is important can be accomplished through sensitivity analyses, in which you vary one parameter's value, keeping other parameter values fixed, and then report how your answer changes as that parameter is varied.

(2) **Don't add lots of free parameters.** There should be a good reason why a feature or parameter is added to a model.

(3) **Try to keep the structure of the model as close as possible to models that are widely used in the literature.** This means that we try to address questions using standard models or the minimum departure from standard models that are required.

(4) **Bring as much evidence as possible to bear on the choice of parameter values.** Too often, we see calibrated models using a parameter that is chosen so that the model matches some feature in the data. However, it may turn out that the parameter value that is chosen is totally at variance with the data. For example, suppose you are studying asset pricing. In representative agent models. We know that the standard model does not do so well in terms of risk premia. Suppose you put in habit formation into the utility function, as in Campbell and Cochrane (JPE). This helps fix the asset pricing implications of the model, but has the negative implication that the business cycle/production predictions of the model are way off.

Nonstationary Time Series

Many economic time series have trends. That is, they tend to grow over time, rather than fluctuate around a constant mean. For example, real GNP, consumption, investment, employment, productivity all grow over time. These variables are nonstartionary. There is an enormous literature in time series devoted to the analysis of nonstationary series. This covers testing for stationarity, the economic implications of nonstationary behavior, and hypothesis testing and forecasting with nonstationary data.

We will now consider 2 different models of nonstationary in a univariate model:

$$y_t = \mu + \gamma t + u_t$$
$$A(L)u_t = \varepsilon_t$$

$$y_t = \mu + y_{t-1} + u_t$$
$$A(L)u_t = \varepsilon_t$$

The first process is called a **trend stationary process**, because it is stationary net of the trend. That is, we will assume that the roots of the coefficient matrix A lie outside the unit circle. Thus, the proper stationary inducing transformation for this process is to remove the γt component:

$$y_t^* = y_t - \gamma t = \mu + u_t$$

The second process is called a **difference stationary**, or integrated process, because it is stationary after first differencing. We will also assume that the roots of the coefficient matrix A lie outside the unit circle. Note that the prooper stationary inducing transformation for this process is to substract y_{t-1} from both sides of the equation:

$$\Delta y_{t-1} = \mu + u_t$$

In general, statistical inference about the parameters in these models will depend on whether the source of non-stationarity is due to a determinstic trend, or whether it is due to a "unit root". This latter process is also called integrated, because it can be written as the accuulated history of all past shocks:

$$y_t = T\mu + \sum_{i=0}^{T} u_{t-i}$$

Note a key difference between the two classes of models regarding the impact of a shock. In the trend stationary model, shocks only have temporary effects. That is, since all the roots of A(L) are outside the unit circle, then the effects of the shock will die out. However, in the difference stationary model, shocks have permanent effects, as evidenced by the above equation. This literally means that shocks that occurred at the beginning of time are still having an effect!

Let's look at the unit root process in a bit more detail.

Consider the stochastic process:

$$y_t = \mu + \gamma t + u_t$$

where u_t follows an ARMA (p,q) process:

$$(1 - \phi_1 L - \ldots)u_t = (1 + \theta_1 L + ..)\varepsilon_t$$

Suppose that $\theta(L)$ is invertible. Now facorize the AR piece as:

$$(1 - \phi_1 L - ..) \ = (1 - \lambda_1 L)(1 - \lambda_2 L)\ldots$$

We also have:

$$u_t = \frac{1 + \theta_1 L + \ldots}{(1 - \lambda_1 L)(1 - \lambda_2 L)\ldots} = \psi(L)\varepsilon_t$$

Now, if all $\lambda's$ are outside the unit circle, then the process is trend stationary, and $\sum(|\ \psi\ |)\ < \infty$

Alternatively, suppose that $\lambda_1 = 1,$ and that all the other roots were outside the unit circle. Then to produce a stationary process, we have:

$$(1 - L)u_t = \frac{1 + \theta_1 L + \ldots}{(1 - \lambda_2 L)(1 - \lambda_3 L)\ldots} = \psi^*(L)\varepsilon_t$$

where $\sum(|\ \psi^*\ |)\ < \infty$

thus, if we first differenced, then we get:

$$(1 - L)y_t = \gamma + \psi^*(L)\varepsilon_t$$

In this case, y_t is a unit root process, that is rendered stationary by differencing. Note that if both λ_1 and λ_2 were equal to 1, then we would need to difference the series twice to achieve stationarity:

$$(1 - L)^2 u_t = \frac{1 + \theta_1 L + \ldots}{(1 - \lambda_3 L)(1 - \lambda_4 L)\ldots} = \psi^*(L)\varepsilon_t$$

$$(1 - L)^2 y_t = \gamma + \psi^*(L)\varepsilon_t$$

There are few time series in economics that require two differences to become stationary.

