Time Series for Macroeconomics and Finance

John H. Cochrane[1]
Graduate School of Business
University of Chicago
5807 S. Woodlawn.
Chicago IL 60637
(773) 702-3059
john.cochrane@gsb.uchicago.edu

Spring 1997; Pictures added Jan 2005

[1]I thank Giorgio DeSantis for many useful comments on this manuscript.

Contents

Chapter 1

Preface

These notes are intended as a *text* rather than as a *reference.* A text is what you read in order to learn something. A reference is something you look back on after you know the outlines of a subject in order to get difficult theorems exactly right.

The organization is quite different from most books, which really are intended as references. Most books first state a general theorem or apparatus, and then show how applications are special cases of a grand general structure. That's how we organize things that we already know, but that's not how we learn things. We learn things by getting familiar with a bunch of examples, and then seeing how they fit together in a more general framework. And the point is the "examples"—knowing how to *do* something.

Thus, for example, I start with linear ARMA models constructed from normal iid errors. Once familiar with these models, I introduce the concept of stationarity and the Wold theorem that shows how such models are in fact much more general. But that means that the discussion of ARMA processes is not as general as it is in most books, and many propositions are stated in much less general contexts than is possible.

I make no effort to be encyclopedic. One function of a text (rather than a reference) is to decide what an average reader—in this case an average first year graduate student in economics—really needs to know about a subject, and what can be safely left out. So, if you want to know everything about a subject, consult a reference, such as Hamilton's (1993) excellent book.

Chapter 2

What is a time series?

Most data in macroeconomics and finance come in the form of *time series*—a set of repeated observations of the same variable, such as GNP or a stock return. We can write a time series as

$$\{x_1, x_2, \ldots x_T\} \text{ or } \{x_t\},\ t = 1, 2, \ldots T$$

We will treat x_t as a *random variable*. In principle, there is nothing about time series that is arcane or different from the rest of econometrics. The only difference with standard econometrics is that the variables are subscripted t rather than i. For example, if y_t is generated by

$$y_t = x_t\beta + \epsilon_t,\ E(\epsilon_t \mid x_t) = 0,$$

then OLS provides a consistent estimate of β, just as if the subscript was "i" not "t".

The word "time series" is used interchangeably to denote a *sample* $\{x_t\}$, such as GNP from 1947:1 to the present, and a *probability model* for that sample–a statement of the joint distribution of the random variables $\{x_t\}$.

A possible probability model for the joint distribution of a time series $\{x_t\}$ is

$$x_t = \epsilon_t,\ \epsilon_t \sim \text{i.i.d. } \mathcal{N}(0, \sigma_\epsilon^2)$$

i.e, x_t normal and independent over time. However, time series are typically *not* iid, which is what makes them interesting. For example, if GNP today is unusually high, GNP tomorrow is also likely to be unusually high.

It would be nice to use a nonparametric approach–just use histograms to characterize the joint density of $\{.., x_{t-1}, x_t, x_{t+1}, \ldots\}$. Unfortunately, we will not have enough data to follow this approach in macroeconomics for at least the next 2000 years or so. Hence, time-series consists of interesting parametric *models* for the joint distribution of $\{x_t\}$. The models impose structure, which you must evaluate to see if it captures the features you think are present in the data. In turn, they reduce the estimation problem to the estimation of a few *parameters* of the time-series model.

The first set of models we study are *linear ARMA models*. As you will see, these allow a convenient and flexible way of studying time series, and capturing the extent to which series can be forecast, i.e. variation over time in conditional means. However, they don't do much to help model variation in conditional variances. For that, we turn to ARCH models later on.

Chapter 3

ARMA models

3.1 White noise

The building block for our time series models is the *white noise* process, which I'll denote ϵ_t. In the least general case,

$$\epsilon_t \sim \text{i.i.d. } N(0, \sigma_\epsilon^2)$$

Notice three implications of this assumption:

1. $E(\epsilon_t) = E(\epsilon_t \mid \epsilon_{t-1}, \epsilon_{t-2} \ldots) = E(\epsilon_t \mid \text{all information at } t-1) = 0.$
2. $E(\epsilon_t \epsilon_{t-j}) = \text{cov}(\epsilon_t \epsilon_{t-j}) = 0$
3. var (ϵ_t) =var $(\epsilon_t \mid \epsilon_{t-1}, \epsilon_{t-2}, \ldots)$ =var $(\epsilon_t \mid \text{all information at } t-1) = \sigma_\epsilon^2$

The first and second properties are the absence of any *serial correlation* or *predictability*. The third property is *conditional homoskedasticity* or a constant conditional variance.

Later, we will generalize the building block process. For example, we may assume property 2 and 3 without normality, in which case the ϵ_t need not be independent. We may also assume the first property only, in which case ϵ_t is a *martingale difference sequence*.

By itself, ϵ_t is a pretty boring process. If ϵ_t is unusually high, there is no tendency for ϵ_{t+1} to be unusually high or low, so it does not capture the interesting property of persistence that motivates the study of time series. More realistic models are constructed by taking combinations of ϵ_t.

3.2 Basic ARMA models

Most of the time we will study a class of models created by taking linear combinations of white noise. For example,

$$\begin{array}{rc} \text{AR(1):} & x_t = \phi x_{t-1} + \epsilon_t \\ \text{MA(1):} & x_t = \epsilon_t + \theta \epsilon_{t-1} \\ \text{AR(p):} & x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \epsilon_t \\ \text{MA(q):} & x_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots \theta_q \epsilon_{t-q} \\ \text{ARMA(p,q):} & x_t = \phi_1 x_{t-1} + ... + \epsilon_t + \theta \epsilon_{t-1+...} \end{array}$$

As you can see, each case amounts to a recipe by which you can construct a sequence $\{x_t\}$ given a sequence of realizations of the white noise process $\{\epsilon_t\}$, and a starting value for x.

All these models are mean zero, and are used to represent deviations of the series about a mean. For example, if a series has mean $\bar{x}$ and follows an AR(1)

$$(x_t - \bar{x}) = \phi(x_{t-1} - \bar{x}) + \epsilon_t$$

it is equivalent to

$$x_t = (1 - \phi)\bar{x} + \phi x_{t-1} + \epsilon_t.$$

Thus, constants absorb means. I will generally only work with the mean zero versions, since adding means and other deterministic trends is easy.

3.3 Lag operators and polynomials

It is easiest to represent and manipulate ARMA models in *lag operator* notation. The lag operator moves the index back one time unit, i.e.

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

More formally, L is an *operator* that takes one whole time series $\{x_t\}$ and produces another; the second time series is the same as the first, but moved backwards one date. From the definition, you can do fancier things:

$$L^2 x_t = LLx_t = Lx_{t-1} = x_{t-2}$$

$$L^j x_t = x_{t-j}$$
$$L^{-j} x_t = x_{t+j}.$$

We can also define *lag polynomials*, for example

$$a(L)x_t = (a_0 L^0 + a_1 L^1 + a_2 L^2)x_t = a_0 x_t + a_1 x_{t-1} + a_2 x_{t-2}.$$

Using this notation, we can rewrite the ARMA models as

$$\begin{array}{rc} \text{AR(1):} & (1-\phi L)x_t = \epsilon_t \\ \text{MA(1):} & x_t = (1+\theta L)\epsilon_t \\ \text{AR(p):} & (1+\phi_1 L + \phi_2 L^2 + \ldots + \phi_p L^p)x_t = \epsilon_t \\ \text{MA(q):} & x_t = (1+\theta_1 L + \ldots \theta_q L^q)\epsilon_t \end{array}$$

or simply

$$\begin{array}{rc} \text{AR:} & a(L)x_t = \epsilon_t \\ \text{MA:} & x_t = b(L)\epsilon \\ \text{ARMA:} & a(L)x_t = b(L)\epsilon_t \end{array}$$

3.3.1 Manipulating ARMAs with lag operators.

ARMA models are not unique. A time series with a given joint distribution of $\{x_0, x_1, \ldots x_T\}$ can usually be represented with a variety of ARMA models. It is often convenient to work with different representations. For example, 1) the shortest (or only finite length) polynomial representation is obviously the easiest one to work with in many cases; 2) AR forms are the easiest to estimate, since the OLS assumptions still apply; 3) moving average representations express x_t in terms of a linear combination of independent right hand variables. For many purposes, such as finding variances and covariances in sec. 4 below, this is the easiest representation to use.

3.3.2 AR(1) to MA(∞) by recursive substitution

Start with the AR(1)

$$x_t = \phi x_{t-1} + \epsilon_t.$$

Recursively substituting,

$$x_t = \phi(\phi x_{t-2} + \epsilon_{t-1}) + \epsilon_t = \phi^2 x_{t-2} + \phi\epsilon_{t-1} + \epsilon_t$$

$$x_t = \phi^k x_{t-k} + \phi^{k-1}\epsilon_{t-k+1} + \ldots + \phi^2\epsilon_{t-2} + \phi\epsilon_{t-1} + \epsilon_t$$

Thus, an AR(1) can always be expressed as an ARMA(k,k-1). More importantly, if $\mid \phi \mid < 1$ so that $\lim_{k\to\infty} \phi^k x_{t-k} = 0$, then

$$x_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$$

so the AR(1) can be expressed as an MA(∞).

3.3.3 AR(1) to MA(∞) with lag operators.

These kinds of manipulations are much easier using lag operators. To invert the AR(1), write it as

$$(1 - \phi L)x_t = \epsilon_t.$$

A natural way to "invert" the AR(1) is to write

$$x_t = (1 - \phi L)^{-1}\epsilon_t.$$

What meaning can we attach to $(1 - \phi L)^{-1}$? We have only defined polynomials in L so far. Let's try using the expression

$$(1 - z)^{-1} = 1 + z + z^2 + z^3 + \ldots \text{ for } \mid z \mid < 1$$

(you can prove this with a Taylor expansion). This expansion, with the hope that $\mid \phi \mid < 1$ implies $\mid \phi L \mid < 1$ in some sense, suggests

$$x_t = (1 - \phi L)^{-1}\epsilon_t = (1 + \phi L + \phi^2 L^2 + \ldots)\epsilon_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$$

which is the same answer we got before. (At this stage, treat the lag operator as a suggestive notation that delivers the right answer. We'll justify that the method works in a little more depth later.)

Note that we can't always perform this inversion. In this case, we required $\mid \phi \mid < 1$. Not all ARMA processes are *invertible* to a representation of x_t in terms of current and past ϵ_t.

3.3.4 AR(p) to MA(∞), MA(q) to AR(∞), factoring lag polynomials, and partial fractions

The AR(1) example is about equally easy to solve using lag operators as using recursive substitution. Lag operators shine with more complicated models. For example, let's invert an AR(2). I leave it as an exercise to try recursive substitution and show how hard it is.

To do it with lag operators, start with

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon_t.$$

$$(1 - \phi_1 L - \phi_2 L^2) x_t = \epsilon_t$$

I don't know any expansion formulas to apply directly to $(1-\phi_1 L-\phi_2 L^2)^{-1}$, but we can use the $1/(1-z)$ formula by *factoring the lag polynomial.* Thus, find λ_1 and λ_2 such that.

$$(1 - \phi_1 L - \phi_2 L^2) = (1 - \lambda_1 L)(1 - \lambda_2 L)$$

The required vales solve

$$\lambda_1 \lambda_2 = -\phi_2$$

$$\lambda_1 + \lambda_2 = \phi_1.$$

(Note λ_1 and λ_2 may be equal, and they may be complex.)

Now, we need to invert

$$(1 - \lambda_1 L)(1 - \lambda_2 L) x_t = \epsilon_t.$$

We do it by

$$x_t = (1-\lambda_1 L)^{-1}(1-\lambda_2 L)^{-1}\epsilon_t$$

$$x_t = (\sum_{j=0}^{\infty}\lambda_1^j L^j)(\sum_{j=0}^{\infty}\lambda_2^j L^j)\epsilon_t.$$

Multiplying out the polynomials is tedious, but straightforward.

$$(\sum_{j=0}^{\infty}\lambda_1^j L^j)(\sum_{j=0}^{\infty}\lambda_2^j L^j) = (1+\lambda_1 L+\lambda_1^2 L^2+\ldots)(1+\lambda_2 L+\lambda_2^2 L^2+\ldots) =$$

$$1+(\lambda_1+\lambda_2)L+(\lambda_1^2+\lambda_1\lambda_2+\lambda_2^2)L^2+\ldots = \sum_{j=0}^{\infty}(\sum_{k=0}^{j}\lambda_1^k\lambda_2^{j-k})L^j$$

There is a prettier way to express the MA(∞). Here we use the *partial fractions* trick. We find a and b so that

$$\frac{1}{(1-\lambda_1 L)(1-\lambda_2 L)} = \frac{a}{(1-\lambda_1 L)} + \frac{b}{(1-\lambda_2 L)} = \frac{a(1-\lambda_2 L)+b(1-\lambda_1 L)}{(1-\lambda_1 L)(1-\lambda_2 L)}.$$

The numerator on the right hand side must be 1, so

$$a+b=1$$

$$\lambda_2 a+\lambda_1 b=0$$

Solving,

$$b=\frac{\lambda_2}{\lambda_2-\lambda_1}, a=\frac{\lambda_1}{\lambda_1-\lambda_2},$$

so

$$\frac{1}{(1-\lambda_1 L)(1-\lambda_2 L)} = \frac{\lambda_1}{(\lambda_1-\lambda_2)}\frac{1}{(1-\lambda_1 L)} + \frac{\lambda_2}{(\lambda_2-\lambda_1)}\frac{1}{(1-\lambda_2 L)}.$$

Thus, we can express x_t as

$$x_t = \frac{\lambda_1}{\lambda_1-\lambda_2}\sum_{j=0}^{\infty}\lambda_1^j\epsilon_{t-j} + \frac{\lambda_2}{\lambda_2-\lambda_1}\sum_{j=0}^{\infty}\lambda_2^j\epsilon_{t-j}.$$

$$x_t = \sum_{j=0}^{\infty} (\frac{\lambda_1}{\lambda_1 - \lambda_2}\lambda_1^j + \frac{\lambda_2}{\lambda_2 - \lambda_1}\lambda_2^j)\epsilon_{t-j}$$

This formula should remind you of the solution to a second-order difference or differential equation–the response of x to a shock ϵ is a sum of two exponentials, or (if the λ are complex) a mixture of two damped sine and cosine waves.

AR(p)'s work exactly the same way. Computer programs exist to find roots of polynomials of arbitrary order. You can then multiply the lag polynomials together or find the partial fractions expansion. Below, we'll see a way of writing the $AR(p)$ as a *vector* $AR(1)$ that makes the process even easier.

Note again that not every $AR(2)$ can be inverted. We require that the $\lambda's$ satisfy $\mid \lambda \mid < 1$, and one can use their definition to find the implied allowed region of ϕ_1 and ϕ_2. Again, until further notice, we will only use invertible ARMA models.

Going from MA to $AR(\infty)$ is now obvious. Write the MA as

$$x_t = b(L)\epsilon_t,$$

and so it has an $AR(\infty)$ representation

$$b(L)^{-1}x_t = \epsilon_t.$$

3.3.5 Summary of allowed lag polynomial manipulations

In summary. one can manipulate lag polynomials pretty much just like regular polynomials, as if L was a number. (We'll study the theory behind them later, and it is based on replacing L by z where z is a complex number.) Among other things,

1) We can multiply them

$$a(L)b(L) = (a_0 + a_1L + ..)(b_0 + b_1L + ..) = a_0b_0 + (a_0b_1 + b_0a_1)L + \dots$$

2) They commute:

$$a(L)b(L) = b(L)a(L)$$

(you should prove this to yourself).

3) We can raise them to positive integer powers

$$a(L)^2 = a(L)a(L)$$

4) We can invert them, by factoring them and inverting each term

$$a(L) = (1 - \lambda_1 L)(1 - \lambda_2 L)\dots$$

$$a(L)^{-1} = (1 - \lambda_1 L)^{-1}(1 - \lambda_2 L)^{-1}\dots = \sum_{j=0}^{\infty} \lambda_1^j L^j \sum_{j=0}^{\infty} \lambda_2^j L^j \dots =$$

$$= c_1(1 - \lambda_1 L)^{-1} + c_2(1 - \lambda_2 L)^{-1} ...$$

We'll consider roots greater than and/or equal to one, fractional powers, and non-polynomial functions of lag operators later.

3.4 Multivariate ARMA models.

As in the rest of econometrics, multivariate models look just like univariate models, with the letters reinterpreted as vectors and matrices. Thus, consider a *multivariate time series*

$$x_t = \begin{bmatrix} y_t \\ z_t \end{bmatrix}.$$

The building block is a multivariate white noise process, $\epsilon_t \tilde{}$ iid $N(0, \Sigma)$, by which we mean

$$\epsilon_t = \begin{bmatrix} \delta_t \\ \nu_t \end{bmatrix};\ E(\epsilon_t) = 0,\ E(\epsilon_t \epsilon_t') = \Sigma = \begin{bmatrix} \sigma_\delta^2 & \sigma_{\delta\nu} \\ \sigma_{\delta\nu} & \sigma_\nu^2 \end{bmatrix},\ E(\epsilon_t \epsilon_{t-j}') = 0.$$

(In the section on orthogonalizing VAR's we'll see how to start with an even simpler building block, δ and ν uncorrelated or $\Sigma = I$.)

The AR(1) is $x_t = \phi x_{t-1} + \epsilon_t$. Reinterpreting the letters as appropriately sized matrices and vectors,

$$\left[\begin{array}{c} y_t \\ z_t \end{array} \right] = \left[\begin{array}{cc} \phi_{yy} & \phi_{yz} \\ \phi_{zy} & \phi_{zz} \end{array} \right] \left[\begin{array}{c} y_{t-1} \\ z_{t-1} \end{array} \right] + \left[\begin{array}{c} \delta_t \\ \nu_t \end{array} \right]$$

or

$$y_t = \phi_{yy} y_{t-1} + \phi_{yz} z_{t-1} + \delta_t$$
$$z_t = \phi_{zy} y_{t-1} + \phi_{zz} z_{t-1} + \nu_t$$

Notice that *both* lagged y and lagged z appear in each equation. Thus, the vector AR(1) captures cross-variable dynamics. For example, it could capture the fact that when M1 is higher in one quarter, GNP tends to be higher the following quarter, as well as the facts that if GNP is high in one quarter, GNP tends to be higher the following quarter.

We can write the vector AR(1) in lag operator notation, $(I - \phi L)x_t = \epsilon_t$ or

$$A(L)x_t = \epsilon_t.$$

I'll use capital letters to denote such *matrices of lag polynomials.*

Given this notation, it's easy to see how to write multivariate ARMA models of arbitrary orders:

$$A(L)x_t = B(L)\epsilon_t,$$

where

$$A(L) = I - \Phi_1 L - \Phi_2 L^2 \ldots; \; B(L) = I + \Theta_1 L + \Theta_2 L^2 + \ldots, \; \Phi_j = \left[\begin{array}{cc} \phi_{j,yy} & \phi_{j,yz} \\ \phi_{j,zy} & \phi_{j,zz} \end{array} \right],$$

and similarly for Θ_j. The way we have written these polynomials, the first term is I, just as the scalar lag polynomials of the last section always start with 1. Another way of writing this fact is $A(0) = I$, $B(0) = I$. As with Σ, there are other equivalent representations that do not have this feature, which we will study when we orthogonalize VARs.

We can invert and manipulate multivariate ARMA models in obvious ways. For example, the MA(∞)representation of the multivariate AR(1) must be

$$(I - \Phi L)x_t = \epsilon_t \Leftrightarrow x_t = (I - \Phi L)^{-1}\epsilon_t = \sum_{j=0}^{\infty} \Phi^j \epsilon_{t-j}$$

More generally, consider inverting an arbitrary AR(p),

$$A(L)x_t = \epsilon_t \Leftrightarrow x_t = A(L)^{-1}\epsilon_t.$$

We can interpret the matrix inverse as a product of sums as above, or we can interpret it with the matrix inverse formula:

$$\begin{bmatrix} a_{yy}(L) & a_{yz}(L) \\ a_{zy}(L) & a_{zz}(L) \end{bmatrix} \begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \delta_t \\ \nu_t \end{bmatrix} \Rightarrow$$

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = (a_{yy}(L)a_{zz}(L) - a_{zy}(L)a_{yz}(L))^{-1} \begin{bmatrix} a_{zz}(L) & -a_{yz}(L) \\ -a_{zy}(L) & a_{yy}(L) \end{bmatrix} \begin{bmatrix} \delta_t \\ \nu_t \end{bmatrix}$$

We take inverses of scalar lag polynomials as before, by factoring them into roots and inverting each root with the $1/(1-z)$ formula.

Though the above are useful ways to think about what inverting a matrix of lag polynomials means, they are not particularly good algorithms for *doing it*. It is far simpler to simply simulate the response of x_t to shocks. We study this procedure below.

The name *vector autoregression* is usually used in the place of "vector ARMA" because it is very uncommon to estimate moving average terms. Autoregressions are easy to estimate since the OLS assumptions still apply, where MA terms have to be estimated by maximum likelihood. Since every MA has an AR(∞) representation, pure autoregressions can approximate vector MA processes, if you allow enough lags.

3.5 Problems and Tricks

There is an enormous variety of clever tricks for manipulating lag polynomials beyond the factoring and partial fractions discussed above. Here are a few.

1. You can invert finite-order polynomials neatly by matching representations. For example, suppose $a(L)x_t = b(L)\epsilon_t$, and you want to find the moving average representation $x_t = d(L)\epsilon_t$. You could try to crank out $a(L)^{-1}b(L)$ directly, but that's not much fun. Instead, you could find $d(L)$ from $b(L)\epsilon_t = a(L)x_t = a(L)d(L)\epsilon_t$, hence

$$b(L) = a(L)d(L),$$

and matching terms in L^j to make sure this works. For example, suppose $a(L) = a_0 + a_1 L$, $b(L) = b_0 + b_1 L + b_2 L^2$. Multiplying out $d(L) = (a_o + a_1 L)^{-1}(b_0 + b_1 L + b_2 L^2)$ would be a pain. Instead, write

$$b_0 + b_1 L + b_2 L^2 = (a_0 + a_1 L)(d_0 + d_1 L + d_2 L^2 + ...).$$

Matching powers of L,

$$\begin{array}{rc} b_0 = & a_0 d_0 \\ b_1 = & a_1 d_0 + a_0 d_1 \\ b_2 = & a_1 d_1 + a_0 d_2 \\ 0 = & a_1 d_{j+1} + a_0 d_j; \quad j \geq 3. \end{array}$$

which you can easily solve recursively for the d_j. (Try it.)

Chapter 4

The autocorrelation and autocovariance functions.

4.1 Definitions

The *autocovariance* of a series x_t is defined as

$$\gamma_j = \text{cov}(x_t, x_{t-j})$$

(Covariance is defined as cov $(x_t, x_{t-j}) = E(x_t - E(x_t))(x_{t-j} - E(x_{t-j}))$, in case you forgot.) Since we are specializing to ARMA models without constant terms, $E(x_t) = 0$ for all our models. Hence

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

Note $\gamma_0 = \text{var}(x_t)$

A related statistic is the correlation of x_t with x_{t-j} or *autocorrelation*

$$\rho_j = \gamma_j/\text{var}(x_t) = \gamma_j/\gamma_0.$$

My notation presumes that the covariance of x_t and x_{t-j} is the same as that of x_{t-1} and x_{t-j-1}, etc., i.e. that it depends only on the separation between two xs, j, and not on the absolute date t. You can easily verify that invertible ARMA models posses this property. It is also a deeper property called *stationarity*, that I'll discuss later.

We constructed ARMA models in order to produce interesting models of the joint distribution of a time series $\{x_t\}$. Autocovariances and autocorrelations are one obvious way of *characterizing* the joint distribution of a time series so produced. The correlation of x_t with x_{t+1} is an obvious measure of how persistent the time series is, or how strong is the tendency for a high observation today to be followed by a high observation tomorrow.

Next, we calculate the autocorrelations of common ARMA processes, both to characterize them, and to gain some familiarity with manipulating time series.

4.2 Autocovariance and autocorrelation of ARMA processes.

White Noise.

Since we assumed $\epsilon_t \sim iid\, \mathcal{N}(0, \sigma_\epsilon^2)$, it's pretty obvious that

$$\gamma_0 = \sigma_\epsilon^2, \gamma_j = 0 \text{ for } j \neq 0$$

$$\rho_0 = 1, \rho_j = 0 \text{ for } j \neq 0.$$

MA(1)

The model is:

$$x_t = \epsilon_t + \theta\epsilon_{t-1}$$

Autocovariance:

$$\gamma_0 = var(x_t) = var(\epsilon_t + \theta\epsilon_{t-1}) = \sigma_\epsilon^2 + \theta^2\sigma_\epsilon^2 = (1+\theta^2)\sigma_\epsilon^2$$

$$\gamma_1 = E(x_t x_{t-1}) = E((\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2}) = E(\theta\epsilon_{t-1}^2) = \theta\sigma_\epsilon^2$$

$$\gamma_2 = E(x_t x_{t-2}) = E((\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2}) = 0$$

$$\gamma_3, \ldots = 0$$

Autocorrelation:

$$\rho_1 = \theta/(1+\theta^2); \quad \rho_2, \ldots = 0$$

MA(2)

Model:

$$x_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$$

Autocovariance:

$$\gamma_0 = E[(\epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2})^2] = (1 + \theta_1^2 + \theta_2^2)\sigma_\epsilon^2$$

$$\gamma_1 = E[(\epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2})(\epsilon_{t-1} + \theta_1 \epsilon_{t-2} + \theta_2 \epsilon_{t-3})] \ = (\theta_1 + \theta_1\theta_2)\sigma_\epsilon^2$$

$$\gamma_2 = E[(\epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2})(\epsilon_{t-2} + \theta_1 \epsilon_{t-3} + \theta_2 \epsilon_{t-4})] \ = \ \theta_2\sigma_\epsilon^2$$

$$\gamma_3, \gamma_4, \ldots = 0$$

Autocorrelation:

$$\rho_0 = 1$$

$$\rho_1 = (\theta_1 + \theta_1\theta_2)/(1 + \theta_1^2 + \theta_2^2)$$

$$\rho_2 = \ \theta_2/(1 + \theta_1^2 + \theta_2^2)$$

$$\rho_3, \rho_4, \ldots = 0$$

MA(q), MA(∞)

By now the pattern should be clear: MA(q) processes have q autocorrelations different from zero. Also, it should be obvious that if

$$x_t = \theta(L)\epsilon_t = \sum_{j=0}^{\infty} (\theta_j L^j)\epsilon_t$$

then

$$\gamma_0 = \text{var}(x_t) = \left(\sum_{j=0}^{\infty} \theta_j^2\right) \sigma_\epsilon^2$$

$$\gamma_k = \sum_{j=0}^{\infty} \theta_j \theta_{j+k} \sigma_\epsilon^2$$

and formulas for ρ_j follow immediately.

There is an important lesson in all this. Calculating second moment properties is easy for MA processes, since all the covariance terms ($E(\epsilon_j \epsilon_k)$) drop out.

AR(1)

There are two ways to do this one. First, we might use the $MA(\infty)$ representation of an $AR(1)$, and use the MA formulas given above. Thus, the model is

$$(1-\phi L)x_t = \epsilon_t \Rightarrow x_t = (1-\phi L)^{-1}\epsilon_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$$

so

$$\gamma_0 = \left(\sum_{j=0}^{\infty} \phi^{2j} \right) \sigma_\epsilon^2 \; = \; \frac{1}{1-\phi^2}\sigma_\epsilon^2; \;\; \rho_0 = 1$$

$$\gamma_1 = \left(\sum_{j=0}^{\infty} \phi^j \phi^{j+1} \right) \sigma_\epsilon^2 = \phi \left(\sum_{j=0}^{\infty} \phi^{2j} \right) \sigma_\epsilon^2 = \frac{\phi}{1-\phi^2}\sigma_\epsilon^2; \;\; \rho_1 = \phi$$

and continuing this way,

$$\gamma_k = \frac{\phi^k}{1-\phi^2}\sigma_\epsilon^2; \;\; \rho_k = \phi^k.$$

There's another way to find the autocorrelations of an $AR(1)$, which is useful in its own right.

$$\gamma_1 = E(x_t x_{t-1}) = E((\phi x_{t-1} + \epsilon_t)(x_{t-1})) = \phi\sigma_x^2; \;\; \rho_1 = \phi$$

$$\gamma_2 = E(x_t x_{t-2}) = E((\phi^2 x_{t-2} + \phi\epsilon_{t-1} + \epsilon_t)(x_{t-2})) = \phi^2\sigma_x^2; \;\; \rho_2 = \phi^2$$

$$\ldots$$

$$\gamma_k = E(x_t x_{t-k}) = E((\phi^k x_{t-k} + \epsilon \ldots)(x_{t-k}) = \phi^k\sigma_x^2; \;\; \rho_k = \phi^k$$

AR(p); Yule-Walker equations

This latter method turns out to be the easy way to do AR(p)'s. I'll do an AR(3), then the principle is clear for higher order AR's

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \epsilon_t$$

multiplying both sides by x_t, $x_{t-1}, ...$, taking expectations, and dividing by γ_0, we obtain

$$1 = \phi_1\rho_1 + \phi_2\rho_2 + \phi_3\rho_3 + \sigma_\epsilon^2/\gamma_0$$

$$\rho_1 = \phi_1 + \phi_2\rho_1 + \phi_3\rho_2$$

$$\rho_2 = \phi_1\rho_1 + \phi_2 + \phi_3\rho_1$$

$$\rho_3 = \phi_1\rho_2 + \phi_2\rho_1 + \phi_3$$

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \phi_3\rho_{k-3}$$

The second, third and fourth equations can be solved for ρ_1, ρ_2 and ρ_3. Then each remaining equation gives ρ_k in terms of ρ_{k-1} and ρ_{k-2}, so we can solve for all the ρs. Notice that the ρs follow the same difference equation as the original x's. Therefore, past ρ_3, the ρ's follow a mixture of damped sines and exponentials.

The first equation can then be solved for the variance,

$$\sigma_x^2 = \gamma_0 = \frac{\sigma_\epsilon^2}{1 - (\phi_1\rho_1 + \phi_2\rho_2 + \phi_3\rho_3)}$$

4.2.1 Summary

The pattern of autocorrelations as a function of lag — ρ_j as a function of j — is called the *autocorrelation function.* The $MA(q)$ process has q (potentially) non-zero autocorrelations, and the rest are zero. The $AR(p)$ process has p (potentially) non-zero autocorrelations with no particular pattern, and then the autocorrelation function dies off as a mixture of sines and exponentials.

One thing we learn from all this is that ARMA models are capable of capturing very complex patters of temporal correlation. Thus, they are a useful and interesting class of models. In fact, they can capture *any* valid autocorrelation! If all you care about is autocorrelation (and not, say, third moments, or nonlinear dynamics), then ARMAs are as general as you need to get!

Time series books (e.g. Box and Jenkins ()) also define a *partial autocorrelation function*. The j'th partial autocorrelation is related to the coefficient

on x_{t-j} of a regression of x_t on $x_{t-1}, x_{t-2}, \ldots, x_{t-j}$. Thus for an $AR(p)$, the $p+1$th and higher *partial* autocorrelations are zero. In fact, the partial autocorrelation function behaves in an exactly symmetrical fashion to the autocorrelation function: the partial autocorrelation of an $MA(q)$ is damped sines and exponentials after q.

Box and Jenkins () and subsequent books on time series aimed at forecasting advocate inspection of autocorrelation and partial autocorrelation functions to "identify" the appropriate "parsimonious" AR, MA or ARMA process. I'm not going to spend any time on this, since the procedure is not much followed in economics anymore. With rare exceptions (for example Rosen (), Hansen and Hodrick(1981)) economic theory doesn't say much about the orders of AR or MA terms. Thus, we use short order ARMAs to approximate a process which probably is "really" of infinite order (though with small coefficients). Rather than spend a lot of time on "identification" of the precise ARMA process, we tend to throw in a few extra lags just to be sure and leave it at that.

4.3 A fundamental representation

Autocovariances also turn out to be useful because they are the first of three *fundamental* representations for a time series. ARMA processes with normal iid errors are linear combinations of normals, so the resulting $\{x_t\}$ are normally distributed. Thus the joint distribution of an ARMA time series is fully characterized by their mean (0) and covariances $E(x_t x_{t-j})$. (Using the usual formula for a multivariate normal, we can write the joint probability density of $\{x_t\}$ using only the covariances.) In turn, all the statistical properties of a series are described by its joint distribution. Thus, once we know the autocovariances we know *everything* there is to know about the process. Put another way,

> *If two processes have the same autocovariance function, they are the* same *process.*

This was not true of ARMA representations—an AR(1) is the same as a (particular) MA(∞), etc.

This is a useful fact. Here's an example. Suppose x_t is composed of two *unobserved components* as follows:

$$y_t = \nu_t + \alpha\nu_{t-1}; \quad z_t = \delta_t; \quad x_t = y_t + z_t$$

ν_t, δ_t iid, independent of each other. What ARMA process does x_t follow?

One way to solve this problem is to find the autocovariance function of x_t, then find an ARMA process with the same autocovariance function. Since the autocovariance function is fundamental, this must be an ARMA representation for x_t.

$$\text{var}(x_t) = \text{var}(y_t) + \text{var}(z_t) = (1+\alpha^2)\sigma_\nu^2 + \sigma_\delta^2$$

$$E(x_t x_{t-1}) = E[(\nu_t + \delta_t + \alpha\nu_{t-1})(\nu_{t-1} + \delta_{t-1} + \alpha\nu_{t-2})] = \alpha\sigma_\nu^2$$

$$E(x_t x_{t-k}) = 0, \quad k \geq 1.$$

γ_0 and γ_1 nonzero and the rest zero is the autocorrelation function of an MA(1), so we must be able to represent x_t as an MA(1). Using the formula above for the autocorrelation of an MA(1),

$$\gamma_0 = (1+\theta^2)\sigma_\epsilon^2 = (1+\alpha^2)\sigma_\nu^2 + \sigma_\delta^2$$

$$\gamma_1 = \theta\sigma_\epsilon^2 = \alpha\sigma_\nu^2.$$

These are two equations in two unknowns, which we can solve for θ and σ_ϵ^2, the two parameters of the MA(1) representation $x_t = (1+\theta L)\epsilon_t$.

Matching fundamental representations is one of the most common tricks in manipulating time series, and we'll see it again and again.

4.4 Admissible autocorrelation functions

Since the autocorrelation function is fundamental, it might be nice to generate time series processes by picking autocorrelations, rather than specifying (non-fundamental) ARMA parameters. But not every collection of numbers is the autocorrelation of a process. In this section, we answer the question,

when is a set of numbers $\{1, \rho_1, \rho_2, \ldots\}$ the autocorrelation function of an ARMA process?

Obviously, correlation coefficients are less that one in absolute value, so choices like 2 or -4 are ruled out. But it turns out that $\mid \rho_j \mid \leq 1$ though necessary, is not sufficient for $\{\rho_1, \rho_2, \ldots\}$ to be the autocorrelation function of an ARMA process.

The extra condition we must impose is that the variance of any random variable is positive. Thus, it must be the case that

$$\text{var}(\alpha_0 x_t + \alpha_1 x_{t-1} + \ldots) \geq 0 \text{ for all } \{\alpha_0, \alpha_1, \ldots.\}.$$

Now, we can write

$$\text{var}(\alpha_0 x_t + \alpha_1 x_{t-1}) = \gamma_0 [\alpha_0 \ \alpha_1] \begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} \geq 0.$$

Thus, the matrices

$$\begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{bmatrix}$$

etc. must all be positive semi-definite. This is a stronger requirement than $\mid \rho \mid \leq 1$. For example, the determinant of the second matrix must be positive (as well as the determinants of its principal minors, which implies $\mid \rho_1 \mid \leq 1$ and $\mid \rho_2 \mid \leq 1$), so

$$1 + 2\rho_1^2 \rho_2 - 2\rho_1^2 - \rho_2^2 \geq 0 \Rightarrow (\rho_2 - (2\rho_1^2 - 1))(\rho_2 - 1) \leq 0$$

We know $\rho_2 \leq 1$ already so,

$$\rho_2 - (2\rho_1^2 - 1) \geq 0 \ \Rightarrow \ \rho_2 \geq 2\rho_1^2 - 1. \Rightarrow -1 \leq \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \leq 1$$

Thus, ρ_1 and ρ_2 must lie[1] in the parabolic shaped region illustrated in figure 4.1.

[1]To get the last implication,

$$2\rho_1^2 - 1 \leq \rho_2 \leq 1 \Rightarrow -(1 - \rho_1^2) \leq \rho_2 - \rho_1^2 \leq 1 - \rho_1^2 \Rightarrow -1 \leq \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \leq 1.$$

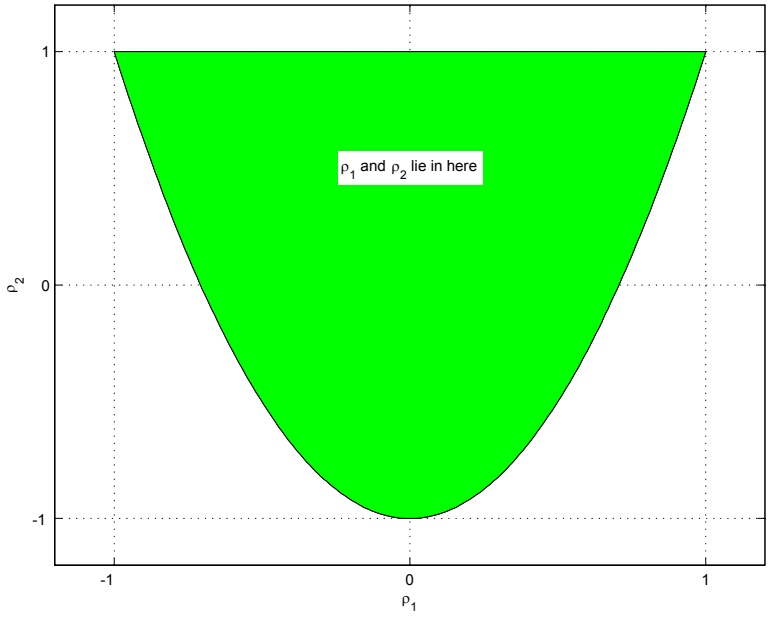

Figure 4.1:

For example, if $\rho_1 = .9$, then $2(.81) - 1 = .62 \leq \rho_2 \leq 1$.

Why go through all this algebra? There are two points: 1) it is *not* true that any choice of autocorrelations with $\mid \rho_j \mid \leq 1$ (or even < 1) is the autocorrelation function of an ARMA process. 2) The restrictions on ρ are very complicated. This gives a reason to want to pay the set-up costs for learning the spectral representation, in which we *can* build a time series by arbitrarily choosing quantities like ρ.

There are two limiting properties of autocorrelations and autocovariances as well. Recall from the Yule-Walker equations that autocorrelations eventually die out exponentially.

> 1) *Autocorrelations are bounded by an exponential.* $\exists\ \lambda > 0\ s.t. |\gamma_j| < \lambda^j$

Since exponentials are square summable, this implies

2) *Autocorrelations are square summable*, $\sum_{j=0}^{\infty} \gamma_j^2 < \infty$.

We will discuss these properties later.

4.5 Multivariate auto- and cross correlations.

As usual, you can remember the multivariate extension by reinterpreting the same letters as appropriate vectors and matrices.

With a vector time series

$$x_t = \begin{bmatrix} y_t \\ z_t \end{bmatrix}$$

we define the autocovariance function as

$$\Gamma_j = E(x_t x'_{t-j}) = \begin{bmatrix} E(y_t y_{t-j}) & E(y_t z_{t-j}) \\ E(z_t y_{t-j}) & E(z_t z_{t-j}) \end{bmatrix}$$

The terms $E(y_t z_{t-j})$ are called *cross-covariances*. Notice that Γ_j does *not* $= \Gamma_{-j}$. Rather, $\Gamma_j = \Gamma'_j$ or $E(x_t x'_{t-j}) = [E(x_t x'_{t+j})]'$. (You should stop and verify this fact from the definition, and the fact that $E(y_t z_{t-j}) = E(z_t y_{t+j})$.)

Correlations are similarly defined as

$$\begin{bmatrix} E(y_t y_{t-j})/\sigma_y^2 & E(y_t z_{t-j})/\sigma_y \sigma_z \\ E(z_t y_{t-j})/\sigma_y \sigma_z & E(z_t z_{t-j})/\sigma_z^2 \end{bmatrix},$$

i.e., we keep track of autocorrelations and cross-correlations.

Chapter 5

Prediction and Impulse-Response Functions

One of the most interesting things to do with an ARMA model is form predictions of the variable given its past. I.e., we want to know what is $E(x_{t+j}$ |all information available at time $t)$. For the moment, "all information" will be all past values of the variable, and all past values of the shocks. I'll get more picky about what is in the information set later. For now, the question is to find

$$E_t(x_{t+j}) \equiv E(x_{t+j} \mid x_t, x_{t-1}, x_{t-2}, \ldots \epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \ldots)$$

We might also want to know how certain we are about the prediction, which we can quantify with

$$var_t(x_{t+j}) \equiv var(x_{t+j} \mid x_t, x_{t-1}, x_{t-2}, \ldots \epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \ldots).$$

The left hand side introduces a short (but increasingly unfashionable) notation for conditional moments.

Prediction is interesting for a variety of reasons. First, it is one of the few rationalizations for time-series to be a subject of its own, divorced from economics. Atheoretical forecasts of time series are often useful. One can simply construct ARMA or VAR models of time series and crank out such forecasts. The pattern of forecasts is also, like the autocorrelation function, an interesting characterization of the behavior of a time series.

5.1 Predicting ARMA models

As usual, we'll do a few examples and then see what general principles underlie them.

AR(1)

For the $AR(1)$, $x_{t+1} = \phi x_t + \epsilon_{t+1}$, we have

$$\begin{array}{rcccc} E_t(x_{t+1}) = & & E_t(\phi x_t + \epsilon_{t+1}) & = & \phi x_t \\ E_t(x_{t+2}) = & & E_t(\phi^2 x_t + \phi\epsilon_{t+1} + \epsilon_{t+2}) & = & \phi^2 x_t \\ E_t(x_{t+k}) = & & & = & \phi^k x_t. \end{array}$$

Similarly,

$$\begin{array}{rcccc} \mathrm{var}_t(x_{t+1}) = & & \mathrm{var}_t(\phi x_t + \epsilon_{t+1}) & = & \sigma_\epsilon^2 \\ \mathrm{var}_t(x_{t+2}) = & & \mathrm{var}(\phi^2 x_t + \ \phi\epsilon_{t+1} + \epsilon_{t+2}) & = & (1+\phi^2)\sigma_\epsilon^2 \\ \mathrm{var}_t(x_{t+k}) = & & ... & = & (1 + \phi^2 + \phi^4 + .. + \phi^{2(k-1)})\sigma_\epsilon^2 \end{array}$$

These means and variances are plotted in figure 5.1.

Notice that

$$\lim_{k\to\infty} E_t(x_{t+k}) = 0 = E(x_t)$$

$$\lim_{k\to\infty} \ \mathrm{var}_t(x_{t+k}) = \sum_{j=0}^{\infty} \phi^{2j}\sigma_\epsilon^2 = \frac{1}{1-\phi^2}\sigma_\epsilon^2 = \mathrm{var}(x_t).$$

The unconditional moments are limits of the conditional moments. In this way, we can think of the unconditional moments as the limit of conditional moments of x_t as of time $t \to -\infty$, or the limit of conditional moments of x_{t+j} as the horizon $j \to \infty$.