From here on, we will refer to trend stationary processes as TS, and difference stationary processes as DS

Is it Important to Distinguish Between TS and DS Processes?

A number of years ago, Larry Christiano and Marty Eichenbaum wrote a paper titled "Unit Roots: Do We Know and Do We Care?". Let's pursue this in more detail and figure out when the differences are important, and how easily we can determine whether data are best characterized as TS or DS.

Does it matter for forecasting? Yes

TS and DS models have different implications for forecasting, particularly long-run forecasting. Consider a s-period ahead forecast for y from a TS process. The forecast error is given by:

$$y_{t+s} - y_{t+s|t} = \{\mu + \gamma(t+s) + \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \ldots \psi_{s-1}\varepsilon_{t+1}\}$$

The mean squared forecast error is:

$$E[y_{t+s} - y_{t+s|t}]^2 = \{1 + \psi_1^2 + \ldots \psi_{s-1}^2\}\sigma^2$$

Note that as we take s far into the future, the mean square forecast error tends to the unconditional variance of the stochastic process $\psi(L)\varepsilon_t$. Thus, the mean square forecast error is bounded above the unconditional variance. This means that long-run forecasts from a TS process revert to the trend line.

There is a big difference in the DS case. Consider the mean square forecast error:

$$y_{t+s} - y_{t+s|t} = \{\Delta y_{t+s} + \ldots . \Delta y_{t+1}\} - \{\Delta y_{t+s|t} + \ldots . \Delta y_{t+1|t}\}$$

Now, because shocks have permanent effects, the mean square error grows over the forecast:

$$E[y_{t+s} - y_{t+s|t}]^2 = \{1 + (1 + \psi_1^2) + (1 + \psi_1^2 + \psi_2^2) + \ldots\}.\sigma^2$$

Thus, every time we make an error in forecasting the process one period ahead, that error will continue to have effects s periods in the future. Thus, whether a process is DS or TS has important implications for long-run forecasting.

Does it matter for substantive economics? Probably not.

Charles Nelson and Charles Plosser "Trends and Random Walks", Journal of Monetary Economics, 1982, is a paper that captured a lot of attention in the 1980s and 1990s. This paper argued that most macroeconomic time series were best characterized as unit root processes rather than trend stationary processes. Their paper went on to argue that this had important implications for business cycles. In particular, they argued that since shocks seem to have permanent - rather than transitory - effects, that business cycles were largeley due to real shocks rather than monetary shocks. Their reasoning is fairly simple. They argued that monetary business cycle models predict that monetary shocks should have transitory effects on output and employment. This prediction of the theory was at variance with their finding. In contrast, real business cycle models predict that permanent shocks to technology will have permanent effects on output.

There are 2 basic reasons why economists no longer hold this view.

(1) The near observational equivalence of the processes:

Consider 2 processes - the first is a TS process and the second is a DS process:

$$y_t = \gamma t + u_t$$
$$u_t = .99u_{t-1} + \varepsilon_t$$

$$y_t = \mu + u_t$$
$$u_t = u_{t-1} + \varepsilon_t$$

Note that one process is trend stationary but with the shocks having very very long persistence. The other process is DS, with pemanent shocks. But for short-run (e.g. business cycle frequencies) issues, the processes will behave roughly the same.

(2) Processes with permanent and temporary components

A reasonable interpretation of macro data is that there are transitory shocks (e.g. strikes, wars, bad weather. monetary shocks) and there are permanent shocks (new innovations, permanent changes in rules, regulations, taxes..)

Under this interpretation, the question becomes what is the relative size of permanent vs. transitory shocks. One way to try to sort this out is using a variance ratio test that is due to John Cochran (1988, JPE, pp. 893-920). Consider TS and DS processes:

$$y_t = \mu + \gamma t + \rho y_{t-1} + u_t, -1 < \rho < 1$$
$$A(L)u_t = \varepsilon_t$$

$$y_t = \mu + y_{t-1} + u_t$$
$$A(L)u_t = \varepsilon_t$$

Now, consider taking variances of long-differences:

$$(1/k)var(y_t - y_{t-k})$$

For the random walk model, we have:

$$(1/k)var(y_t - y_{t-k}) = \sigma_\varepsilon^2$$

For the trend-stationary model, we have as k tends to infitinity:

$$(1/k)var(y_t - y_{t-k}) = 0$$

This suggests a way to measure the relative size of the permanent shock. The idea is to form the ratio of the variance of the kth difference relative to the variance of of the first difference:

$$perm = \frac{(1/k)var(y_t - y_{t-k})}{(var(y_t - y_{t-1})}$$

For U.S. GNP, the size of the permanent component is about 30% of the total variation in GNP. This says that a lot of the variation is due to a permanent component. What one could do is form the ratio of k period variances to 2, 3, 4, etc. period variances to get an idea of the persistence of the transitory component.