MA

Forecasting MA models is similarly easy. Since

$$x_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \ldots,$$

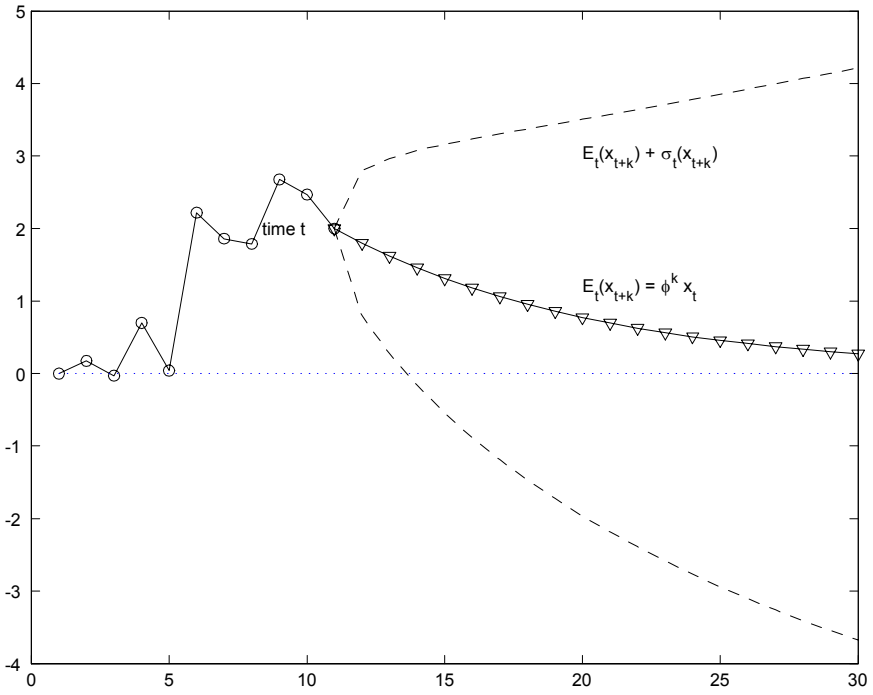

Figure 5.1: AR(1) forecast and standard deviation

we have

$$E_t(x_{t+1}) = E_t(\epsilon_{t+1} + \theta_1\epsilon_t + \theta_2\epsilon_{t-1} + \ldots) = \theta_1\epsilon_t + \theta_2\epsilon_{t-1} + \ldots$$

$$E_t(x_{t+k}) = E_t(\epsilon_{t+k} + \theta_1\epsilon_{t+k-1} + \ldots + \theta_k\epsilon_t + \theta_{k+1}\epsilon_{t-1} + \ldots) = \theta_k\epsilon_t + \theta_{k+1}\epsilon_{t-1} + \ldots$$

$$var_t(x_{t+1}) = \sigma_\epsilon^2$$

$$var_t(x_{t+k}) = (1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_{k-1}^2)\sigma_\epsilon^2$$

AR's and ARMA's

The general principle in cranking out forecasts is to exploit the facts that $E_t(\epsilon_{t+j}) = 0$ and $\text{var}_t(\epsilon_{t+j}) = \sigma_\epsilon^2$ for $j > 0$. You express x_{t+j} as a sum of

things known at time t and shocks between t and $t+j$.

$$x_{t+j} = \{\text{function of } \epsilon_{t+j}, \epsilon_{t+j-1}, ..., \epsilon_{t+1}\} + \{\text{function of } \epsilon_t, \epsilon_{t-1}, ..., x_t, x_{t-1}, ...\}$$

The things known at time t define the conditional mean or forecast and the shocks between t and t+j define the conditional variance or forecast error. Whether you express the part that is known at time t in terms of x's or in terms of ϵ's is a matter of convenience. For example, in the AR(1) case, we could have written $E_t(x_{t+j}) = \phi^j x_t$ or $E_t(x_{t+j}) = \phi^j \epsilon_t + \phi^{j+1}\epsilon_{t-1} + ...$ Since $x_t = \epsilon_t + \phi\epsilon_{t-1} + ...$, the two ways of expressing $E_t(x_{t+j})$ are obviously identical.

It's easiest to express forecasts of AR's and ARMA's analytically (i.e. derive a formula with $E_t(x_{t+j})$ on the left hand side and a closed-form expression on the right) by inverting to their $MA(\infty)$ representations. To find forecasts numerically, it's easier to use the state space representation described later to recursively generate them.

Multivariate ARMAs

Multivariate prediction is again exactly the same idea as univariate prediction, where all the letters are reinterpreted as vectors and matrices. As usual, you have to be a little bit careful about transposes and such.

For example, if we start with a vector $MA(\infty), x_t = B(L)$, we have

$$E_t(x_{t+j}) = B_j\epsilon_t + B_{j+1}\epsilon_{t-1} + \dots$$

$$\text{var}_t(x_{t+j}) = \Sigma + B_1\Sigma B_1' + \dots + B_{j-1}\Sigma B_{j-1}'.$$

5.2 State space representation

The AR(1) is particularly nice for computations because both forecasts and forecast error variances can be found recursively. This section explores a really nice trick by which any process can be mapped into a vector AR(1), which leads to easy programming of forecasts (and lots of other things too.)

5.2.1 ARMAs in vector AR(1) representation

For example, start with an ARMA(2,1)

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \theta_1 \epsilon_{t-1}.$$

We map this into

$$\begin{bmatrix} y_t \\ y_{t-1} \\ \epsilon_t \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \theta_1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \epsilon_{t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} [\epsilon_t]$$

which we write in AR(1) form as

$$x_t = Ax_{t-1} + Cw_t.$$

It is sometimes convenient to redefine the C matrix so the variance-covariance matrix of the shocks is the identity matrix. To to this, we modify the above as

$$C = \begin{bmatrix} \sigma_\epsilon \\ 0 \\ \sigma_\epsilon \end{bmatrix} \quad E(w_t w'_t) = I.$$

5.2.2 Forecasts from vector AR(1) representation

With this vector AR(1) representation, we can find the forecasts, forecast error variances and the impulse response function either directly or with the corresponding vector $MA(\infty)$ representation $x_t = \sum_{j=0}^{\infty} A^j C w_{t-j}$. Either way, forecasts are

$$E_t(x_{t+k}) = A^k x_t$$

and the forecast error variances are[1]

$$x_{t+1} - E_t(x_{t+1}) = Cw_{t+1} \Rightarrow \text{var}_t(x_{t+1}) = CC'$$

$$x_{t+2} - E_t(x_{t+2}) = Cw_{t+2} + ACw_{t+1} \Rightarrow \text{var}_t(x_{t+2}) = CC' + ACC'A'$$

[1]In case you forgot, if x is a vector with covariance matrix Σ and A is a matrix, then $var(Ax) = A\Sigma A'$.

$$\text{var}_t(x_{t+k}) = \sum_{j=0}^{k-1} A^j CC' A^{j'}$$

These formulas are particularly nice, because they can be computed *recursively,*

$$E_t(x_{t+k}) = AE_t(x_{t+k-1})$$
$$\text{var}_t(x_{t+k}) = CC' + A\text{var}_t(x_{t+k-1})A'.$$

Thus, you can program up a string of forecasts in a simple do loop.

5.2.3 VARs in vector AR(1) representation.

The multivariate forecast formulas given above probably didn't look very appetizing. The easy way to do computations with VARs is to map them into a vector AR(1) as well. Conceptually, this is simple—just interpret x_t above as a vector $[y_t \ z_t]'$. Here is a concrete example. Start with the prototype VAR,

$$y_t = \phi_{yy1} y_{t-1} + \phi_{yy2} y_{t-2} + \ldots + \phi_{yz1} z_{t-1} + \phi_{yz2} z_{t-2} + \ldots + \epsilon_{yt}$$

$$z_t = \phi_{zy1} y_{t-1} + \phi_{zy2} y_{t-2} + \ldots + \phi_{zz1} z_{t-1} + \phi_{zz2} z_{t-2} + \ldots + \epsilon_{zt}$$

We map this into an AR(1) as follows.

$$\begin{bmatrix} y_t \\ z_t \\ y_{t-1} \\ z_{t-1} \\ \vdots \end{bmatrix} = \begin{bmatrix} \phi_{yy1} & \phi_{yz1} & \phi_{yy2} & \phi_{yz2} & \\ \phi_{zy1} & \phi_{zz1} & \phi_{zy2} & \phi_{zz2} & \\ 1 & 0 & 0 & 0 & \ldots \\ 0 & 1 & 0 & 0 & \\ & & \ldots & & \ddots \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \\ y_{t-2} \\ z_{t-2} \\ \vdots \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} \epsilon_{yt} \\ \epsilon_{zt} \end{bmatrix}$$

i.e.,

$$x_t = Ax_{t-1} + \epsilon_t, \ E(\epsilon_t \epsilon_t') = \Sigma,$$

Or, starting with the vector form of the VAR,

$$x_t = \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + ... + \epsilon_t,$$

$$\begin{bmatrix} x_t \\ x_{t-1} \\ x_{t-2} \\ \vdots \end{bmatrix} = \begin{bmatrix} \Phi_1 & \Phi_2 & \dots \\ I & 0 & \dots \\ 0 & I & \dots \\ \dots & \dots & \ddots \end{bmatrix} \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ x_{t-3} \\ \vdots \end{bmatrix} + \begin{bmatrix} I \\ 0 \\ 0 \\ \vdots \end{bmatrix} [\epsilon_t]$$

Given this AR(1) representation, we can forecast both y and z as above. Below, we add a small refinement by choosing the C matrix so that the shocks are orthogonal, $E(\epsilon\epsilon') = I$.

Mapping a process into a vector AR(1) is a very convenient trick, for other calculations as well as forecasting. For example, Campbell and Shiller (199x) study present values, i.e. $E_t(\sum_{j=1}^{\infty} \lambda^j x_{t+j})$ where x = dividends, and hence the present value should be the price. To compute such present values from a VAR with x_t as its first element, they map the VAR into a vector AR(1). Then, the computation is easy: $E_t(\sum_{j=1}^{\infty} \lambda^j x_{t+j}) = (\sum_{j=1}^{\infty} \lambda^j A^j) x_t = (I - \lambda A)^{-1} x_t$. Hansen and Sargent (1992) show how an unbelievable variety of models beyond the simple ARMA and VAR I study here can be mapped into the vector AR(1).

5.3 Impulse-response function

The impulse response function is the path that x follows if it is kicked by a single unit shock ϵ_t, i.e., $\epsilon_{t-j} = 0, \epsilon_t = 1, \epsilon_{t+j} = 0$. This function is interesting for several reasons. First, it is another characterization of the behavior of our models. Second, and more importantly, it allows us to start thinking about "causes" and "effects". For example, you might compute the response of GNP to a shock to money in a GNP-M1 VAR and interpret the result as the "effect" on GNP of monetary policy. I will study the cautions on this interpretation extensively, but it's clear that it's interesting to learn how to calculate the impulse-response.

For an $AR(1)$, recall the model is $x_t = \phi x_{t-1} + \epsilon_t$ or $x_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$. Looking at the MA($\infty$) representation, we see that the impulse-response is

ϵ_t :	0	0	1	0	0	0	0
x_t :	0	0	1	ϕ	ϕ^2	ϕ^3	...

Similarly, for an $MA(\infty)$, $x_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}$,

$$\begin{array}{lccccccc} \epsilon_t: & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ x_t: & 0 & 0 & 1 & \theta & \theta_2 & \theta_3 & ... \end{array}.$$

As usual, vector processes work the same way. If we write a vector $MA(\infty)$ representation as $x_t = B(L)\epsilon_t$, where $\epsilon_t \equiv [\epsilon_{yt} \ \epsilon_{zt}]'$ and $B(L) \equiv B_0 + B_1 L + ...$, then $\{B_0, B_1, ...\}$ define the impulse-response function. Precisely, $B(L)$ means

$$B(L) = \begin{bmatrix} b_{yy}(L) & b_{yz}(L) \\ b_{zy}(L) & b_{zz}(L) \end{bmatrix},$$

so $b_{yy}(L)$ gives the response of y_{t+k} to a unit y shock ϵ_{yt}, $b_{yz}(L)$ gives the response of y_{t+k} to a unit z shock, etc.

As with forecasts, $MA(\infty)$ representations are convenient for studying impulse-responses analytically, but mapping to a vector $AR(1)$ representation gives the most convenient way to calculate them in practice. Impulse-response functions for a vector $AR(1)$ look just like the scalar $AR(1)$ given above: for

$$x_t = Ax_{t-1} + C\epsilon_t,$$

the response function is

$$C, AC, A^2C, ..., A^kC, ..$$

Again, this can be calculated recursively: just keep multiplying by A. (If you want the response of y_t, and not the whole state vector, remember to multiply by $[1\ 0\ 0\ \dots\ ]'$ to pull off y_t, the first element of the state vector.)

While this looks like the same kind of trivial response as the $AR(1)$, remember that A and C are matrices, so this simple formula can capture the complicated dynamics of *any* finite order ARMA model. For example, an $AR(2)$ can have an impulse response with decaying sine waves.

5.3.1 Facts about impulse-responses

Three important properties of impulse-responses follow from these examples:

1. The $MA(\infty)$ representation is the *same thing* as the impulse-response function.

This fact is very useful. To wit:

2. The easiest way to calculate an $MA(\infty)$ representation is to simulate the impulse-response function.

Intuitively, one would think that impulse-responses have something to do with forecasts. The two are related by:

3. The impulse response function is the same as $E_t(x_{t+j}) - E_{t-1}(x_{t+j})$.

Since the ARMA models are linear, the response to a unit shock if the value of the series is zero is the same as the response to a unit shock on top of whatever other shocks have hit the system. This property is *not* true of nonlinear models!

Chapter 6

Stationarity and Wold representation

6.1 Definitions

In calculating the moments of ARMA processes, I used the fact that the moments do not depend on the calendar time:

$$E(x_t) = E(x_s) \text{ for all } t \text{ and } s$$

$$E(x_t x_{t-j}) = E(x_s x_{s-j}) \text{ for all } t \text{ and } s.$$

These properties are true for the invertible ARMA models, as you can show directly. But they reflect a much more important and general property, as we'll see shortly. Let's define it:

Definitions:

A process $\{x_t\}$ is *strongly stationary* or *strictly stationary* if the joint probability distribution function of $\{x_{t-s}, .., x_t, \ldots x_{t+s}\}$ is independent of t for all s.

A process x_t is *weakly stationary* or *covariance stationary* if $E(x_t), E(x_t^2)$ are finite and $E(x_t x_{t-j})$ depends only on j and not on t.

Note that

1. Strong stationarity does *not* $\Rightarrow$ weak stationarity. $E(x_t^2)$ must be finite. For example, an iid Cauchy process is strongly, but not covariance, stationary.

2. Strong stationarity *plus* $E(x_t), E(x_t x) < \infty \Rightarrow$ weak stationarity

3. Weak stationarity does *not* $\Rightarrow$ strong stationarity. If the process is not normal, other moments ($E(x_t x_{t-j} x_{t-k})$) *might* depend on t, so the process might not be strongly stationary.

4. Weak stationarity *plus* normality $\Rightarrow$ strong stationarity.

Strong stationarity is useful for proving some theorems. For example, a nonlinear (measurable) function of a strongly stationary variable is strongly stationary; this is not true of covariance stationarity. For most purposes, weak or covariance stationarity is enough, and that's the concept I will use through the rest of these notes.

Stationarity is often misunderstood. For example, if the conditional covariances of a series vary over time, as in ARCH models, the series can still be stationary. The definition merely requires that the unconditional covariances are not a function of time. Many people use "nonstationary" interchangeably with "has a unit root". That is one form of nonstationarity, but there are lots of others. Many people say a series is "nonstationary" if it has breaks in trends, or if one thinks that the time-series process changed over time. If we think that the trend-break or structural shift occurs at one point in time, no matter how history comes out, they are right. However, if a series is subject to occasional stochastic trend breaks or shifts in structure, then the unconditional covariances will no longer have a time index, and the series can be stationary.

6.2 Conditions for stationary ARMA's

Which ARMA processes are stationary? First consider the MA processes $x_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}$. Recalling the formula for the variance, $var(x_t) = \sum_{j=0}^{\infty} \theta_j^2 \sigma_\epsilon^2$,

we see that *Second moments exist if and only if the MA coefficients are square summable,*

$$\text{Stationary MA} \Leftrightarrow \sum_{j=0}^{\infty} \theta_j^2 < \infty.$$

If second moments exist, it's easy to see that they are independent of the t index.

Consider the $AR(1)$. Inverting it to an MA, we get

$$x_t = \sum_{j=0}^{k} \phi^j \epsilon_{t-j} + \phi^k x_{t-k}.$$

Clearly, unless $\mid \phi \mid < 1$, we are not going to get square summable MA coefficients, and hence the variance of x will again not be finite.

More generally, consider factoring an AR

$$A(L)x_t = \epsilon_t = (1 - \lambda_1 L)(1 - \lambda_2 L) \ldots x_t = \epsilon_t.$$

For the variance to be finite, *The AR lag polynomial must be invertible* or $\mid \lambda_i \mid < 1$ for all i. A more common way of saying this is to factor A(L) somewhat differently,

$$A(L) = constant(L - \zeta_1)(L - \zeta_2) \ldots .$$

the ζ_i are the *roots of the lag polynomial*, since A(z) = 0 when $z = \zeta_i$. We can rewrite the last equation as

$$A(L) = constant(-\zeta_1)(1 - \frac{1}{\zeta_1} L)(-\zeta_2)(1 - \frac{1}{\zeta_2} L) \ldots$$

Thus the roots ζ and the λs must be related by

$$\zeta_i = \frac{1}{\lambda_i}.$$

Hence, the rule "all $\mid \lambda \mid < 1$" means "all $\mid \zeta \mid > 1$", or since λ and ζ can be complex,

> AR's are stationary if all roots of the lag polynomial lie outside the unit circle, i.e. if the lag polynomial is invertible.

Both statements of the requirement for stationarity are equvalent to

> ARMAs are stationary if and only if the impluse-response function eventually decays exponentially.

Stationarity does not require the MA polynomial to be invertible. That means something else, described next.

6.3 Wold Decomposition theorem

The above definitions are important because they define the range of "sensible" ARMA processes (invertible AR lag polynomials, square summable MA lag polynomials). Much more importantly, they are useful to enlarge our discussion past ad-hoc linear combinations of iid Gaussian errors, as assumed so far. Imagine *any* stationary time series, for example a non-linear combination of serially correlated lognormal errors. It turns out that, so long as the time series is covariance stationary, it has a linear ARMA representation! Therefore, the ad-hoc ARMA models we have studied so far turn out to be a much more general class than you might have thought. This is an enormously important fact known as the

> **Wold Decomposition Theorem:** Any mean zero covariance stationary process $\{x_t\}$ can be represented in the form
>
> $$x_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j} + \eta_t$$
>
> where

1. $\epsilon_t \equiv x_t - P(x_t \mid x_{t-1}, x_{t-2}, \ldots..)$.
2. $P(\epsilon_t | x_{t-1}, x_{t-2}, \ldots.) = 0,\ E(\epsilon_t x_{t-j}) = 0,\ E(\epsilon_t) = 0,\ E(\epsilon_t^2) = \sigma_\epsilon^2$ (same for all t), $E(\epsilon_t \epsilon_s) = 0$ for all $t \neq s$,
3. All the roots of $\theta(L)$ are on or outside the unit circle, i.e. (unless there is a unit root) the MA polynomial is *invertible*.

4. $\sum_{j=0}^{\infty} \theta_j^2 < \infty,\ \theta_0 = 1$

5. $\{\theta_j\}$ and $\{\epsilon_s\}$ are unique.

6. η_t is linearly deterministic, i.e.$\eta_t = P(\eta_t \mid x_{t-1}, \ldots)$.

Property 1) just defines ϵ_t as the linear forecast errors of x_t. $P(\circ|\circ)$ denotes projection, i.e. the fitted value of a linear regression of the left hand variable on the right hand variables. Thus, if we start with such a regression, say $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \epsilon_t$, 1) merely solves this regression for ϵ_t. Properties 2) result from the definition of ϵ_t as regression errors, and the fact from 1) that we can recover the ϵ_t from current and lagged x's, so ϵ_t is orthogonal to lagged ϵ as well as lagged x. Property 3) results from the fact that we can recover ϵ_t from current and lagged x. If $\theta(L)$ was not invertible, then we couldn't solve for ϵ_t in terms of current and lagged x.Property 4) comes from stationarity. If the θ were not square summable, the variance would be infinite. Suppose we start with an AR(1) plus a sine wave. The resulting series is covariance stationary. Property 6) allows for things like sine waves in the series. We usually specify that the process $\{x_t\}$ is *linearly regular*, which just means that deterministic components η_t have been removed.

Sargent (1987), Hansen and Sargent (1991) and Anderson (1972) all contain proofs of the theorem. The proofs are very interesting, and introduce you to the Hilbert space machinery, which is the hot way to think about time series. The proof is not that mysterious. All the theorem says is that x_t can be written as a sum of its forecast errors. If there was a time zero with information I_0 and $P(x_j \mid I_0) = 0$, this would be obvious:

$$x_1 = x_1 - P(x_1 \mid I_0) = \epsilon_1$$

$$x_2 = x_2 - P(x_2 \mid x_1, I_0) + P(x_2 \mid x_1, I_0) = \epsilon_2 + \theta_1 x_1 = \epsilon_2 + \theta_1 x_1 = \epsilon_2 + \theta_1 \epsilon_1$$

$$x_3 = x_3 - P(x_3 \mid x_2, x_1, I_0) + P(x_3 \mid x_2, x_1, I_0) = \epsilon_3 + \phi_1 x_2 + \phi_2 x_1$$

$$= \epsilon_3 + \phi_1(\epsilon_2 + \theta_1 x_1) + \phi_2 x_1 = \epsilon_3 + \phi_1 \epsilon_2 + (\phi_1 \theta_1)\epsilon_1$$

and so forth. You can see how we are getting each x as a linear function of past linear prediction errors. We could do this even for nonstationary x; the stationarity of x means that the coefficients on lagged ϵ are independent of the time index.

6.3.1 What the Wold theorem does not say

Here are a few things the Wold theorem does *not* say:

1) The ϵ_t need *not* be normally distributed, and hence need *not* be iid.

2) Though $P(\epsilon_t \mid x_{t-j}) = 0$, it need not be true that $E(\epsilon_t \mid x_{t-j}) = 0$. The projection operator $P(x_t \mid x_{t-1}, \ldots)$ finds the best guess of x_t (minimum squared error loss) from *linear* combinations of past x_t, i.e. it fits a linear regression. The conditional expectation operator $E(x_t \mid x_{t-1}, \ldots)$ is equivalent to finding the best guess of x_t using linear *and all nonlinear* combinations of past x_t, i.e., it fits a regression using all linear *and nonlinear* transformations of the right hand variables. Obviously, the two are different.

3) The shocks ϵ need not be the "true" shocks to the system. If the true x_t is not generated by linear combinations of past x_t plus a shock, then the Wold shocks will be different from the true shocks.

4) Similarly, the Wold $MA(\infty)$ is *a* representation of the time series, one that fully captures its second moment properties, but not *the* representation of the time series. Other representations may capture deeper properties of the system. The uniqueness result only states that the Wold representation is the unique *linear* representation where the shocks are *linear* forecast errors. Non-linear representations, or representations in terms of non-forecast error shocks are perfectly possible.

Here are some examples:

A) *Nonlinear dynamics.* The true system may be generated by a nonlinear difference equation $x_{t+1} = g(x_t, x_{t-1}, \ldots) + \eta_{t+1}$. Obviously, when we fit a linear approximation as in the Wold theorem, $x_t = P(x_t \mid x_{t-1}, x_{t-2}, \ldots) + \epsilon_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots \epsilon_t$, we will find that $\epsilon_t \neq \eta_t$. As an extreme example, consider the random number generator in your computer. This is a *deterministic* nonlinear system, $\eta_t = 0$. Yet, if you fit arbitrarily long AR's to it, you will get errors! This is another example in which $E(.)$ and $P(.)$ are not the same thing.

B) *Non-invertible shocks.* Suppose the true system is generated by

$$x_t = \eta_t + 2\eta_{t-1}. \ \eta_t \text{ iid}, \sigma_\eta^2 = 1$$

This is a stationary process. But the MA lag polynomial is not invertible

(we can't express the η shocks as x forecast errors), so it can't be the Wold representation. To find the Wold representation of the same process, match autocorrelation functions to a process $x_t = \epsilon_t + \theta\epsilon_{t-1}$:

$$E(x_t^2) = (1+4) = 5 = (1+\theta^2)\sigma_\epsilon^2$$

$$E(x_t x_{t-1}) = 2 = \theta\sigma_\epsilon^2$$

Solving,

$$\frac{\theta}{1+\theta^2} = \frac{2}{5} \Rightarrow \theta = \{2 \text{ or } 1/2\}$$

and

$$\sigma_\epsilon^2 = 2/\theta = \{1 \text{ or } 4\}$$

The original model $\theta = 2, \sigma_\eta^2 = 1$ is one possibility. But $\theta = 1/2, \sigma_\epsilon^2 = 4$ works as well, and that root is invertible. The Wold representation is *unique*: if you've found one invertible MA, it must be *the* Wold representation.

Note that the impulse-response function of the original model is 1, 2; while the impulse response function of the Wold representation is 1, 1/2. Thus, the Wold representation, which is what you would recover from a VAR does *not* give the "true" impulse-response.

Also, the Wold errors ϵ_t are recoverable from a linear function of current and pas x_t. $\epsilon_t = \sum_{j=0}^{\infty}(-.5)^j x_{t-j}$The true shocks are not. In this example, the true shocks are linear functions of future x_t: $\eta_t = \sum_{j=1}^{\infty}(-.5)^j x_{t+j}$!

This example holds more generally: *any* $MA(\infty)$ can be reexpressed as an *invertible* $MA(\infty)$.

6.4 The Wold MA(∞) as another fundamental representation

One of the lines of the Wold theorem stated that the Wold $MA(\infty)$ representation was *unique*. This is a convenient fact, because it means that the $MA(\infty)$ representation in terms of linear forecast errors (with the autocorrelation function and spectral density) is another *fundamental* representation. If two time series have the same Wold representation, they are *the same time series* (up to second moments/linear forecasting).

This is the same property that we found for the autocorrelation function, and can be used in the same way.

Chapter 7

VARs: orthogonalization, variance decomposition, Granger causality

7.1 Orthogonalizing VARs

The impulse-response function of a VAR is slightly ambiguous. As we will see, you can represent a time series with arbitrary linear combinations of any set of impulse responses. "Orthogonalization" refers to the process of selecting one of the many possible impulse-response functions that you find most interesting to look at. It is also technically convenient to transform VARs to systems with orthogonal error terms.

7.1.1 Ambiguity of impulse-response functions

Start with a VAR expressed in vector notation, as would be recovered from regressions of the elements of x_t on their lags:

$$A(L)x_t = \epsilon_t,\ A(0) = I,\ E(\epsilon_t \epsilon_t') = \Sigma. \tag{7.1}$$

Or, in moving average notation,

$$x_t = B(L)\epsilon_t,\ B(0) = I,\ E(\epsilon_t \epsilon_t') = \Sigma \tag{7.2}$$

where $B(L) = A(L)^{-1}$. Recall that $B(L)$ gives us the response of x_t to unit impulses to each of the elements of ϵ_t. Since $A(0) = I$, $B(0) = I$ as well.

But we could calculate instead the responses of x_t to new shocks that are *linear combinations* of the old shocks. For example, we could ask for the response of x_t to unit movements in ϵ_{yt} and $\epsilon_{zt} + .5\epsilon_{yt}$. (Just *why* you might want to do this might not be clear at this point, but bear with me.) This is easy to do. Call the new shocks η_t so that $\eta_{1t} = \epsilon_{yt},\ \eta_{2t} = \epsilon_{zt} + .5\epsilon_{yt}$, or

$$\eta_t = Q\epsilon_t,\ Q = \begin{bmatrix} 1 & 0 \\ .5 & 1 \end{bmatrix}.$$

We can write the moving average representation of our VAR in terms of these new shocks as $x_t = B(L)Q^{-1}Q\epsilon_t$ or

$$x_t = C(L)\eta_t. \tag{7.3}$$

where $C(L) = B(L)Q^{-1}$. $C(L)$ gives the response of x_t to the new shocks η_t. As an equivalent way to look at the operation, you can see that $C(L)$ is a linear combination of the original impulse-responses $B(L)$.

So which linear combinations should we look at? Clearly the data are no help here—the representations (7.2) and (7.3) are observationally equivalent, since they produce the same series x_t. We have to decide which linear combinations we think are the most *interesting.* To do this, we state a set of assumptions, called *orthogonalization assumptions,* that uniquely pin down the linear combination of shocks (or impulse-response functions) that we find most interesting.

7.1.2 Orthogonal shocks

The first, and almost universal, assumption is that *the shocks should be orthogonal* (uncorrelated). If the two shocks ϵ_{yt} and ϵ_{zt} are correlated, it doesn't make much sense to ask "what if ϵ_{yt} has a unit impulse" with no change in ϵ_{zt}, since the two usually come at the same time. More precisely, we would like to start thinking about the impulse-response function in causal terms—the "effect" of money on GNP, for example. But if the money shock is correlated with the GNP shock, you don't know if the response you're seeing is the response of GNP to money, or (say) to a technology shocks that happen to

come at the same time as the money shock (maybe because the Fed sees the GNP shock and accommodates it). Additionally, it is convenient to rescale the shocks so that they have a unit variance.

Thus, we want to pick Q so that $E(\eta_t \eta_t') = I$. To do this, we need a Q such that

$$Q^{-1}Q^{-1\prime} = \Sigma$$

With that choice of Q,

$$E(\eta_t \eta_t') = E(Q\epsilon_t \epsilon_t' Q') = Q\Sigma Q' = I$$

One way to construct such a Q is via the *Choleski decomposition*. (Gauss has a command CHOLESKI that produces this decomposition.)

Unfortunately there are many different Q's that act as "square root" matrices for Σ. (Given one such Q, you can form another, Q^*, by $Q^* = RQ$, where R is an orthogonal matrix, $RR' = I$. $Q^*Q^{*\prime} = RQQ'R' = RR' = I$.) Which of the many possible Q's should we choose?

We have exhausted our possibilities of playing with the error term, so we now specify desired properties of the moving average $C(L)$ instead. Since $C(L) = B(L)Q^{-1}$, specifying a desired property of $C(L)$ can help us pin down Q. To date, using "theory" (in a very loose sense of the word) to specify features of $C(0)$ and $C(1)$ have been the most popular such assumptions. Maybe you can find other interesting properties of $C(L)$ to specify.

7.1.3 Sims orthogonalization—Specifying $C(0)$

Sims (1980) suggests we specify properties of $C(0)$, which gives the instantaneous response of each variable to each orthogonalized shock η. In our original system, (7.2) $B(0) = I$. This means that each shock only affects its own variable contemporaneously. Equivalently, $A(0) = I$—in the autoregressive representation (7.1), neither variable appears contemporaneously in the other variable's regression.

Unless Σ is diagonal (orthogonal shocks to start with), every diagonalizing matrix Q will have off-diagonal elements. Thus, $C(0)$ *cannot* $= I$. This means that some shocks *will* have effects on more than one variable. Our job is to specify this pattern.

Sims suggests that we choose a *lower triangular* $C(0)$,

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} C_{0yy} & 0 \\ C_{0zy} & C_{0zz} \end{bmatrix} \begin{bmatrix} \eta_{1t} \\ \eta_{2t} \end{bmatrix} + C_1\eta_{t-1} + ...$$

As you can see, this choice means that the second shock η_{2t} does not affect the first variable, y_t, contemporaneously. Both shocks can affect z_t contemporaneously. Thus, all the contemporaneous correlation between the original shocks ϵ_t has been folded into C_{0zy}.

We can also understand the orthogonalization assumption in terms of the implied autoregressive representation. In the original VAR, $A(0) = I$, so contemporaneous values of each variable do not appear in the other variable's equation. A lower triangular $C(0)$ implies that contemporaneous y_t appears in the z_t equation, but z_t does not appear in the y_t equation. To see this, call the orthogonalized autoregressive representation $D(L)x_t = \eta_t$, i.e., $D(L) = C(L)^{-1}$. Since the inverse of a lower triangular matrix is also lower triangular, $D(0)$ is lower triangular, i.e.

$$\begin{bmatrix} D_{0yy} & 0 \\ D_{0zy} & D_{0zz} \end{bmatrix} \begin{bmatrix} y_t \\ z_t \end{bmatrix} + D_1x_{t-1} + ... = \eta_t$$

or

$$\begin{array}{lllll} D_{0yy}y_t = & & -D_{1yy}y_{t-1} - D_{1yz}z_{t-1} & +\eta_{1t} \\ D_{0zz}z_t = & -D_{0zy}y_t & -D_{1zy}y_{t-1} - D_{1zz}z_{t-1} & +\eta_{2t} \end{array}. \tag{7.4}$$

As another way to understand Sims orthogonalization, note that it is *numerically* equivalent to estimating the system by OLS with contemporaneous y_t in the z_t equation, but not vice versa, and then scaling each equation so that the error variance is one. To see this point, remember that OLS estimates produce residuals that are uncorrelated with the right hand variables by construction (this is their defining property). Thus, suppose we run OLS on

$$\begin{array}{lllll} y_t = & & a_{1yy}y_{t-1} + .. + a_{1yz}z_{t-1} + .. + \eta_{yt} \\ z_t = & a_{0zy}y_t + & a_{1zy}y_{t-1} + .. + a_{1zz}z_{t-1} + .. + \eta_{zt} \end{array} \tag{7.5}$$

The first OLS residual is defined by $\eta_{yt} = y_t - E(y_t \mid y_{t-1}, .., z_{t-1}, ..)$ so η_{yt} is a linear combination of $\{y_t, y_{t-1}, .., z_{t-1}, ..\}$.OLS residuals are orthogonal to right hand variables, so η_{zt} is orthogonal to any linear combination of $\{y_t, y_{t-1}, .., z_{t-1}, ..\}$, by construction. Hence, η_{yt} and η_{zt} are uncorrelated

with each other. a_{0zy} captures all of the contemporaneous correlation of news in y_t and news in z_t.

In summary, one can uniquely specify Q and hence which linear combination of the original shocks ϵ you will use to plot impulse-responses by the requirements that 1) the errors are orthogonal and 2) the instantaneous response of one variable to the other shock is zero. Assumption 2) is equivalent to 3) The VAR is estimated by OLS with contemporaneous y in the z equation but not vice versa.

Happily, the Choleski decomposition produces a lower triangular Q. Since

$$C(0) = B(0)Q^{-1} = Q^{-1},$$

the Choleski decomposition produces the Sims orthogonalization already, so you don't have to do any more work. (You do have to decide what order to put the variables in the VAR.)

Ideally, one tries to use economic theory to decide on the order of orthogonalization. For example, (reference) specifies that the Fed cannot see GNP until the end of the quarter, so money cannot respond within the quarter to a GNP shock. As another example, Cochrane (1993) specifies that the instantaneous response of consumption to a GNP shock is zero, in order to identify a movement in GNP that consumers regard as transitory.

7.1.4 Blanchard-Quah orthogonalization–restrictions on $C(1)$.

Rather than restrict the *immediate* response of one variable to another shock, Blanchard and Quah (1988) suggest that it is interesting to examine shocks defined so that the *long-run* response of one variable to another shock is zero. If a system is specified in changes, $\Delta x_t = C(L)\eta_t$, then $C(1)$ gives the long-run response of the levels of x_t to η shocks. Blanchard and Quah argued that "demand" shocks have no long-run effect on GNP. Thus, they require $C(1)$ to be lower diagonal in a VAR with GNP in the first equation. We find the required orthogonalizing matrix Q from $C(1) = B(1)Q^{-1}$.

7.2 Variance decompositions

In the orthogonalized system we can compute an accounting of forecast error variance: what percent of the k step ahead forecast error variance is due to which variable. To do this, start with the moving average representation of an orthogonalized VAR

$$x_t = C(L)\eta_t,\ E(\eta_t\eta_t') = I.$$

The one step ahead forecast error variance is

$$\epsilon_{t+1} = x_{t+1} - E_t(x_{t+1}) = C_0\eta_{t+1} = \begin{bmatrix} c_{yy,0} & c_{yz,0} \\ c_{zy,0} & c_{zz,0} \end{bmatrix} \begin{bmatrix} \eta_{y,t+1} \\ \eta_{z,t+1} \end{bmatrix}.$$

(In the right hand equality, I denote $C(L) = C_0 + C_1L + C_2L^2 + ...$ and the elements of $C(L)$ as $c_{yy,0} + c_{yy,1}L + c_{yy,2}L^2 + ...$, etc.) Thus, since the η are uncorrelated and have unit variance,

$$\text{var}_t(y_{t+1}) = c_{yy,0}^2\sigma^2(\eta_y) + c_{yz,0}^2\sigma^2(\eta_z) = c_{yy,0}^2 + c_{yz,0}^2$$

and similarly for z. Thus, $c_{yy,0}^2$ gives the amount of the one-step ahead forecast error variance of y due to the η_y shock, and $c_{yz,0}^2$ gives the amount due to the η_z shock. (Actually, one usually reports fractions $c_{yy,0}^2/(c_{yy,0}^2 + c_{yz,0}^2)$.)

More formally, we can write

$$\text{var}_t(x_{t+1}) = C_0C_0'.$$

Define

$$I_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},\ I_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},\ \text{etc.}$$

Then, the part of the one step ahead forecast error variance due to the first (y) shock is $C_0I_1C_0'$, the part due to the second (z) shock is $C_0I_2C_0'$, etc. Check for yourself that these parts add up, i.e. that

$$C_0C_0' = C_0\ I_1C_0' + C_0\ I_2C_0' + \ldots$$

You can think of I_τ as a new covariance matrix in which all shocks but the τth are turned off. Then, the total variance of forecast errors must be equal to the part due to the τth shock, and is obviously $C_0I_\tau C_0'$.

Generalizing to k steps ahead is easy.

$$x_{t+k} - E_t(x_{t+k}) = C_0\eta_{t+k} + C_1\eta_{t+k-1} + \ldots + C_{k-1}\eta_{t+1}$$

$$\text{var}_t(x_{t+k}) = C_0C_0' + C_1C_1' + \ldots + C_{k-1}C_{k-1}'$$

Then

$$v_{k,\tau} = \sum_{j=0}^{k-1} C_j I_\tau C_j'$$

is the variance of k step ahead forecast errors due to the τth shock, and the variance is the sum of these components, $\text{var}_t(x_{t+k}) = \sum_\tau v_{k,\tau}$.

It is also interesting to compute the decomposition of the actual variance of the series. Either directly from the MA representation, or by recognizing the variance as the limit of the variance of k-step ahead forecasts, we obtain that the contribution of the τth shock to the variance of x_t is given by

$$v_\tau = \sum_{j=0}^{\infty} C_j I_\tau C_j'$$

and $\text{var}(x_{t+k}) = \sum_\tau v_\tau$.

7.3 VAR's in state space notation

For many of these calculations, it's easier to express the VAR as an $AR(1)$ in state space form. The only refinement relative to our previous mapping is how to include orthogonalized shocks η. Since $\eta_t = Q\epsilon_t$, we simply write the VAR

$$x_t = \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + ... + \epsilon_t$$

as

$$\begin{bmatrix} x_t \\ x_{t-1} \\ x_{t-2} \\ \vdots \end{bmatrix} = \begin{bmatrix} \Phi_1 & \Phi_2 & \ldots \\ I & 0 & \ldots \\ 0 & I & \ldots \\ \ldots & & \ddots \end{bmatrix} \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ x_{t-3} \\ \vdots \end{bmatrix} + \begin{bmatrix} Q^{-1} \\ 0 \\ 0 \\ \vdots \end{bmatrix} [\eta_t]$$

$$x_t = Ax_{t-1} + C\eta_t, \quad E(\eta_t\eta_t') = I$$

The impulse-response function is $C,\ AC,\ A^2C, ...$ and can be found recursively from

$$IR_0 = C,\ IR_j = A\, IR_{j-1}\,.$$

If Q^{-1} is lower diagonal, then only the first shock affects the first variable, as before. Recall from forecasting AR(1)'s that

$$\mathrm{var}_t(x_{t+j}) = \sum_{j=0}^{k-1} A^j CC' A'^{\,j}.$$

Therefore,

$$v_{k,\tau} = \sum_{j=0}^{k-1} A^j C\, I_\tau\, C' A'^{\,j}$$

gives the variance decomposition—the contribution of the τth shock to the k-step ahead forecast error variance. It too can be found recursively from

$$v_{i,\tau} = CI_\tau C',\ \ v_{k,t} = Av_{k-1,t}A'.$$

7.4 Tricks and problems:

1. Suppose you multiply the original VAR by an arbitrary lower triangular Q. This produces a system of the same form as (7.4). Why would OLS (7.5) not recover this system, instead of the system formed by multiplying the original VAR by the inverse of the Choleski decomposition of Σ?

2.Suppose you start with a given orthogonal representation,

$$x_t = C(L)\eta_t,\ E(\eta_t\eta_t') = I.$$

Show that you can transform to other orthogonal representations of the shocks by an orthogonal matrix—a matrix Q such that $QQ' = I$.

3. Consider a two-variable cointegrated VAR. y and c are the variables, $(1-L)y_t$ and $(1-L)c_t$, and $y_t - c_t$ are stationary, and c_t is a random walk. Show that in this system, Blanchard-Quah and Sims orthogonalization produce the same result.

4. Show that the Sims orthogonalization is equivalent to requiring that the one-step ahead forecast error variance of the first variable is all due to the first shock, and so forth.

Answers:

1. The OLS regressions of (7.5) do not (necessarily) produce a diagonal covariance matrix, and so are *not* the same as OLS estimates, even though the same number of variables are on the right hand side. Moral: watch the properties of the error terms as well as the properties of $C(0)$ or $B(0)$!

2. We want to transform to shocks ξ_t, such that $E(\xi_t \xi_t') = I$. To do it, $E(\xi_t \xi_t') = E(Q\eta_t \eta_t' Q') = QQ'$, which had better be I. Orthogonal matrices rotate vectors without stretching or shrinking them. For example, you can verify that

$$Q = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

rotates vectors counterclockwise by θ. This requirement means that the columns of Q must be orthogonal, and that if you multiply Q by two orthogonal vectors, the new vectors will still be orthogonal. Hence the name.

3. Write the $y,\ c$ system as $\Delta x_t = B(L)\epsilon_t$. $y,\ c$ cointegrated implies that c and y have the same long-run response to any shock—$B_{cc}(1) = B_{yc}(1),\ B_{cy}(1) = B_{yy}(1)$. A random walk means that the immediate response of c to any shock equals its long run response, $B_{ci}(0) = B_{ci}(1),\ i = c,\ y$. Hence, $B_{cy}(0) = B_{cy}(1)$. Thus, $B(0)$ is lower triangular if and only if $B(1)$ is lower triangular.

c a random walk is sufficient, but only the weaker condition $B_{ci}(0) = B_{ci}(1),\ i = c,\ y$ is necessary. c's response to a shock could wiggle, so long as it ends at the same place it starts.

4. If $C(0)$ is lower triangular, then the upper left hand element of $C(0)C(0)'$ is $C(0)_{11}^2$.

7.5 Granger Causality

It might happen that one variable has no response to the shocks in the other variable. This particular pattern in the impulse-response function has attracted wide attention. In this case we say that the shock variable fails to Granger cause the variable that does not respond.

The first thing you learn in econometrics is a caution that putting x on the right hand side of $y = x\beta + \epsilon$ doesn't mean that x "causes" y. (The convention that causes go on the right hand side is merely a hope that one set of causes—x—might be orthogonal to the other causes ϵ.) Then you learn that "causality" is not something you can test for statistically, but must be known a priori.

Granger causality attracted a lot of attention because it turns out that there is a limited sense in which we can test whether one variable "causes" another and vice versa.

7.5.1 Basic idea

The most natural definition of "cause" is that causes should precede effects. But this need not be the case in time-series.

Consider an economist who windsurfs.[1] Windsurfing is a tiring activity, so he drinks a beer afterwards. With W = windsurfing and B = drink a beer, a time line of his activity is given in the top panel of figure 7.1. Here we have no difficulty determining that windsurfing causes beer consumption.

But now suppose that it takes 23 hours for our economist to recover enough to even open a beer, and furthermore let's suppose that he is lucky enough to live somewhere (unlike Chicago) where he can windsurf every day. Now his time line looks like the middle panel of figure 7.1. It's still true that W causes B, but B precedes W every day. The "cause precedes effects" rule would lead you to believe that drinking beer causes one to windsurf!

How can one sort this out? The problem is that both B and W are regular events. If one could find an *unexpected* W, and see whether an *unexpected* B follows it, one could determine that W causes B, as shown in the bottom

[1] The structure of this example is due to George Akerlof.

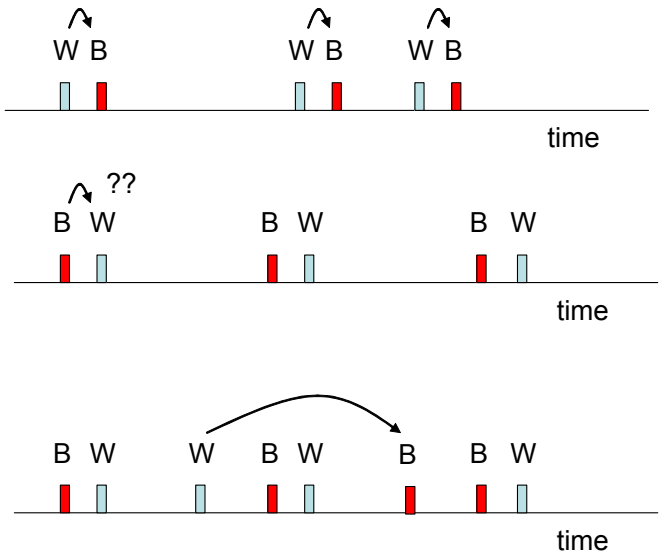

Figure 7.1:

panel of figure 7.1. So here is a possible definition: if an *unexpected* W forecasts B then we know that W "causes" B. This will turn out to be one of several equivalent definitions of Granger causality.