Does it Matter for Hypothesis Testing? Yes.

It turns out that stationarity plays a key role in testing hypotheses, particularly those based on asymptotic normality. In particular, using normal distribution theory can lead to substantial

bias in test statistics. For example, if one includes a computer-generated random walk in a VAR model and uses the standard distribution theory to test whether the computer generated random walk Granger causes output, it turns out that it does. The bias can be as large as 30% for a 5% nominal test.

Testing for Stationarity

Testing for a unit root is a bit complicated, as it depends on the deterministic regressors in the model. Let's begin with the simplest case, which is a first order AR with white noise disturbances:

$$y_t = \phi y_{t-1} + \varepsilon_t$$

The basic test

Testing the hypothesis inolves estimating an equation that looks like the following:

$$\Delta y_t = \tau' DR_t + \pi y_{t-1} + u_t$$

where DR_t is a vector of deterministic components. Normally this will include a constant term. It also may include a deterministic time trend. These 2 elements are by far the most common encountered in applied work. Once we select the deterministic component, we will then wish to test the unit root hypothesis using a "t"-type test:

$$H_0 : \pi = 0$$

Some points to keep in mind

1. The asymptotic distribution for t_π depends on the set of deterministic regressors included. the test statistic can be found in Wayne Fuller's 1976 textbook. (See the handout from Fuller's book).

2. Suppose that DR omits a variable that is growing at least as fast a rate as any of the other elements in DR. Then under the null of a unit root, the t statistic can be normalized so that its distribution is standard normal.

3. Point 2 sounds good, but wait.....It also turns out that if DR omits a variable from the DGP that is growing at a rate at last as fast as any of the elements in DR, then the power of the test statistic goes to 0!

So the moral of the story is to include all relevant deterministic regressors...

But wait - too much of a good thing can be bad!

4. Adding deterministic regressors beyond those in the DGP reduces power....(this is not surprising – the same issue arises in virtually all testing situations...)

5. Testing when the process is a general ARMA structure. The Dickey-Fuller tests described in Fuller's book need to be modified when the process has serial correlation. This involves including lagged values of the differenced process in the regression so that the residual is white noise. This is known as the augmented Dickey-Fuller test:

$$\Delta y_t = \tau DR_t + \pi y_{t-1} + \sum_{i=1}^{k} \alpha_i \Delta y_{t-i} + u_t$$

One would then use the Dickey-Fuller statistic to test the null hypothesis.

How to choose k?

Pick a maximum number of lags; call it $k_{\max}$. Estimate the regression, and check to see if the last lag is significant using a t-test. If it is, choose $k = k_{\max}$. If not, then drop the last lag and re-estimate the autoregression. Continue until you get a significant lag. If there are no significant lags, then k =0.

Multivariate time series/Cointegration

We will now consider stationarity issues with multiple time series. The classic reference for this is Rob Engle and Clive Granger "Co-integration and Error Correction: Representation, Estimation and Testing", Econometrica, March, 1987, 251-276. We will consider an "n" dimensiional time series of the following form for the "ith" variable:

$$y_{it} = TD_{it} + Z_{it}$$
$$A_i(L)Z_{it} = B_i(L)e_{it}$$

where TD is the deterministic component. We will assume that y contains at most one unit root, and that all other roots lie outside the unit circle. We will also assume that the trend component consists of a constant and a linear time trend.

We will first discuss cointegration. **Cointegration has implications for the efficient specification of VARs.**

The basic idea behind cointegration is as follows. Suppose there are 2 variables, z and x. Suppose both z and x each have one unit root so that they are difference stationary. However, suppose that there exists a linear combination of z and x such that that deviation from this linear combination is stationary:

$$z_t \sim I(1)$$
$$x_t \sim I(1)$$
$$u_t = z_t - \beta x_t \sim I(0),$$

where I(1) means a stochastic process is integrated of order 1 (that is, it needs to be differenced once to achieve stationarity), and I(0) means a variable does not need to be differenced - it already is stationary. Thus, cointegration in this case means the two variables share the same common stochastic trend. **In other words, each process may drift stochastically, but they never drift very far apart from each other.**

More formally, we have the following:

A vector of variables as denoted above is cointegrated if there exists at least one nonzero n-element vector β_i such that $\beta_i' y_t$ is trend stationary. β_i is called a cointegrating vector. If there are r linearly independent vectors, then y is cointegrated with cointregrating rank r. We then define the (n x r) matrix of cointegrating vectors β.