7.5.2 Definition, autoregressive representation

Definition: w_t *Granger causes* y_t *if* w_t *helps to forecast* y_t*, given past* y_t*.*

Consider a vector autoregression

$$y_t = a(L)y_{t-1} + b(L)w_{t-1} + \delta_t$$

$$w_t = c(L)y_{t-1} + d(L)w_{t-1} + \nu_t$$

our definition amounts to: *w_t does not Granger cause y_t if $b(L) = 0$*, i.e. if the vector autoregression is equivalent to

$$\begin{array}{rlll} y_t = & a(L)y_{t-1} & & +\delta_t \\ w_t = & c(L)y_{t-1} & +d(L)w_{t-1} & +\nu_t \end{array}$$

We can state the definition alternatively in the autoregressive representation

$$\begin{bmatrix} y_t \\ w_t \end{bmatrix} = \begin{bmatrix} a(L) & b(L) \\ c(L) & d(L) \end{bmatrix} \begin{bmatrix} y_{t-1} \\ w_{t-1} \end{bmatrix} + \begin{bmatrix} \delta_t \\ \nu_t \end{bmatrix}$$

$$\begin{bmatrix} I - La(L) & -Lb(L) \\ -Lc(L) & I - Ld(L) \end{bmatrix} \begin{bmatrix} y_t \\ w_t \end{bmatrix} = \begin{bmatrix} \delta_t \\ \nu_t \end{bmatrix}$$

$$\begin{bmatrix} a^*(L) & b^*(L) \\ c^*(L) & d^*(L) \end{bmatrix} \begin{bmatrix} y_t \\ w_t \end{bmatrix} = \begin{bmatrix} \delta_t \\ \nu_t \end{bmatrix}$$

Thus, *w does not Granger cause y iff $b^*(L) = 0$, or if the autoregressive matrix lag polynomial is lower triangular.*

7.5.3 Moving average representation

We can invert the autoregressive representation as follows:

$$\begin{bmatrix} y_t \\ w_t \end{bmatrix} = \frac{1}{a^*(L)d^*(L) - b^*(L)d^*(L)} \begin{bmatrix} d^*(L) & -b^*(L) \\ -c^*(L) & a^*(L) \end{bmatrix} \begin{bmatrix} \delta_t \\ \nu_t \end{bmatrix}$$

Thus*, w does not Granger cause y if and only if the Wold moving average matrix lag polynomial is lower triangular.* This statement gives another interpretation: if w does not Granger cause y, then y is a function of its shocks only and does not respond to w shocks. w is a function of both y shocks and w shocks.

Another way of saying the same thing is that *w does not Granger cause y if and only if y's bivariate Wold representation is the same as its univariate Wold representation,* or *w does not Granger cause y if the projection of y on past y and w is the same as the projection of y on past y alone.*

7.5.4 Univariate representations

Consider now the pair of univariate Wold representations

$$y_t = e(L)\xi_t \;\; \xi_t = y_t - P(y_t \mid y_{t-1}, y_{t-2}, \ldots);$$

$$w_t = f(L)\mu_t \;\; \mu_t = w_t - P(w_t \mid w_{t-1}, w_{t-2}, \ldots);$$

(I'm recycling letters: there aren't enough to allow every representation to have its own letters and shocks.) I repeated the properties of ξ and μ to remind you what I mean.

w_t does not Granger cause y_t if $E(\mu_t \xi_{t+j}) = 0$ for all $j > 0$. In words, w_t Granger causes y_t if the *univariate* innovations of w_t are correlated with (and hence forecast) the *univariate* innovations in y_t. In this sense, our original idea that w_t causes y_t if its movements precede those of y_t was true iff it applies to *innovations*, not the level of the series.

Proof: If w does not Granger cause y then the bivariate representation is

$$\begin{array}{rll} y_t = & a(L)\delta_t & \\ w_t = & c(L)\delta_t & +d(L)\nu_t \end{array}$$

The second line must equal the univariate representation of w_t

$$w_t = c(L)\delta_t + d(L)\nu_t = f(L)\mu_t$$

Thus, μ_t is a linear combination of current and past δ_t and ν_t. Since δ_t is the *bivariate* error, $E(\delta_t \mid y_{t-1} \ldots w_{t-1} \ldots) = E(\delta_t \mid \delta_{t-1} \ldots \nu_{t-1} \ldots) = 0$. Thus, δ_t is uncorrelated with lagged δ_t and ν_t, and hence lagged μ_t.

If $E(\mu_t \xi_{t+j}) = 0$, then past μ do not help forecast ξ, and thus past μ do not help forecast y given past y. Since one can solve for $w_t = f(L)\mu_t$(w and μ span the same space) this means past w do not help forecast y given past y.

□

7.5.5 Effect on projections

Consider the projection of w_t on the entire y process,

$$w_t = \sum_{j=-\infty}^{\infty} b_j y_{t-j} + \epsilon_t$$

Here is the fun fact:

The projection of w_t on the entire y process is equal to the projection of w_t on current and past y alone, ($b_j = 0$ for $j < 0$ if and only if w does not Granger cause y.

Proof: 1) w does not Granger cause y $\Rightarrow$one sided. If w does not Granger cause y, the bivariate representation is

$$y_t = a(L)\delta_t$$

$$w_t = d(L)\delta_t + e(L)\nu_t$$

Remember, all these lag polynomials are one-sided. Inverting the first,

$$\delta_t = a(L)^{-1} y_t$$

substituting in the second,

$$w_t = d(L)a(L)^{-1} y_t + e(L)\nu_t.$$

Since δ and ν are orthogonal at all leads and lags (we assumed contemporaneously orthogonal as well) $e(L)\nu_t$ is orthogonal to y_t at all leads and lags. Thus, the last expression is the projection of w on the entire y process. Since $d(L)$ and $a(L)^{-1}$ are one sided the projection is one sided in current and past y.

2) One sided $\Rightarrow$ w does not Granger cause y . Write the univariate representation of y as $y_t = a(L)\xi_t$ and the projection of w on the whole y process

$$w_t = h(L)y_t + \eta_t$$

The given of the theorem is that $h(L)$ is one sided. Since this is the projection on the whole y process, $E(y_t \eta_{t-s}) = 0$ for all s.

η_t is potentially serially correlated, so it has a univariate representation

$$\eta_t = b(L)\delta_t.$$

Putting all this together, y and z have a joint representation

$$\begin{array}{rcl} y_t = & a(L)\xi_t & \\ w_t = & h(L)a(L)\xi_t & +b(L)\delta_t \end{array}$$

It's not enough to make it look right, we have to check the properties. $a(L)$ and $b(L)$ are one-sided, as is $h(L)$ by assumption. Since η is uncorrelated with y at all lags, δ is uncorrelated with ξ at all lags. Since ξ and δ have the right correlation properties and $[y\ w]$ are expressed as one-sided lags of them, we have the bivariate Wold representation.

□

7.5.6 Summary

w does not Granger cause y if

1) Past w do not help forecast y given past y —Coefficients in on w in a regression of y on past y and past w are 0.

2) The autoregressive representation is lower triangular.

3) The bivariate Wold moving average representation is lower triangular.

4) Proj(w_t |all y_t) = Proj(w_t |current and past y)

5) Univariate innovations in w are not correlated with subsequent univariate innovations in y.

6) The response of y to w shocks is zero

One could use any definition as a test. The easiest test is simply an F-test on the w coefficients in the VAR. Monte Carlo evidence suggests that this test is also the most robust.

7.5.7 Discussion

It is not necessarily the case that one pair of variables must Granger cause the other and not vice versa. We often find that each variable responds to the other's shock (as well as its own), or that there is feedback from each variable to the other.

The first and most famous application of Granger causality was to the question "does money growth cause changes in GNP?" Friedman and Schwartz (19) documented a correlation between money growth and GNP, and a tendency for money changes to lead GNP changes. But Tobin (19) pointed out that, as with the windsurfing example given above, a phase lead and a correlation may not indicate causality. Sims (1972) applied a Granger causality test, which answers Tobin's criticism. In his first work, Sims found that money Granger causes GNP but not vice versa, though he and others have found different results subsequently (see below).

Sims also applied the last representation result to study regressions of GNP on money,

$$y_t = \sum_{j=0}^{\infty} b_j m_{t-j} + \delta_t.$$

This regression is known as a "St. Louis Fed" equation. The coefficients were interpreted as the response of y to changes in m; i.e. if the Fed sets $m, \{b_j\}$ gives the response of y. Since the coefficients were "big", the equations implied that constant money growth rules were desirable.

The obvious objection to this statement is that the coefficients may reflect reverse causality: the Fed sets money in anticipation of subsequent economic growth, or the Fed sets money in response to past y. In either case, the error term δ is correlated with current and lagged m's so OLS estimates of the b's are inconsistent.

Sims (1972) ran causality tests essentially by checking the pattern of correlation of univariate shocks, and by running regressions of y on past *and future* m, and testing whether coefficients on future m are zero. He concluded that the "St. Louis Fed" equation is correctly specified after all. Again, as we see next, there is now some doubt about this proposition. Also, even if correctly estimated, the projection of y on all m's is not necessarily the answer to "what if the Fed changes m".

	Explained by shocks to		
Var. of	M1	IP	WPI
M1	97	2	1
IP	37	44	18
WPI	14	7	80

Table 7.1: Sims variance accounting

7.5.8 A warning: why "Granger causality" is not "Causality"

"Granger causality" is not causality in a more fundamental sense because of the possibility of other variables. If x leads to y with one lag but to z with two lags, then y will Granger cause z in a bivariate system$-$ $-$ y will help forecast z since it reveals information about the "true cause" x. But it does not follow that if you change y (by policy action), then a change in z will follow. The weather forecast Granger causes the weather (say, rainfall in inches), since the forecast will help to forecast rainfall amount given the time-series of past rainfall. But (alas) shooting the forecaster will not stop the rain. The reason is that forecasters use a lot more information than past rainfall.

This wouldn't be such a problem if the estimated pattern of causality in macroeconomic time series was stable over the inclusion of several variables. But it often is. A beautiful example is due to Sims (1980). Sims computed a VAR with money, industrial production and wholesale price indices. He summarized his results by a 48 month ahead forecast error variance, shown in table 7.1

The first row verifies that $M1$ is exogenous: it does not respond to the other variables' shocks. The second row shows that $M1$ "causes" changes in IP, since 37% of the 48 month ahead variance of IP is due to $M1$ shocks. The third row is a bit of a puzzle: WPI also seems exogenous, and not too influenced by $M1$.

Table 7.2 shows what happens when we add a further variable, the interest rate. Now, the second row shows a substantial response of money to interest

	Explained by shocks to			
Var of	R	M1	WPI	IP
R	50	19	4	28
M1	56	42	1	1
WPI	2	32	60	6
IP	30	4	14	52

Table 7.2: Sims variance accounting including interest rates

rate shocks. It's certainly not exogenous, and one could tell a story about the Fed's attempts to smooth interest rates. In the third row, we now find that M *does* influence WPI. And, worst of all, the fourth row shows that M does *not* influence IP; the *interest rate* does. Thus, *interest rate* changes seem to be the driving force of real fluctuations, and money just sets the price level! However, later authors have interpreted these results to show that interest rates are in fact the best indicators of the Fed's monetary stance.

Notice that Sims gives an economic measure of feedback (forecast error variance decomposition) rather than F-tests for Granger causality. Since the first flush of optimism, economists have become less interested in the pure hypothesis of no Granger causality at all, and more interested in simply quantifying how much feedback exists from one variable to another. And sensibly so.

Any variance can be broken down by frequency. Geweke (19) shows how to break the variance decomposition down by frequency, so you get measures of feedback at each frequency. This measure can answer questions like "does the Fed respond to low or high frequency movements in GNP?", etc.

7.5.9 Contemporaneous correlation

Above, I assumed where necessary that the shocks were orthogonal. One can expand the definition of Granger causality to mean that *current* and past w do not predict y given past y. This means that the orthogonalized MA is lower triangular. Of course, this version of the definition will depend on the order of orthogonalization. Similarly, when thinking of Granger causality in

terms of impulse response functions or variance decompositions you have to make one or the other orthogonalization assumption.

Intuitively, the problem is that one variable may affect the other so quickly that it is within the one period at which we observe data. Thus, we can't use our statistical procedure to see whether contemporaneous correlation is due to y causing w or vice-versa. Thus, the orthogonalization assumption is equivalent to an assumption about the direction of *instantaneous* causality.

Chapter 8

Spectral Representation

The third fundamental representation of a time series is its *spectral density.* This is just the Fourier transform of the autocorrelation/ autocovariance function. If you don't know what that means, read on.

8.1 Facts about complex numbers and trigonometry

8.1.1 Definitions

Complex numbers are composed of a real part plus an imaginary part, $z = A + Bi$, where $i = (-1)^{1/2}$. We can think of a complex number as a point on a plane with reals along the x axis and imaginary numbers on the y axis as shown in figure 8.1.

Using the identity $e^{i\theta} = \cos\theta + i\sin\theta$, we can also represent complex numbers in *polar notation* as $z = Ce^{i\theta}$ where $C = (A^2+B^2)^{1/2}$ is the *amplitude* or *magnitude,* and $\theta = \tan^{-1}(B/A)$ is the *angle* or *phase.* The length C of a complex number is also denoted as its norm $\mid z \mid$.

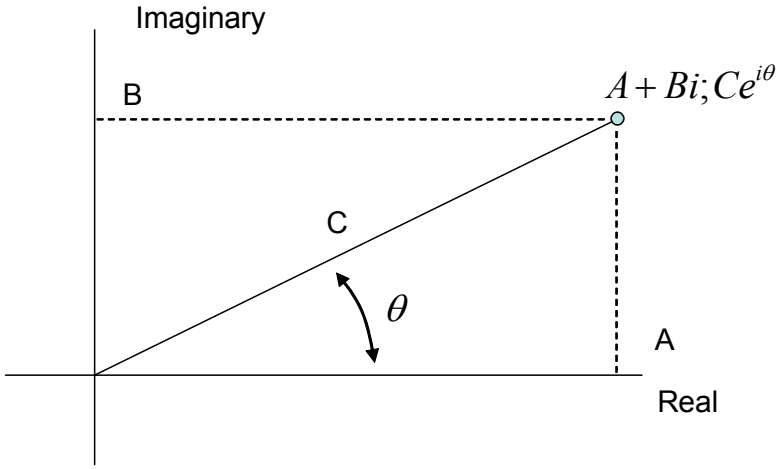

Figure 8.1: Graphical representation of the complex plane.

8.1.2 Addition, multiplication, and conjugation

To add complex numbers, you add each part, as you would any vector

$$(A + Bi) + (C + Di) = (A + C) + (B + D)i.$$

Hence, we can represent addition on the complex plane as in figure 8.2

You multiply them just like you'd think:

$$(A + Bi)(C + Di) = AC + ADi + BCi + BDi = (AC - BD) + (AD + BC)i.$$

Multiplication is easier to see in polar notation

$$De^{i\theta_1}Ee^{i\theta_2} = DEe^{i(\theta_1+\theta_2)}$$

Thus, multiplying two complex numbers together gives you a number whose magnitude equals the product of the two magnitudes, and whose angle (or phase) is the sum of the two angles, as shown in figure 8.3. Angles are denoted in radians, so $\pi = 180^0$, etc.

The *complex conjugate* * is defined by

$$(A + Bi)^* = A - Bi \text{ and } (Ae^{i\theta})^* = Ae^{-i\theta}.$$

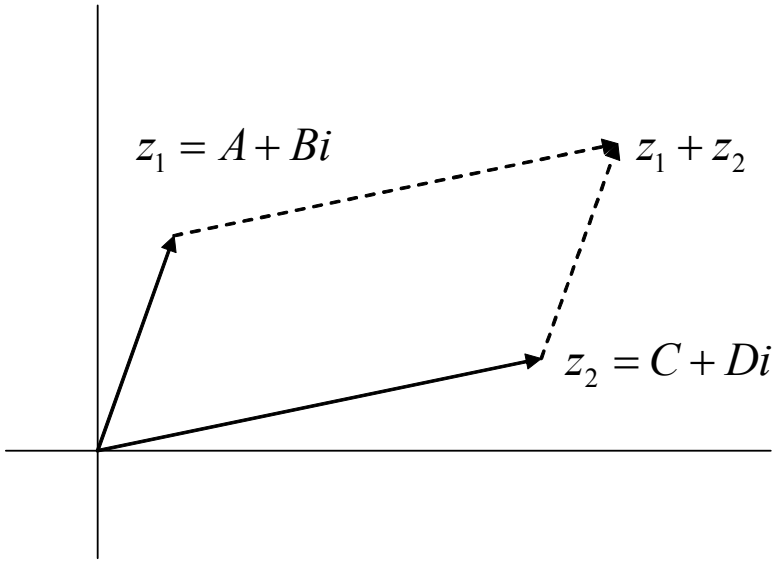

Figure 8.2: Complex addition

This operation simply flips the complex vector about the real axis. Note that $zz^* =\mid z \mid^2$, and $z + z^* = 2Re(z)$ is real..

8.1.3 Trigonometric identities

From the identity

$$e^{i\theta} = \cos\theta + i\sin\theta,$$

two useful identities follow

$$\cos\theta = (e^{i\theta} + e^{-i\theta})/2$$
$$\sin\theta = (e^{i\theta} - e^{-i\theta})/2i$$

8.1.4 Frequency, period and phase

Figure 8.4 reminds you what sine and cosine waves look like.

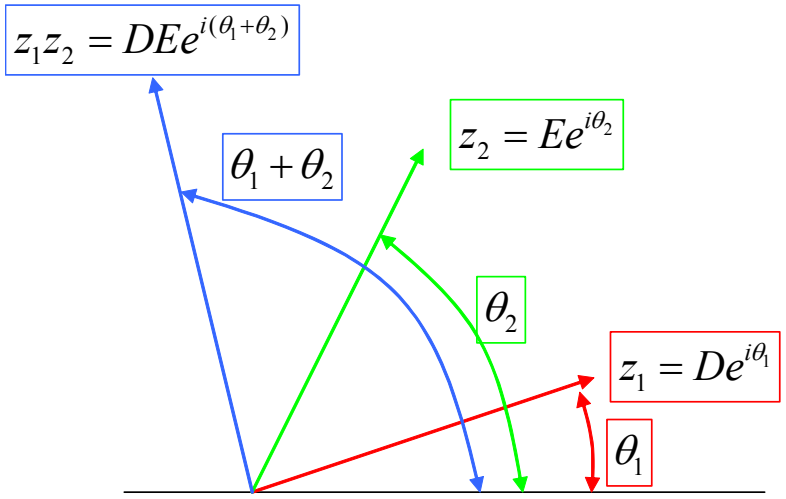

Figure 8.3: Complex multiplication

The *period* λ is related to the *frequency* ω by $\lambda = 2\pi/\omega$. The period λ is the amount of time it takes the wave to go through a whole cycle. The frequency ω is the angular speed measured in radians/time. The *phase* is the angular amount ϕ by which the sine wave is shifted. Since it is an angular displacement, the *time* shift is ϕ/ω.

8.1.5 Fourier transforms

Take any series of numbers $\{x_t\}$. We define its *Fourier transform* as

$$x(\omega) = \sum_{t=-\infty}^{\infty} e^{-i\omega t} x_t$$

Note that this operation transforms a series, a function of t, to a complex-valued function of ω. Given $x(\omega)$, we can recover x_t, by the *inverse Fourier transform*

$$x_t = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{+i\omega t} x(\omega) d\omega.$$

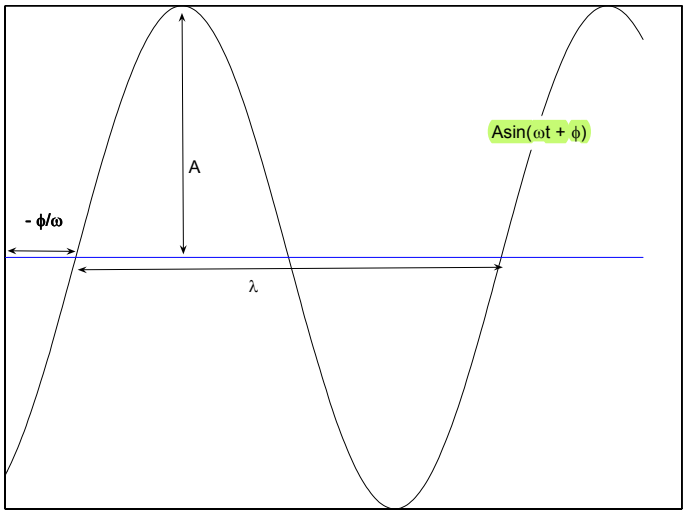

Figure 8.4: Sine wave amplitude A, period λ and frequency ω.

Proof: Just substitute the definition of $x(\omega)$ in the inverse transform, and verify that we get x_t back.

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{+i\omega t}\left(\sum_{\tau=-\infty}^{\infty} e^{-i\omega\tau}x_\tau\right)d\omega = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} x_\tau \int_{-\pi}^{\pi} e^{+i\omega t}e^{-i\omega\tau}d\omega$$

$$= \sum_{\tau=-\infty}^{\infty} x_\tau \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega(t-\tau)}d\omega$$

Next, let's evaluate the integral.

$$t=\tau \Rightarrow \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega(t-\tau)}d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} d\omega = 1,$$

$$t-\tau=1 \Rightarrow \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega(t-\tau)}d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega}d\omega = 0$$

since the integral of sine or cosine all the way around the circle is zero. The same point holds for any $t \neq \tau$, thus (this is another important fact about complex numbers)

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega(t-\tau)}d\omega = \delta(t-\tau) = \begin{cases} 1 \text{ if } t-\tau = 0 \\ 0 \text{ if } t-\tau \neq 0 \end{cases}$$

Picking up where we left off,

$$\sum_{\tau=-\infty}^{\infty} x_\tau \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega(t-\tau)}d\omega = \sum_{\tau=-\infty}^{\infty} x_\tau \delta(t-\tau) = x_t.$$

□

The inverse Fourier transform expresses x_t as a sum of sines and cosines at each frequency ω. We'll see this explicitly in the next section.

8.1.6 Why complex numbers?

You may wonder why complex numbers pop up in the formulas, since all economic time series are real (i.e., the sense in which they contain imaginary numbers has nothing to do with the square root of -1). The answer is that they don't have to: we can do all the analysis with only real quantities. However, it's simpler with the complex numbers. The reason is that we always have to keep track of *two* real quantities, and the complex numbers do this for us by keeping track of a real and imaginary part in one symbol. The answer is always real.

To see this point, consider what a more intuitive inverse Fourier transform might look like:

$$x_t = \frac{1}{\pi}\int_0^{\pi} \mid x(\omega) \mid \cos(\omega t + \phi(\omega))d\omega$$

Here we keep track of the amplitude $\mid x(\omega) \mid$ (a real number) and phase $\phi(\omega)$ of components at each frequency ω. It turns out that this form is exactly the same as the one given above. In the complex version, the magnitude of $x(\omega)$ tells us the amplitude of the component at frequency ω, the phase of

$x(\omega)$ tells use the phase of the component at frequency ω, and we don't have to carry the two around separately. But which form you use is really only a matter of convenience.

Proof:

Writing $x(\omega) = |x(\omega)|\ e^{i\phi(\omega)}$,

$$x_t = \frac{1}{2\pi}\int_{-\pi}^{\pi} x(\omega)e^{i\omega t}d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mid x(\omega) \mid e^{i(\omega t+\phi(\omega))}d\omega$$

$$= \frac{1}{2\pi}\int_{0}^{\pi} \left(\mid x(\omega) \mid e^{i(\omega t+\phi(\omega))} + \mid x(-\omega) \mid e^{i(-\omega t+\phi(-\omega))}\right) d\omega.$$

But $x(\omega) = x(-\omega)^*$ (to see this, $x(-\omega) = \sum_t e^{i\omega t}x_t = \left(\sum_t e^{-i\omega t}x_t\right)^* = x(\omega)^*$), so $\mid x(-\omega) \mid = \mid x(\omega) \mid$, $\phi(-\omega) = -\phi(\omega)$. Continuing,

$$x_t = \frac{1}{2\pi}\int_{0}^{\pi} \mid x(\omega) \mid \left(e^{i(\omega t+\phi(\omega))} + e^{-i(\omega t+\phi(\omega))}\right) d\omega = \frac{1}{\pi}\int_{0}^{\pi} \mid x(\omega) \mid \cos(\omega t+\phi(\omega))d\omega.$$

□

As another example of the inverse Fourier transform interpretation, suppose $x(\omega)$ was a spike that integrates to one (a delta function) at ω and $-\omega$. Since $\sin(-\omega t) = -\sin(\omega t)$, we have $x_t = 2\cos(\omega t)$.

8.2 Spectral density

The spectral density is defined as the Fourier transform of the autocovariance function

$$S(\omega) = \sum_{j=-\infty}^{\infty} e^{-i\omega j}\gamma_j$$

Since γ_j is symmetric, $S(\omega)$ is real

$$S(\omega) = \gamma_0 + 2\sum_{j=1}^{\infty} \gamma_j \cos(j\omega)$$

The formula shows that, again, we could define the spectral density using real quantities, but the complex versions are prettier. Also, notice that the symmetry $\gamma_j = \gamma_{-j}$ means that $S(\omega)$ is symmetric: $S(\omega) = S(-\omega)$, and real.

Using the inversion formula, we can recover γ_j from $S(\omega)$.

$$\gamma_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{+i\omega j} S(\omega) d\omega.$$

Thus, the spectral density is an autocovariance generating function. In particular,

$$\gamma_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega) d\omega$$

This equation interprets the spectral density as a decomposition of the variance of the process into uncorrelated components at each frequency ω (if they weren't uncorrelated, their variances would not sum without covariance terms). We'll come back to this interpretation later.

Two other sets of units are sometimes used. First, we could divide everything by the variance of the series, or, equivalently, Fourier transform the autocorrelation function. Since $\rho_j = \gamma_j/\gamma_0$,

$$f(\omega) = S(\omega)/\gamma_0 = \sum_{j=-\infty}^{\infty} e^{-i\omega j} \rho_j$$

$$\rho_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{+i\omega j} f(\omega) d\omega.$$

$$1 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\omega) d\omega.$$

$f(\omega)/2\pi$ looks just like a probability density: it's real, positive and integrates to 1. Hence the terminology "spectral density". We can define the corresponding distribution function

$$F(\omega) = \int_{-\pi}^{\omega} \frac{1}{2\pi} f(\nu)\, d\nu. \text{ where } F(-\pi) = 0, F(\pi) = 1 \; F \text{ increasing}$$

This formalism is useful to be precise about cases with deterministic components and hence with "spikes" in the density.

8.2.1 Spectral densities of some processes

White noise

$$x_t = \epsilon_t$$
$$\gamma_0 = \sigma_\epsilon^2, \gamma_j = 0 \text{ for } j > 0$$
$$S(\omega) = \sigma_\epsilon^2 = \sigma_x^2$$

The spectral density of white noise is flat.

MA(1)

$$x_t = \epsilon_t + \theta\epsilon_{t-1}$$
$$\gamma_0 = (1+\theta^2)\sigma_\epsilon^2,\ \gamma_1 = \theta\sigma_\epsilon^2,\ \gamma_j = 0 \text{ for } j > 1$$

$$S(\omega) = (1+\theta^2)\sigma_\epsilon^2 + 2\theta\sigma_\epsilon^2\cos\omega = (1+\theta^2+2\theta\cos\omega)\sigma_\epsilon^2 = \gamma_0(1+\ \frac{2\theta}{1+\theta^2}\cos\omega)$$

Hence, $f(\omega) = S(\omega)/\gamma_0$ is

$$f(\omega) = 1 + \frac{2\theta}{1+\theta^2}\cos\omega$$

Figure 8.5 graphs this spectral density.

As you can see, "smooth" MA(1)'s with $\theta > 0$ have spectral densities that emphasize low frequencies, while "choppy" ma(1)'s with $\theta < 0$ have spectral densities that emphasize high frequencies.

Obviously, this way of calculating spectral densities is not going to be very easy for more complicated processes (try it for an $AR(1)$.) A by-product of the filtering formula I develop next is an easier way to calculate spectral densities.

8.2.2 Spectral density matrix, cross spectral density

With $x_t = [y_t\ z_t]'$, we defined the variance-covariance matrix $\Gamma_j = E(x_t x_{t-j}')$, which was composed of auto- and cross-covariances. The *spectral density matrix* is defined as

$$S_x(\omega) = \sum_{j=-\infty}^{\infty} e^{-i\omega j}\Gamma_j = \begin{bmatrix} \sum_j e^{-i\omega j}\gamma_y(j) & \sum_j e^{-i\omega j}E(y_t z_{t-j}) \\ \sum_j e^{-i\omega j}E(z_t y_{t-j}) & \sum_j e^{-i\omega j}\gamma_z(j) \end{bmatrix}$$

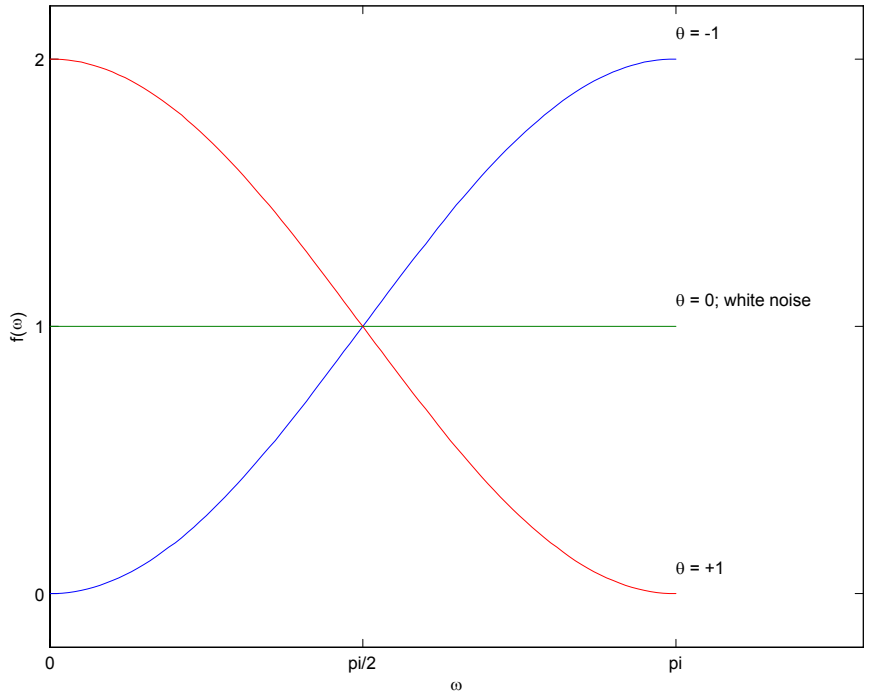

Figure 8.5: MA(1) spectral density

You may recognize the diagonals as the spectral densities of y and z. The off-diagonals are known as the *cross-spectral densities.*

$$S_{yz}(\omega) = \sum_{j=-\infty}^{\infty} e^{-i\omega j} E(y_t z_{t-j}).$$

Recall that we used the symmetry of the autocovariance function $\gamma_j = \gamma_{-j}$ to show that the spectral density is real and symmetric in ω. However, it is *not* true that $E(y_t z_{t-j}) = E(z_t y_{t-j})$ so the cross-spectral density need not be real, symmetric, or positive. It does have the following symmetry property:

$$S_{yz}(\omega) = [S_{zy}(\omega)]^* = S_{zy}(-\omega)$$

Proof:

$$S_{yz}(\omega) = \sum_{j=-\infty}^{\infty} e^{-ij\omega} E(y_t z_{t-j}) = \sum_{j=-\infty}^{\infty} e^{-ij\omega} E(z_t y_{t+j}) =$$

$$\sum_{k=-\infty}^{\infty} e^{ik\omega} E(z_t y_{t-k}) = [S_{zy}(\omega)]^* = S_{zy}(-\omega).$$

□

As with any Fourier transform, we can write the corresponding inverse transform

$$E(y_t z_{t-j}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\omega j} S_{yz}(\omega) d\omega$$

and, in particular,

$$E(y_t z_t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{yz}(\omega) d\omega.$$

The cross-spectral density decomposes the covariance of two series into the covariance of components at each frequency ω. While the spectral density is real, the cross-spectral density may be complex. We can characterize the relation between two sine or cosine waves at frequency ω by the product of their amplitudes, which is the magnitude of $\mathrm{S}_{yz}(\omega)$, and the phase shift between them, which is the phase of $S_{yz}(\omega)$.

8.2.3 Spectral density of a sum

Recall that the variance of a sum is $\mathrm{var}(a+b) = \mathrm{var}(a) + \mathrm{var}(b) + 2\mathrm{cov}(a,b)$. Since spectral densities are variances of "components at frequency ω" they obey a similar relation,

$$S_{x+y}(\omega) = S_x(\omega) + S_y(\omega) + S_{xy}(\omega) + S_{yx}(\omega)$$

$$= S_x(\omega) + S_y(\omega) + (S_{xy}(\omega) + S_{xy}(\omega)^*)$$

$$= S_x(\omega) + S_y(\omega) + 2Re(S_{xy}(\omega))$$

where $Re(x)$ = the real part of x.

Proof: As usual, use the definitions and algebra:

$$S_{x+y}(\omega) = \sum_j e^{-i\omega j} E[(x_t + y_t)(x_{t-j} + y_{t-j})] =$$

$$= \sum_j e^{-i\omega j} \left(E(x_t x_{t-j}) + E(y_t y_{t-j}) + E(x_t y_{t-j}) + E(y_t x_{t-j})\right) =$$

$$= S_x(\omega) + S_y(\omega) + S_{xy}(\omega) + S_{yx}(\omega).$$

□

In particular, if x and y are uncorrelated at all leads and lags, their spectral densities add.

$$E(x_t y_{t-j}) = 0 \text{ for all } j \Rightarrow S_{x+y}(\omega) = S_x(\omega) + S_y(\omega)$$

8.3 Filtering

8.3.1 Spectrum of filtered series

Suppose we form a series y_t by *filtering* a series x_t, i.e. by applying a moving average

$$y_t = \sum_{j=-\infty}^{\infty} b_j x_{t-j} = b(L)x_t.$$

It would be nice to characterize the process y_t given the process x_t.

We could try to derive the autocovariance function of y_t given that of x_t. Let's try it:

$$\gamma_k(y) = E(y_t y_{t-k}) = E\left(\sum_{j=-\infty}^{\infty} b_j x_{t-j} \sum_{l=-\infty}^{\infty} b_l x_{t-k-l}\right)$$

$$= \sum_{j,l} b_j b_l E(x_{t-j} x_{t-k-l}) = \sum_{j,l} b_j b_l \gamma_{k+l-j}(x)$$

This is not a pretty convolution.

However, the formula for the *spectral density* of y given the spectral density of x turns out to be very simple:

$$S_y(\omega) = \mid b(e^{-i\omega}) \mid^2 S_x(\omega).$$

$b(e^{-i\omega})$ is a nice notation for the Fourier transform of the b_j coefficients. $b(L) = \sum_j b_j L^j$, so $\sum_j e^{-i\omega j} b_j = b(e^{-i\omega})$.

Proof: Just plug in definitions and go.

$$S_y(\omega) = \sum_k e^{-i\omega k} \gamma_k(y) = \sum_{k,j,l} e^{-i\omega k} b_j b_l \gamma_{k+l-j}(x)$$

Let $h = k + l - j$, so $k = h - l + j$

$$S_y(\omega) = \sum_{h,j,l} e^{-i\omega(h-l+j)} b_j b_l \gamma_h(x) = \sum_j e^{-i\omega j} b_j \sum_l e^{+i\omega l} b_l \sum_h e^{-i\omega h} \gamma_h(x)$$

$$= b(e^{-i\omega}) b(e^{+i\omega}) S_x(\omega) = \mid b(e^{-i\omega}) \mid^2 S_x(\omega).$$

The last equality results because $b(z^*) = b(z)^*$ for polynomials.

□

The filter $y_t = b(L)x_t$ is a complex dynamic relation. Yet the filtering formula looks just like the scalar formula $y = bx \Rightarrow \text{var}(y) = b^2 \text{ var}(x)$. This starts to show you why the spectral representation is so convenient: operations with a difficult dynamic structure in the time domain (convolutions) are just multiplications in the frequency domain.

8.3.2 Multivariate filtering formula

The vector version of the filtering formula is a natural extension of the scalar version.

$$y_t = B(L)x_t \Rightarrow S_y(\omega) = B(e^{-i\omega}) S_x(\omega) B(e^{i\omega})'.$$

This looks just like the usual variance formula: if x is a vector-valued random variable with covariance matrix Σ and $y = Ax$, then the covariance matrix of y is $A\Sigma A'$.

8.3.3 Spectral density of arbitrary MA(∞)

Since the $MA(\infty)$ representation expresses any series as a linear filter of white noise, an obvious use of the filtering formula is a way to derive the spectral density of any ARMA expressed in Wold representation,

$$x_t = \theta(L)\epsilon_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}.$$

By the filtering formula,

$$S_x(\omega) = \theta(e^{-i\omega})\theta(e^{+i\omega})\sigma_\epsilon^2$$

Now you know how to find the spectral density of *any* process. For example, the $MA(1)$, $x_t = (1+\theta L)\epsilon_t$ gives

$$S_x(\omega) = (1+\theta e^{-i\omega})(1+\theta e^{i\omega})\sigma_\epsilon^2 = \left(1+\theta(e^{i\omega}+e^{-i\omega})+\theta^2\right)\sigma_\epsilon^2 = (1+2\theta\cos(\omega)+\theta^2)\sigma_\epsilon^2$$

as before.

8.3.4 Filtering and OLS

Suppose

$$y_t = b(L)x_t + \epsilon_t, \quad E(x_t\epsilon_{t-j}) = 0 \text{ for all } j.$$

This is what OLS of y_t on the entire x_t process produces. Then, adding the filtering formula to the fact that spectral densities of uncorrelated processes add, we have

$$S_y(\omega) = S_{b(L)x}(\omega) + S_\epsilon(\omega) = \mid b(e^{-i\omega}) \mid^2 S_x(\omega) + S_\epsilon(\omega).$$

This formula looks a lot like $y_t = x_t\beta + \epsilon_t \Rightarrow \sigma_y^2 = \beta^2\sigma_x^2 + \sigma_\epsilon^2$, and the resulting formula for R^2. Thus, it lets us do an R^2 decomposition — how much of the variance of y at each frequency is due to x and ϵ?

More interestingly, we can relate the cross-spectral density to $b(L)$,

$$S_{yx}(\omega) = b(e^{-i\omega})S_x(\omega).$$

Proof: The usual trick: write out the definition and play with the indices until you get what you want.

$$S_{yx}(\omega) = \sum_k e^{-i\omega k} E(y_t x_{t-k}) = \sum_k e^{-i\omega k} \sum_j b_j E(x_{t-j} x_{t-k})$$

$$= \sum_k e^{-i\omega k} \sum_j b_j \gamma_{k-j}(x)$$

Let $l = k - j$, so $k = l + j$,

$$S_{yx}(\omega) = \sum_l e^{-i\omega(l+j)} \sum_j b_j \gamma_l(x) = \sum_j e^{-i\omega j} b_j \sum_l e^{-i\omega l} \gamma_l(x) = b(e^{-i\omega}) S_x(\omega)$$

□

This formula has a number of uses. First, divide through to express

$$b(e^{-i\omega}) = \frac{S_{yx}(\omega)}{S_x(\omega)}.$$

This looks a lot like $\beta = \text{cov}(y, x)/\text{var}(x)$. Again, the spectral representation reduces complex dynamic representations to simple scalar multiplication, at each frequency ω. Second, you can estimate the lag distribution, or think about lag distributions this way;

$$\hat{b}_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ij\omega} b(e^{-i\omega})\, d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ij\omega} \frac{S_{yx}(\omega)}{S_x(\omega)}\, d\omega$$

is known as "Hannan's inefficient estimator." Third, we can turn the formula around, and use it to help us to understand the cross-spectral density:

$$S_{yx} = b(e^{-i\omega}) S_x(\omega).$$

For example, if $x_t = \epsilon_t$, $\sigma_\epsilon^2 = 1$, and $y_t = x_{t-1}$, then $S_y(\omega) = S_x(\omega) = 1$, but $S_{yx}(\omega) = e^{i\omega}$. The real part of this cross-spectral density is 1.0—x is neither stretched nor shrunk in its transformation to y. But the one-period lag shows up in the complex part of the cross-spectral density—y lags x by ω.

8.3.5 A cosine example

The following example may help with the intuition of these filtering formulas. Suppose $x_t = \cos(\omega t)$. Then we can find by construction that

$$y_t = b(L)x_t = \mid b(e^{-i\omega}) \mid \cos(\omega t + \phi(\omega)).$$

where

$$b(e^{-i\omega}) = \mid b(e^{-i\omega}) \mid e^{i\phi(\omega)}.$$

The quantity $b(e^{-i\omega})$ is known as the *frequency response* of the filter. Its magnitude (or magnitude squared) is called the *gain* of the filter and its angle ϕ is known as the *phase*. Both are functions of frequency.

The spectral representation of the filter shows you what happens to sine waves of each frequency. All you can do to a sine wave is stretch it or shift it, so we can represent what happens to a sine wave at each frequency by two numbers, the gain and phase of the $b(e^{-i\omega})$. The usual representation $b(L)$ shows you what happens in response to a unit impulse. This is, of course, a complex dynamic relation. Then, we can either think of y_t as the sum of impulses times the complex dynamic response to each impulse, or as the sum of sine and cosine waves times the simple gain and phase -

Proof:

$$y_t = \frac{1}{2}\sum_j b_j(e^{i\omega(t-j)} + e^{-i\omega(t-j)}) = \frac{1}{2}\left(e^{i\omega t}\sum_j b_j e^{-i\omega j} + e^{-i\omega t}\sum_j b_j e^{i\omega j}\right) =$$

$$\frac{1}{2}\left(e^{i\omega t}b(e^{-i\omega}) + e^{-i\omega t}b(e^{i\omega})\right) = \mid b(e^{-i\omega}) \mid \cos(\omega t + \phi(\omega)).$$

($\mid b(e^{-i\omega}) \mid = \mid b(e^{i\omega}) \mid$ for polynomial $b(L)$).

□

8.3.6 Cross spectral density of two filters, and an interpretation of spectral density

Here's a final filtering formula, which will help give content to the interpretation of the spectral density as the variance of components at frequency

ω.

$$y_{1t} = b_1(L)x_t,\ y_{2t} = b_2(L)x_t \Rightarrow S_{y_1y_2}(\omega) = b_1(e^{-i\omega})b_2(e^{i\omega})S_x(\omega)$$

Note this formula reduces to $S_y(\omega) =\mid b(e^{-i\omega}) \mid^2 S_x(\omega)$ if $b_1(L) = b_2(L)$.

Proof: As usual, write out the definition, and play with the sum indices.

$$S_{y_1y_2}(\omega) = \sum_j e^{-i\omega j} E(y_{1t}y_{2t-j}) = \sum_j e^{-i\omega j} E\left(\sum_k b_{1k}x_{t-k} \sum_l b_{2l}x_{t-j-l}\right)$$

$$= \sum_j e^{-i\omega j} \sum_{k,l} b_{1k}b_{2l}E(x_{t-k}x_{t-j-l}) = \sum_j e^{-i\omega j} \sum_{k,l} b_{1k}b_{2l}\gamma_{j+l-k}(x)$$

Let $m = j + l - k$, so $j = m - l + k$,

$$\sum_m e^{-i\omega(m-l+k)} \sum_k \sum_l b_{1k}b_{2l}\gamma_m = \sum_k b_{1k}e^{-i\omega k} \sum_l b_{2l}e^{i\omega l} \sum_m e^{-i\omega m}\gamma_m(x).$$

□

Now, suppose $b(L)$ is a *bandpass filter,*

$$b(e^{-i\omega}) = \begin{cases} 1, \mid \omega \mid \in [\alpha, \beta] \\ 0 \text{ elsewhere} \end{cases},$$

as displayed in 8.6. Then, the variance of filtered data gives the average spectral density in the window. (We'll use this idea later to construct spectral density estimates.)

$$\text{var}(b(L)x_t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mid b(e^{-i\omega}) \mid^2 S_x(\omega)d\omega = \frac{1}{2\pi} \int_{|\omega| \in [\alpha,\beta]} S_x(\omega)d\omega.$$

Next, subdivide the frequency range from $-\pi$ to π into nonoverlapping intervals

$$\text{var}(x_t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega)d\omega = \frac{1}{2\pi} \left(\int_{b_1} S_x(\omega)d\omega + \int_{b_2} S_x(\omega)d\omega..\right) =$$

$$\text{var}(b_1(L)x_t) + \text{var}(b_2(L)x_t) + \ldots$$

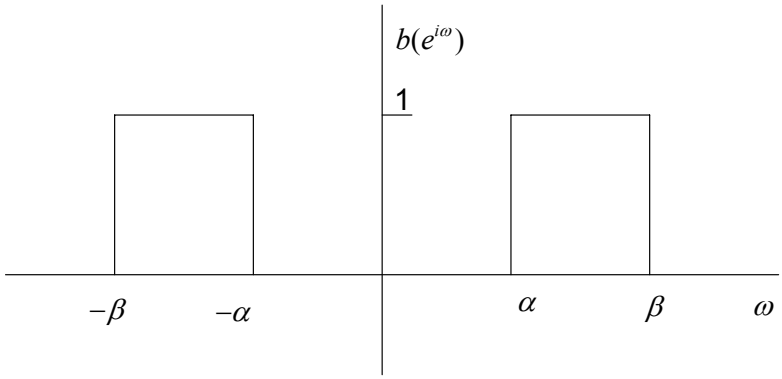

Figure 8.6: Frequency response (all real) of a bandpass filter that includes frequencies ω from α to β.