The cointegrating vectors are identifiable up to a scale factor. That is, if $\beta' y_t$ is I(0), then so is $c\beta_i y_t$ for $c \neq 0$.

If there is at least one integrated variable in the vector y, there cannot be more than n-1 linearly independent cointegrating vector. To see this, suppose there are 2 vairables, one I(0)

and one that is I(1). Since a nonstationary variable cannot be combined with a stationary variable to yield a stationary variale, then there is no linearly independent vector that produces a cointegrated vector. Now, suppose that there are 2 I(1) variables, and 1 I(0) variable. Suppose also that the linear combination:

$$y_{1t} + ay_{2t} \sim I(0)$$

Then the cointegrating vector (a,1) is unique up to a scale factor. To see this, note that if another cointegrating vector existed, it could be combined with the first to produce both variables being I(0).

Is cointegration interesting?

At some level, yes. Cointegration is a way of imposing that variables do not deviate much from each other. For example, consider a stochastic growth model driven by productivity shocks that are a random walk. It may be the case that shocks result in a short run deviation between consumption and income, but that these variables exhibit deviation to a one-time shock only temporarily. Cointegration is also useful from a forecasting standpoint. To see this, note that if "x" is integrated, and if x helps predict y in the long-run, then it follows that the two variables will be cointegrated.

Cointegration Represenations

Lets start with the pth order VAR:

$$A(L)z_t = e_t$$
$$z_t = y_t - \delta t - \kappa$$

Now, consider differencing the VAR:

$$\Delta y_t = \mu + \Pi[y_{t-1} - \delta(t-1)] + \sum_{j=1}^{k} \gamma \Delta y_{t-j} + e_t$$

where k = p-1, $\Pi = \sum_i A_i - I$ and $\gamma_i = -\sum_{i=j+1} A$. To analyze cointegration, we need to search for conditions such that both sides of the above equation stationary. Note that the left hand side is statiionary by construction, since we have assumed only 1 unit root. What about the right hand side? This side is stationary only if $\Pi[y_{t-1} - \delta(t-1)]$ is stationary.

To assess this, let's first suppose that Π is a full rank matrix. Then, for all of the elements to be stationary, we require that all n linearly independent combinations of y_{t-1} formed by the rows of Π to be stationary. If this is the case, then it must be true that all elements of y are trend stationary around δ_t. Here, a standard VAR can be used, with a time trend. But what if the rank of Π is zero? This implies that there are no linear combinations of the variables that are trend statonary. In this case, we need to fit the VAR in first differences.

What about the intermediate case when Π is neither zero rank nor full rank? In this case, there are (n x r) matrices α and β such that $\Pi=\alpha\beta'$ To have $\Pi[y_{t-1} - \delta(t-1)]$ we need that

$\beta'\Pi[y_{t-1} - \delta(t-1)]$ is stationary. Thus, β is the matrix whose columns are the linearly independent cointegrating vectors and the rank of Π is the cointegrating rank of y. Note that there is an identification problem here: the parameters α and β are not identified since for any nonsingular matrix F the matrices αF and $\beta F'^{-1}$ yield the same matrix Π. This leads us to ask how we should interpret these parameters

First, let's try to interpret $\beta' y$. Each column of β can be viewed as the linear long run relationship between the integrated series in $\Pi[y_{t-1} - \delta(t-1)]$.

Now, let's define:

$$z_{t-1} = \beta'\Pi[y_{t-1} - \delta(t-1)]$$

Then we have:

$$\Delta y_t = \mu + \alpha z_{t-1} + \sum_{j=1}^{k} \gamma \Delta y_{t-j} + e_t$$

This is called an error correction model. It implies that the change in y depends not only its own past, but also on the deviation from the previous period, z_{t-1}. This implies that α tells us how fast which the vector returns to its long-run value. The importance of this is that wehn both unit roots and cointegration are present in a VAR, then first-differencing all the variables is misspecified.

In general, VARs fall into 3 categories. (1) rank(Π) = n which means that l variables are trend stationary and a levels VAR is fine, (2) rank $\Pi = 0$, in which case all variable should be first differenced, and (3) 0<rank(Π) < r < n. In this case, there is at least one integrated variable and one cointegrating relation. In this case, one should specify the error correction model.

Testing for Cointegration

There are a number of tests for cointegration. One approach to test for CI between y and x is to use a static regression:

$$y_t = \beta x_t + u_t$$

Thus, we can test if u_t is stationary using the Dickey-Fuller/augmented Dickey-Fuller tests. We can also use a Durbin-Watson test. In particular, if y and x are not cointegrated, then u_t should be integrated, which means that the DF statistic should approach 0. Critical values for the DF test and for some other tests are presented in Engle and Granger's paper.