Since the windows do not overlap, the covariance terms are zero. As the windows get smaller and smaller, the windows approach a delta function, so the components $b_j(L)x_t$ start to look more and more like pure cosine waves at a single frequency. In this way, the spectral density decomposes the variance of the series into the variance of *orthogonal* sine and cosine waves at different frequencies.

8.3.7 Constructing filters

Inversion formula

You may have wondered in the above, how do we know that a filter exits with a given desired frequency response? If it exists, how do you construct it? The answer, of course, is to use the inversion formula,

$$b_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega j} b(e^{-i\omega}) d\omega$$

For example, let's find the moving average representation b_j of the bandpass filter.

$$b_j = \frac{1}{2\pi}\int_{|\omega|\in[\alpha,\beta]} e^{i\omega j} d\omega = \frac{1}{2\pi}\int_{-\beta}^{-\alpha} e^{i\omega j} d\omega + \frac{1}{2\pi}\int_{\alpha}^{\beta} e^{i\omega j} d\omega =$$

$$= \frac{1}{2\pi}\,\frac{e^{i\omega j}}{ij}\Big]_{-\beta}^{-\alpha} + \frac{1}{2\pi}\,\frac{e^{i\omega j}}{ij}\Big]_{\alpha}^{\beta} = \frac{e^{-ij\alpha} - e^{-ij\beta} + e^{ij\beta} - e^{ij\alpha}}{2\pi ij} =$$

$$\frac{\sin(j\beta)}{\pi j} - \frac{\sin(j\alpha)}{\pi j}.$$

Each term is a two-sided infinite order moving average filter. Figure 8.7 plots the filter. Then, you just apply this two-sided moving average to the series.

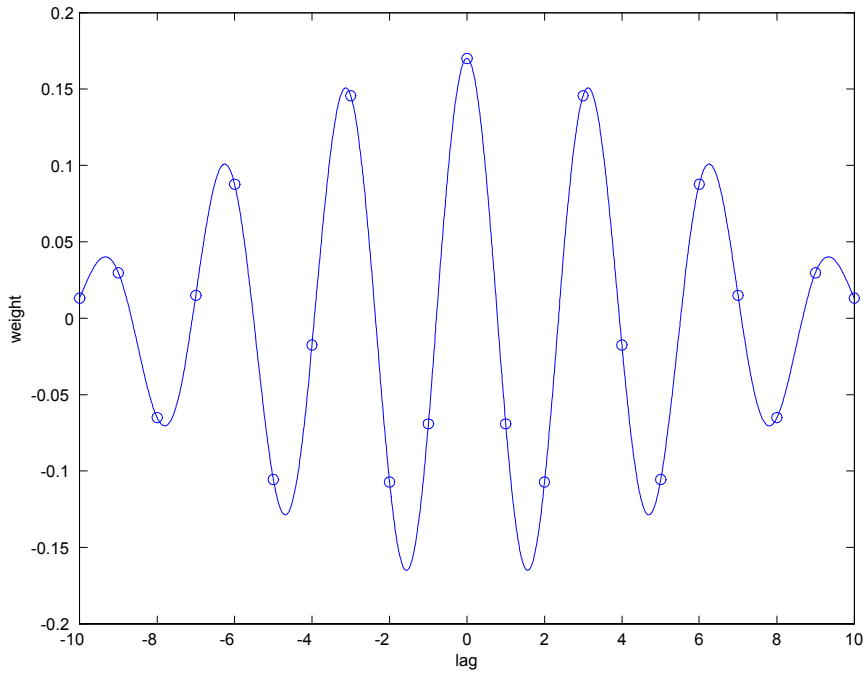

Figure 8.7: Moving average weights of a bandpass filter.

Extracting unobserved components by knowledge of spectra.

Of course, you may not know ahead of time exactly what frequency response $b(e^{-i\omega})$ you're looking for. One situation in which we can give some help is a case of *unobserved components.* In seasonal adjustment, growth/cycle

decompositions, and many other situations, you think of your observed series X_t as composed of two components, x_t and u_t that you do not observe separately. However, you do have some prior ideas about the spectral shape of the two components. Growth is long-run, seasonals are seasonal, etc.

Assume that the two components are uncorrelated at all leads and lags.

$$X_t = x_t + u_t, \quad E(x_t u_s) = 0$$

The only thing you can do is construct a "best guess" of x_t given your data on X_t, i.e., construct a filter that recovers the projection of x_t on the X process,

$$x_t = \sum_{j=-\infty}^{\infty} h_j X_{t-j} + \epsilon_t \Rightarrow \hat{x}_t = \sum_{j=-\infty}^{\infty} h_j X_{t-j}.$$

The parameters h_j are given from inverting

$$h(e^{-i\omega}) = \frac{S_x(\omega)}{S_x(\omega) + S_u(\omega)} = \frac{S_x(\omega)}{S_X(\omega)}$$

Proof: (left as a problem)

This formula is an example of a problem that is much easier to solve in the frequency domain! (A fancy name might be "optimal seasonal (or trend/cycle) extraction".)

8.3.8 Sims approximation formula

Often, (always) we approximate true, infinite-order ARMA processes by finite order models. There is a neat spectral representation of this approximation given by the *Sims approximation* formula: Suppose the true projection of y_t on $\{x_t\}$ is

$$y_t = \sum_{j=-\infty}^{\infty} b_j^0 x_{t-j} + \epsilon_t$$

but a researcher fits by OLS a restricted version,

$$y_t = \sum_{j=-\infty}^{\infty} b_j^1 x_{t-j} + u_t$$

The $\{b_j^1\}$ lie in some restricted space, for example, the MA representations of an ARMA(p,q). In population, OLS (or maximum likelihood with normal iid errors) picks b_j^1 to minimize

$$\int_{-\pi}^{\pi} \mid b^0(e^{-i\omega}) - b^1(e^{-i\omega}) \mid^2 S_x(\omega) d\omega.$$

Proof: (left as a problem)

OLS tries to match an average of $b^0(e^{-i\omega}) - b^1(e^{-i\omega})$ over the entire spectrum from $-\pi$ to π. How hard it tries to match them depends on the spectral density of x. Thus, OLS will sacrifice accuracy of the estimated $b(L)$ is a small frequency window, and/or a window in which $S_x(\omega)$ is small, to get better accuracy in a large window, and/or a window in which $S_x(\omega)$ is large.

8.4 Relation between Spectral, Wold, and Autocovariance representations

We have discussed three different fundamental representations for time series processes: the autocovariance function, the Wold $MA(\infty)$ and the spectral density. The following diagram summarizes the relation between the three representations.

$$\begin{array}{ccccc}
 & & \text{Autoregresions} & & \\
 & & \downarrow & & \\
 & & \text{Wold } MA(\infty) & & \\
 & \nearrow & x_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}, & \nwarrow & \\
\text{Infer AR} & \swarrow & \theta(L) \text{ invertible} & \searrow & S(\omega) = \theta(e^{-i\omega})\theta(e^{i\omega})\sigma_\epsilon^2 \\
\gamma_k = \sum_{j=0}^{\infty} \theta_j \theta_{j+k} \sigma_\epsilon^2 & & & & \sigma_\epsilon^2 = e^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln S(\omega) d\omega} \\
\text{Autocovariance} & \rightarrow & S(\omega) = \sum_{k=-\infty}^{\infty} e^{-ik\omega}\gamma_k & \rightarrow & \text{Spectral density } S(\omega) \\
\gamma_k = E(x_t x_{t-k}) & \leftarrow & \gamma_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-k\omega} S(\omega) d\omega & \leftarrow &
\end{array}$$

Each of the three representations can be estimated directly: we find the Wold MA by running autoregressions and simulating the impulse-response function; we can find the autocovariances by finding sample autocovariances. The next chapter discusses spectral density estimates. From each representation we can derive the others, as shown.

The only procedure we have not discussed so far is how to go from spectral density to Wold. The central insight to doing this is to realize that the roots of the spectral density function are the same as the roots of the Wold moving average, plus the inverses of those roots. If $x_t = \theta(L)\epsilon_t$ =const.$(L-\lambda_1)(L-\lambda_2)...$, so that $\theta(z)$ =const.$(z-\lambda_1)(z-\lambda_2)..$, λ and $\lambda_1, \lambda_2, ...$are roots, then $S_x(z)$ =const.$^2(z-\lambda_1)(z-\lambda_2)...(z^{-1}-\lambda_1)(z^{-1}-\lambda_2)...$so that $\lambda_1, \lambda_2, ...$ *and* $\lambda_1^{-1}, \lambda_2^{-1}...$ are roots. Thus, find roots of the spectral density, those outside the unit circle are the roots of the Wold lag polynomial. To find σ_ϵ^2, either make sure that the integral of the spectral density equals the variance of the Wold representation, or use the direct formula given above.

Chapter 9

Spectral analysis in finite samples

So far, we have been characterizing population moments of a time series. It's also useful to think how spectral representations work with a finite sample of data in hand. Among other things, we need to do this in order to think about how to estimate spectral densities.

9.1 Finite Fourier transforms

9.1.1 Definitions

For a variety of purposes it is convenient to think of finite-sample counterparts to the population quantities we've introduced so far. To start off, think of fourier transforming the *data* x_t rather than the autocovariance function. Thus, let

$$x_\omega = \frac{1}{T^{1/2}} \sum_{t=1}^{T} e^{-i\omega t} x_t$$

Where T = sample size. We can calculate x_ω for arbitrary ω. However, I will mostly calculate it for T $\omega' s$, spread evenly about the unit circle. When the

$\omega's$ are spread evenly in this way, there is an inverse finite fourier transform,

$$x_t = \frac{1}{T^{1/2}} \sum_{\omega} e^{i\omega t} x_\omega$$

Proof: Just like the proof of the infinite size inverse transform.

$$\frac{1}{T^{1/2}} \sum_{\omega} e^{i\omega t} x_\omega = \frac{1}{T^{1/2}} \sum_{\omega} e^{i\omega t} \frac{1}{T^{1/2}} \sum_{j=1}^{T} e^{-i\omega j} x_j = \frac{1}{T} \sum_{j=1}^{T} x_j \sum_{\omega} e^{i\omega(t-j)}.$$

$$\sum_{\omega} e^{i\omega(t-j)} = \begin{cases} T & \text{if } t-j=0 \\ 0 \text{ if } t-j \neq 0 \end{cases}.$$

□

It is handy to put this transformation in matrix notation. Let

$$W = T^{1/2} \begin{bmatrix} e^{-i\omega_1} & e^{-2i\omega_2} & \dots \\ e^{-i\omega_2} & e^{-2i\omega_2} & \dots \\ \dots & & \ddots \end{bmatrix}; \ \mathbf{x}_t = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}; \ \mathbf{x}_\omega = \begin{bmatrix} x_{\omega 1} \\ x_{\omega 2} \\ \vdots \end{bmatrix}$$

Then, we can write the fourier transform and its inverse as

$$\mathbf{x}_\omega = W\mathbf{x}_t, \quad \mathbf{x}_t = W^T \mathbf{x}_\omega$$

where $W^T = (W')^* = (W^*)'$ denotes "complex conjugate and transpose". Note $W^T W = WW^T = I$, i.e., fourier transforming and then inverse transforming gets you back where you started. Matrices W with this property are called unitary matrices.

9.2 Band spectrum regression

9.2.1 Motivation

We very frequently filter series before we look at relations between them—we detrend them, deseasonlaize them, etc. Implicitly, these procedures mean

that we think relations are different at different frequencies, and we want to isolate the frequencies of interest before we look at the relation, rather than build an encompassing model of the relation between series at all frequencies. "Band spectrum regression" is a technique designed to think about these situations. I think that it is not so useful in itself, but I find it a very useful way of thinking about what we're doing when we're filtering, and how relations that differ across frequencies influence much time series work.

Here are some examples of situations in which we think relations vary across frequencies, and for which filtered data have been used:

Example 1: Kydland and Prescott.

Kydland and Prescott simulate a model that gives rise to *stationary* time series patterns for GNP, consumption, etc. However, the GNP, consumption, etc. data are not stationary: not only do they trend upwards but they include interesting patterns at long horizons, such as the "productivity slowdown" of the $70's$, as well as recessions. Kydland and Prescott only designed their model to capture the business cycle correlations of the time series; hence they filtered the series with the Hodrick-Prescott filter (essentially a high-pass filter) before computing covariances.

Example 2: Labor supply.

The Lucas-Prescott model of labor supply makes a distinction between "permanent" and "transitory" changes in wage rates. A transitory change in wage rates has no income effect, and so induces a large increase in labor supply (intertemporal substitution). A permanent increase has an income effect, and so induces a much smaller increase in labor supply. This model might (yes, I know this is static thinking in a dynamic model – this is only a suggestive example!) make time series predictions as illustrated in figure 9.1:

As you can see, we might expect a different relation between wages and labor supply at business cycle frequencies vs. longer horizons.

Example 3: Money supply and interest rates

Conventional textbooks tell you that money growth increases have a short run negative impact on real and hence nominal interest rates, but a long run, one-for-one impact on inflation and hence nominal interest rates. The general view is that the time-series relation between money and interest rates looks

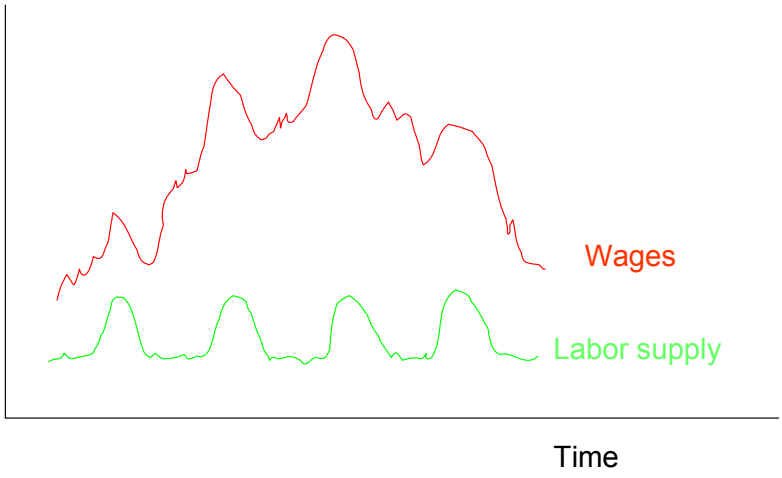

Figure 9.1: Time series of wages and labor supply

something like figure 9.2

Again, the relation between money growth and nominal interest rates depends on the frequency at which you look at it. At high frequencies we expect a negative relation, at low frequencies we expect a positive relation. (I wrote my PhD thesis on this idea, see "The Return of the Liquidity Effect: A Study of the Short Run Relation Between Money Growth and Interest Rates" Journal of Business and Economic Statistics 7 (January 1989) 75-83.)

Example 4: Seasonal adjustment.

Most of the time we use seasonally adjusted data. Seasonal adjustment is a band pass filter that attenuates or eliminates seasonal frequencies. This procedure must reflect a belief that there is a different relation between variables at seasonal and nonseasonal frequencies. For example, there is a tremendous seasonal in nondurable and services consumption growth; the use of seasonally adjusted consumption data in asset pricing reflects a belief that consumers *do not* try to smooth these in asset markets.

WARNING: Though filtering is very common in these situations, it is probably "wrong". The "Right" thing to do is generally specify the *whole* model, at long, short, seasonal, and nonseasonal frequencies, and estimate

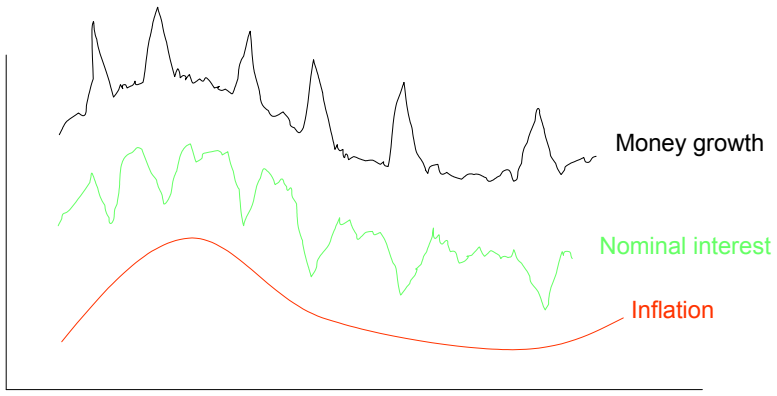

Figure 9.2: Money and interest rates

the entire dynamic system. The underlying economics typically does *not* separate by frequency. An optimizing agent facing a shock with a seasonal and nonseasonal component will set his control variable at *non*seasonal frequencies in a way that reflects the seasonals. "Growth theory" and "business cycle theory" may each make predictions at each others' frequencies, so one cannot really come up with separate theories to explain separate bands of the spectrum. The proper distinction between shocks may be "expected" and "unexpected" rather than "low frequency" vs. "high frequency". Nonetheless, removing frequency bands because you have a model that "only applies at certain frequencies" is very common, even if that characterization of the model's predictions is not accurate.

9.2.2 Band spectrum procedure

Suppose y and x satisfy the OLS assumptions,

$$\mathbf{y}_t = \mathbf{x}_t\beta + \boldsymbol{\epsilon}_t, \quad E(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t') = \sigma^2 I, \quad E(\mathbf{x}_t\boldsymbol{\epsilon}_t') = 0$$

Since W is a unitary matrix, the fourier transformed versions also satisfy the OLS assumptions:

$$W\mathbf{y}_t = W\mathbf{x}_t\beta + W\boldsymbol{\epsilon}_t$$

or

$$\mathbf{y}_\omega = \mathbf{x}_\omega \beta + \boldsymbol{\epsilon}_\omega$$

where

$$E(\boldsymbol{\epsilon}_\omega \boldsymbol{\epsilon}_\omega^T) = E(W \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t' W^T) = \sigma^2 I,$$
$$E(\mathbf{x}_\omega \boldsymbol{\epsilon}_\omega') = E(W \mathbf{x}_t \boldsymbol{\epsilon}_t' W^T) = 0.$$

Thus, why not run OLS using the frequency transformed data,

$$\hat{\beta} = (\mathbf{x}_\omega^T \mathbf{x}_\omega)^{-1} (\mathbf{x}_\omega^T \mathbf{y}_\omega)?$$

In fact, this β is *numerically identical* to the usual OLS β :

$$(\mathbf{x}_\omega^T \mathbf{x}_\omega)^{-1} (\mathbf{x}_\omega^T \mathbf{y}_\omega) = (\mathbf{x}_t' W^T W \mathbf{x}_t)^{-1} (\mathbf{x}_t' W^T W \mathbf{y}_t) = (\mathbf{x}_t' \mathbf{x}_t)^{-1} (\mathbf{x}_t' \mathbf{y}_t).$$

So far, pretty boring. But now, recall what we do with OLS in the presence of a "structural shift". The most common example is war, where you might think that behavioral relationships are different for the data points 1941-1945. (This usually means that you haven't specified your behavior at a deep enough level.) Also, it is common to estimate different relationships in different "policy regimes", such as pre-and post 1971 when the system of fixed exchange rates died, or in 1979-1982 and outside that period, when the Fed (supposedly) followed a nonborrowed reserve rather than interest rate targeting procedure. Thus, suppose you thought

$$y_t = x_t \beta_1 + \epsilon_t \text{ (period A)}$$

$$y_t = x_t \beta_2 + \epsilon_t \text{ (period B)}$$

What do you do? You estimate separate regressions, or you drop from your sample the time periods in which you think β changed from the "true value" that you want to estimate.

The equivalence of the regular and frequency domain OLS assumptions shows you that OLS assumes that β is the same *across frequencies*, as it assumes that β is the same over time. Band spectrum regression becomes useful when you don't think this is true: when you think that the relationship you want to investigate only holds at certain *frequencies.*

The solution to this problem is obvious: drop certain *frequencies* from the frequency transformed regression, i.e. transform the data and then estimate

separate β_1 and β_2; $y_\omega = x_\omega\beta_1 + \epsilon_\omega$ in frequency interval A and $y_\omega = x_\omega\beta_2 + \epsilon_\omega$ in frequency interval B. If you believe the model holds only at certain frequencies, just run $y_\omega = x_\omega\beta + \epsilon_\omega\omega$ in given frequency band.

One way to formalize these operations is to let A be a "selector matrix" that picks out desired frequencies. It has ones on the diagonal corresponding to the "good" frequencies that you want to keep, and zeros elsewhere. Then, if you run

$$A\mathbf{y}_\omega = A\mathbf{x}_\omega\beta + A\boldsymbol{\epsilon}_\omega.$$

Only the "good" frequencies are included. The β you get is

$$\hat{\beta} = (\mathbf{x}_\omega^T A A\mathbf{x}_\omega)^{-1}(\mathbf{x}_\omega^T A A\mathbf{y}_\omega).$$

As you might suspect, *filtering* the data to remove the unwanted frequencies and then running OLS (in time domain) on the filtered data is equivalent to this procedure. To see this, invert the band spectrum regression to time-domain by running

$$W^T A y_\omega = W^T A\mathbf{x}_\omega\beta + W^T A\boldsymbol{\epsilon}_\omega$$

This regression gives numerically identical results. Writing the definition of x_ω, this is in turn equivalent to

$$W^T AW\mathbf{y}_t = W^T AW\mathbf{x}_t\beta + W^T AW\boldsymbol{\epsilon}_t$$

So, what's $W^T AW$? It takes time domain to frequency domain, drops a band of frequencies, and reverts to time domain. Thus it's a *band pass filter!*

There is one warning: If we drop k frequencies in the band-spectrum regression, OLS on the frequency domain data will pick up the fact that only $T - k$ degrees of freedom are left. However, there are still T time-domain observations, so you need to correct standard errors for the lost degrees of freedom. Alternatively, note that ϵ_t is serially uncorrelated, $W^T AW_t$ is serially correlated, so you have to correct for serial correlation of the error in inference.

Of course, it is unlikely that your priors are this sharp–that certain frequencies belong, and certain others don't. More likely, you want to give more weight to some frequencies and less weight to others, so you use a filter with a smoother response than the bandpass filter. Nonetheless, I find it useful to think about the rationale of this kind of procedure with the band-pass result in mind.

9.3 Cramér or Spectral representation

The spectral or Cramér representation makes precise the sense in which the spectral density is a decomposition of the variance of a series into the variance of orthogonal components at each frequency. To do this right (in population) we need to study Brownian motion, which we haven't gotten to yet. But we can make a start with the finite fourier transform and a finite data set.

The inverse fourier transform of the data is

$$x_t = \frac{1}{T^{1/2}} \sum_{\omega} e^{i\omega t} x_\omega = \frac{1}{T^{1/2}} \left(x_0 + 2 \sum_{\omega>0} \mid x_\omega \mid \cos(\omega t + \phi_\omega) \right)$$

where $x_\omega = \mid x_\omega \mid e^{i\phi_\omega}$.Thus it represents the time series x_t as a sum of cosine waves of different frequencies and phase shifts.

One way to think of the inverse fourier transform is that we can draw random variables $\{x_\omega\}$ at the beginning of time, and then let x_t evolve according to the inverse transform. It looks like this procedure produces a deterministic time series x_t, but that isn't true. "Deterministic" means that x_t is perfectly predictable given the history of x, *not* given $\{x_\omega\}$. If you have k $x_t's$ you can only figure out k $x_w's$.

So what are the statistical properties of these random variables x_ω? Since x_t is mean zero, $E(x_\omega) = 0$. The variance and covariances are

$$\lim_{T\to\infty} E(x_\omega x_\lambda^*) = \begin{cases} S_x(\omega) \text{ if } \omega = \lambda \\ 0 \text{ if } \omega \neq \lambda \end{cases}$$

Proof:

$$E(x_\omega x_\lambda^*) = E\frac{1}{T} \sum_t e^{-i\omega t} x_t \sum_j e^{i\lambda j} x_j = E\frac{1}{T} \sum_{t,j} e^{i\lambda j} e^{-i\omega t} x_t x_{j.}$$

setting $j = t - k$, thus $k = j - t$,

$$= E\frac{1}{T} \sum_{t,k} e^{i\lambda(t-k)} e^{-i\omega t} x_t x_{t-k} = \sum_k e^{-i\lambda k} \frac{1}{T} \sum_t e^{i(\lambda-\omega)t} \gamma_k(x)$$

$$= \sum_k e^{-i\lambda k} \gamma_k(x) \frac{1}{T} \sum_t e^{i(\lambda-\omega)t}.$$

If $\omega = \lambda$, the last term is T, so we get

$$\lim_{T\to\infty} E(x_\omega x_\omega^*) = \sum_k e^{-i\omega k}\gamma_k = S_x(\omega)$$

If $\omega \neq \lambda$, the last sum is zero, so

$$\lim_{T\to\infty} E(x_\omega x_\lambda^*) = 0.$$

□

Though it looks like finite-sample versions of these statements also go through, I was deliberately vague about the sum indices. When these do not go from $-\infty$ to ∞, there are small-sample biases.

Thus, when we think of the time series by picking $\{x_\omega\}$ and then generating out $\{x_t\}$ by the inversion formula, the x_ω are *uncorrelated.* However, they are *heteroskedastic*, since the variance of x_ω is the spectral density of x, which varies over ω.

The Cramér or spectral representation takes these ideas to their limit. It is

$$x_t = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega t} dz(\omega)$$

where

$$E(dz(\omega)dz(\lambda)*) = \begin{cases} S_x(\omega)d\omega \text{ for } \omega = \lambda \\ 0 \text{ for } \omega \neq \lambda \end{cases}$$

The dz are increments of a Brownian motion (A little more precisely, the random variable

$$Z(\omega) = \int_{-\pi}^{\omega} dz(\lambda)$$

has uncorrelated increments: $E[Z(.3) - Z(.2))(Z(.2) - Z(.1))] = 0.$) Again, the idea is that we draw this Brownian motion from $-\pi$ to π at the beginning of time, and then fill out the x_t. As before, the past history of x is not sufficient to infer the whole Z path, so x is still indeterministic.

This is what we really mean by "the component at frequency ω" of a non-deterministic series. As you can see, being really precise about it means we have to study Brownian motions, which I'll put off for a bit.

9.4 Estimating spectral densities

We will study a couple of approaches to estimating spectral densities. A reference for most of this discussion is Anderson (1971) p. 501 ff.

9.4.1 Fourier transform sample covariances

The obvious way to estimate the spectral density is to use sample counterparts to the definition. Start with an estimate of the autocovariance function.

$$\hat{\gamma}_k = \frac{1}{T-k}\sum_{t=k+1}^{T} x_t x_{t-k} \text{ or } \hat{\gamma}_k = \frac{1}{T}\sum_{t=k+1}^{T} x_t x_{t-k}.$$

(I'm assuming the x's have mean zero. If not, take out sample means first. This has small effects on what follows.) Both estimates are consistent. The first is unbiased; the second produces positive definite autocovariance sequences in any sample. The first does *not* (necessarily) produce positive definite sequences, and hence positive spectral densities. One can use either; I'll use the second.

We can construct spectral density estimates by fourier transforming these autocovariances:

$$\hat{S}(\omega) = \sum_{k=-(T-1)}^{T-1} e^{-i\omega k}\hat{\gamma}_k$$

9.4.2 Sample spectral density

A second approach is suggested by our definition of the finite fourier transform. Since

$$\lim_{T\to\infty} E(x_\omega x_\omega^*) = S(\omega),$$

why not estimate $S(\omega)$ by the *sample spectral density* [1] $I(\omega)$:

$$\hat{S}(\omega) = I(\omega) = x_\omega x_\omega^* = \frac{1}{T}\left(\sum_{t=1}^{T} e^{-i\omega t} x_t\right)\left(\sum_{t=1}^{T} e^{i\omega t} x_t\right) = \frac{1}{T}\left|\sum_{t=1}^{T} e^{-i\omega t} x_t\right|^2$$

9.4.3 Relation between transformed autocovariances and sample density

The fourier transform of the sample autocovariance is numerically identical to the sample spectral density!

Proof:

$$\frac{1}{T}\left(\sum_{t=1}^{T} e^{-i\omega t} x_t\right)\left(\sum_{t=1}^{T} e^{i\omega t} x_t\right) = \frac{1}{T}\left(\sum_{t=1}^{T}\sum_{=1}^{T} e^{i\omega(j-t)} x_t x_j\right)$$

Let $k = t - j$, so $j = t - k$.

$$= \sum_{k=-(T-1)}^{T-1} e^{-i\omega k}\,\frac{1}{T}\sum_{t=|k|+1}^{T} x_t x_{t-|k|} = \sum_{k=-(T-1)}^{T-1} e^{-i\omega k}\hat{\gamma}_k$$

(To check the limits on the sums, verify that each $x_t x_j$ is still counted once. It may help to plot each combination on a t vs. j grid, and verify that the second system indeed gets each one once. You'll also have to use $x_1 x_{1-(-3)} = x_4 x_1$, etc. when $k < 0$.)

□

Thus, the sample spectral density is the fourier transform of the sample autocovariances,

$$I(\omega) = \sum_{k=-(T-1)}^{T-1} e^{-i\omega k}\hat{\gamma}_k = \sum_{k=-\infty}^{\infty} e^{-i\omega k}\hat{\gamma}_k$$

[1] This quantity is also sometimes called the *periodogram*. Many treatements of the periodogram divide by an extra T. The original periodogram was designed to ferret out pure sine or cosine wave components, which require dividing by an extra T for the periodogram to stay stable as $T \to \infty$. When there are only non-deterministic components, we divide only by one T. I use "sample spectral density" to distinguish the two possibilities.

where the latter equality holds by the convention that $\hat{\gamma}_k = 0$ for $k > T-1$. Applying the inverse fourier transform, we know that

$$\hat{\gamma}_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\omega k} I(\omega) d\omega.$$

Thus, the sample autocovariances (using $1/T$) and sample spectral density have the same relation to each other as the population autocovariance and spectral density.

$$\gamma_k \Leftrightarrow S(\omega)$$

just like

$$\hat{\gamma}_k \Leftrightarrow I(\omega).$$

The sample spectral density is completely determined by its values at T different frequencies, which you might as well take evenly spaced. (Analogously, we defined x_ω for T different ω above, and showed that they carried all the information in a sample of length T.) To see this, note that there are only T autocovariances(including the variance). Thus, we can recover the T autocovariances from T values of the spectral density, and construct the spectral density at any new frequency from the T autocovariances.

As a result, you might suspect that there is also a finite fourier transform relation between $I(\omega)$ and $\hat{\gamma}(\omega)$, and there is,

$$\hat{\gamma}_k = \frac{1}{T} \sum_{\omega} e^{i\omega k} I(\omega).$$

Proof:

$$\frac{1}{T} \sum_{\omega} e^{i\omega k} I(\omega) = \frac{1}{T} \sum_{\omega} e^{i\omega k} \sum_{j} e^{-i\omega j} \hat{\gamma}_j = \frac{1}{T} \sum_{j} \sum_{\omega} e^{i\omega(k-j)} \hat{\gamma}_j$$

$$= \frac{1}{T} \sum_{j} T\delta(k-j)\hat{\gamma}_j = \hat{\gamma}_k.$$

(Since we did not divide by $1/T^{1/2}$ going from $\hat{\gamma}$ to I, we divide by T going back.)

□

9.4.4 Asymptotic distribution of sample spectral density

Here are some facts about the asymptotic distribution of these spectral density estimates:

$$\lim_{T\to\infty} E(I(\omega)) = S(\omega)$$

$$\lim_{T\to\infty} \text{var}(I(\omega)) = \begin{cases} 2S^2(0) \text{ for } \omega = 0 \\ S^2(\omega) \text{ for } \omega \neq 0 \end{cases}$$

$$\lim_{T\to\infty} cov(I(\omega), I(\lambda)) = 0 \text{ for } \mid \omega \mid \neq \mid \lambda \mid$$

$$2I(\omega)/S(\omega) \to \chi^2_2$$

Note that the variance of the sample spectral density does not go to zero, as the variance of most estimators (not yet scaled by $T^{1/2}$) does. The sample spectral density is not consistent, since its distribution does not collapse around the true value. This variance problem isn't just a problem of asymptotic mumbo-jumbo. Plots of the sample spectral density of even very smooth processes show a lot of jumpiness.

There are two ways to understand this inconsistency intuitively. First, recall that $I(\omega) = x_\omega x'_\omega$. Thus, $I(\omega)$ represents one data point's worth of information. To get consistent estimates of anything, you need to include increasing numbers of data points. Second, look at the definition as the sum of sample covariances; the high autocovariances are on average weighted just as much as the low autocovariances. But the last few autocovariances are bad estimates: $\gamma_{T-1} = (x_T x_1)/T$ no matter how big the sample. Thus $I(\omega)$ always contains estimates with very high variance.

9.4.5 Smoothed periodogram estimates

Here is one solution to this problem: Instead of estimating $S(\omega)$ by $I(\omega)$, average $I(\omega)$ over several nearby ω's. Since $S(\omega)$ is a smooth function and adjacent $I(\omega)$ are uncorrelated in large samples, this operation reduces variance without adding too much bias. Of course, how many nearby $I(\omega)$ to include will be a tricky issue, a trade-off between variance and bias. As $T \to \infty$, promise to slowly reduce the range of averaged ω's. You want to

reduce it so that as $T \to \infty$ you are in fact only estimating $S(\omega)$, but you want to reduce it slowly so that larger and larger numbers of x'_ω enter the band as $T \to \infty$.

Precisely, consider a smoothed periodogram estimator $\hat{S}(\omega) = \int_{-\pi}^{\pi} h(\lambda - \omega) I(\lambda) d\lambda$ or $\hat{S}(\omega) = \sum_{\lambda_i} h(\lambda_i - \omega) I(\lambda_i)$ where h is a moving average function as shown in figure 9.3 To make the estimate asymptotically unbiased and its

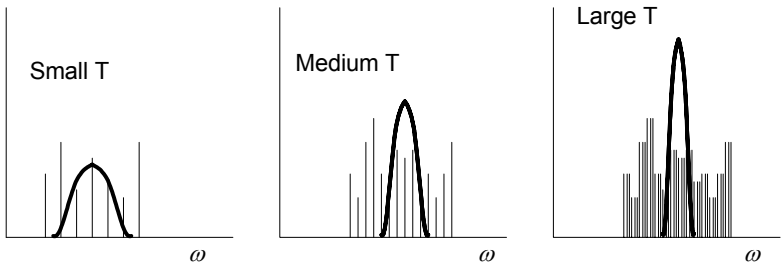

Figure 9.3: Smoothed periodogram estimate as sample size increases. Note that the window size decreases with sample size, but the number of frequencies in the window increases.

variance go to zero (and also consistent) you want to make promises as shown in the figure: Since more and more periodogram ordinates enter as $T \to \infty$, the variance goes to zero. But since the size of the window goes to zero as $T \to \infty$, the bias goes to zero as well.

9.4.6 Weighted covariance estimates

Viewing the problem as the inclusion of poorly measured, high-order autocovariances, we might try to estimate the spectral density by lowering the weight on the high autocovariances. Thus, consider

$$\hat{S}(\omega) = \sum_{k=-(T-1)}^{T-1} e^{-i\omega k} g(k) \hat{\gamma}_k$$

where $g(k)$ is a function as shown in figure 9.4 For example, the Bartlett

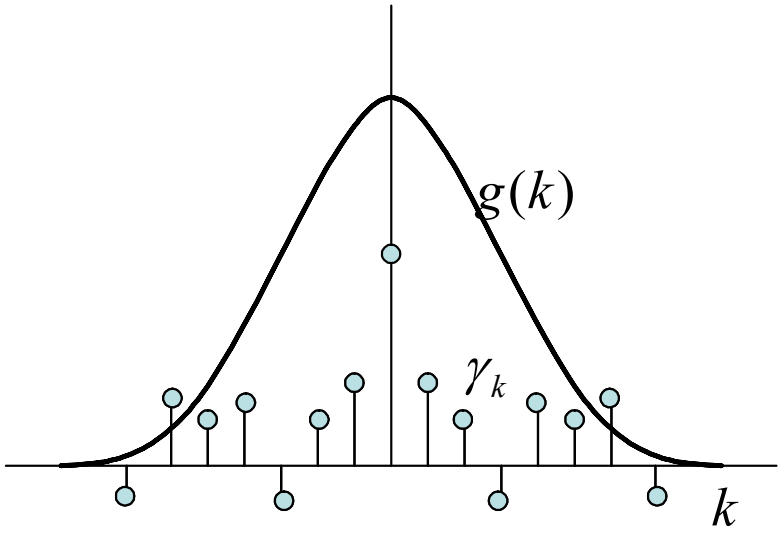

Figure 9.4: Covariance weighting function

window has a triangular shape,

$$\hat{S}(\omega) = \sum_{k=-r}^{r} e^{-i\omega k}(1 - \frac{\mid k \mid}{r})\hat{\gamma}_k$$

Again, there is a trade off between bias and variance. Down weighting the high autocovariances improves variance but introduces bias. Again, one can make promises to make this go away in large samples. Here, one wants to promise that $g(k) \to 1$ as $T \to \infty$ to eliminate bias. Thus, for example, an appropriate set of Bartlett promises is $r \to \infty$ and $r/T \to 0$ as $T \to \infty$; this can be achieved with $r \sim T^{1/2}$.

9.4.7 Relation between weighted covariance and smoothed periodogram estimates

Not surprisingly, these two estimates are related. Every weighted covariance estimate is equivalent to a smoothed periodogram estimate, where the

smoothing function is proportional the fourier transform of the weighting function; and vice-versa.

Proof:

$$\hat{S}(\omega) = \sum_k e^{-i\omega k} g(k) \hat{\gamma}_k = \sum_k e^{-i\omega k} g(k) \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\nu k} I(\lambda) d\lambda =$$

$$\int_{-\pi}^{\pi} \frac{1}{2\pi} \left(\sum_k e^{-i(\omega-\lambda)k} g(k) \right) I(\lambda) d\nu = \int_{-\pi}^{\pi} h(\omega - \lambda) I(\lambda) d\nu$$

Similarly, one can use the finite fourier transform at T frequencies,

$$\hat{S}(\omega) = \sum_k e^{-i\omega k} g(k) \hat{\gamma}_k = \sum_k e^{-i\omega k} g(k) \frac{1}{T} \sum_\nu e^{i\lambda k} I(\lambda) =$$

$$\sum_\nu \left(\frac{1}{T} \sum_k e^{-i(\omega-\lambda)k} g(k) \right) I(\lambda) = \sum_\nu h(\omega - \lambda) I(\lambda)$$

□

9.4.8 Variance of filtered data estimates

A last, equivalent approach is to filter the data with a filter that isolates components in a frequency window, and then take the variance of the filtered series. After all, spectral densities are supposed to be the variance of components at various frequencies. With a suitably chosen filter, this approach is equivalent to weighted periodogram or covariance estimates. Thus, let

$$x_t^f = F(L) x_t$$

Hence,

$$var(x_t^f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mid F(e^{-i\lambda}) \mid^2 S_x(\lambda) d\lambda$$

So all you have to do is pick $F(L)$ so that

$$\frac{1}{2\pi} \mid F(e^{-i\lambda}) \mid^2 = h(\omega - \lambda)$$

Variance ratio estimates of the spectral density at frequency zero are examples of this procedure.

9.4.9 Spectral density implied by ARMA models

All of the above estimates are non-parametric. One reason I did them is to introduce you to increasingly fashionable non-parametric estimation. Of course, one can estimate spectral densities parametrically as well. Fit an AR or ARMA process, find its Wold representation

$$x_t = \theta(L)\epsilon_t$$

and then

$$\hat{S}(\omega) = \mid \theta(e^{-i\omega}) \mid^2 s_\epsilon^2.$$

How well this works depends on the quality of the parametric approximation. Since OLS tries to match the whole frequency range, this technique may sacrifice accuracy in small windows that you might be interested in, to get more accuracy in larger windows you might not care about.

9.4.10 Asymptotic distribution of spectral estimates

The asymptotic distribution of smoothed periodogram / weighted covariance estimates obviously will depend on the shape of the window / weighting function, and the promises you make about how that shape varies as $T \to \infty$. When you want to use one, look it up.

Chapter 10

Unit Roots

10.1 Random Walks

The basic random walk is

$$x_t = x_{t-1} + \epsilon_t;\ E_{t-1}(\epsilon_t) = 0$$

Note the property

$$E_t(x_{t+1}) = x_t.$$

As a result of this property, random walks are popular models for asset prices.

Random walks have a number of interesting properties.

1) The impulse-response function of a random walk is one at all horizons. The impulse-response function of stationary processes dies out eventually.

2) The forecast variance of the random walk grows linearly with the forecast horizon

$$var(x_{t+k} \mid x_t) = var(x_{t+k} - x_t) = k\sigma_\epsilon^2.$$

The forecast error variance of a stationary series approaches a constant, the unconditional variance of that series. Of course, the variance of the random walk is infinite, so in a sense, the same is true.

3) The autocovariances of a random walk aren't defined, strictly speaking. However, you can think of the limit of an $AR(1)$, $x_t = \phi x_{t-1} + \epsilon_t$ as the

autoregression parameter ϕ goes to 1. Then, for a random walk,

$$\rho_j = 1 \text{ for all } j.$$

Thus, a sign of a random walk is that all the estimated autocorrelations are near one, or die out "too slowly".

4) The spectral density (normalized by the variance) of the $AR(1)$ is

$$f(\omega) = [(1 - \phi e^{-i\omega})(1 - \phi e^{i\omega})]^{-1} = \frac{1}{1 + \phi^2 - 2\phi\cos(\omega)}.$$

In the limit $\phi \to 1$ we get

$$f(\omega) = \frac{1}{2(1 - \cos(\omega))}.$$

As $\omega \to 0, S(\omega) \to \infty$. Thus, *the variance of a random walk is primarily due to low-frequency components.* The signature of a random walk is its tendency to wander around at low frequencies.

10.2 Motivations for unit roots

10.2.1 Stochastic trends

One reason macroeconomists got interested in unit roots is the question of how to represent trends in time series. Until the late $70's$ it was common to simply fit a linear trend to log GNP (by OLS), and then define the stochastic part of the time series as deviations from this trend. This procedure let to problems when it seemed like the "trend", "potential" etc. GNP growth rate slowed down. Since the slowdown was not foreseen it was hard to go about business as usual with more complex deterministic trends, such as polynomials. Instead, macroeconomists got interested in *stochastic* trends, and random-walk type processes give a convenient representation of such trends since they wander around at low frequencies.

10.2.2 Permanence of shocks

Once upon a time, macroeconomists routinely detrended data, and regarded business cycles as the (per-force) stationary deviation about that trend. It was wisely accepted that business cycles were short-run (no more than a few years, at most) deviations from trend. However, macroeconomists have recently questioned this time-honored assumption, and have started to wonder whether shocks to GNP might not more closely resemble the permanent shocks of a random walk more than the transitory shocks of the old $AR(2)$ about a linear trend. In the first round of these tests, it was claimed that the permanence of shocks shed light on whether they were "real" ("technology") or "monetary", "supply" or "demand", etc. Now, it's fairly well accepted that nothing of direct importance hangs on the permanence of shocks, but it is still an interesting stylized fact.

At the same time, financial economists got interested in the question of whether stock returns are *less* than perfect random walks. It turns out that the same techniques that are good for quantifying how much GNP *does* behave like a random walk are useful for quantifying the extent to which stock prices *do not* exactly follow a random walk. Again, some authors once thought that these tests were convincing evidence about "efficient markets", but now most recognize that this is not the case.

10.2.3 Statistical issues

At the same time, the statistical issue mentioned above made it look likely that we could have mistaken time series with unit roots for trend stationary time series. This motivated Nelson and Plosser (1982) to test macroeconomic time series for unit roots. They found they could not reject unit roots in most time series. They interpreted this finding as evidence for technology shocks, though Campbell and Mankiw (1987) interpreted the exact same findings as evidence for long-lasting Keynesian stickiness. Whatever the interpretation, we became more aware of the possibility of long run movements in time-series.

Here are a few examples of the statistical issues. These are for motivation only at this stage; we'll look at distribution theory under unit roots in more detail in chapter x.

Distribution of $AR(1)$ estimates

Suppose a series is generated by a random walk

$$y_t = y_{t-1} + \epsilon_t.$$

You might test for a random walks by running

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

by OLS and testing whether $\phi = 1$. However, the assumptions underlying the usual asymptotic distribution theory for OLS estimates and test statistics are violated here, since$x'x/T$ does not converge in probability.

Dickey and Fuller looked at the distribution of this kind of test statistic and found that OLS estimates are biased down (towards stationarity) and OLS standard errors are tighter than the actual standard errors. Thus, it is possible that many series that you would have thought were stationary based on ols regressions were in fact generated by random walks.

Inappropriate detrending

Things get worse with a trend in the model. Suppose the real model is

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

Suppose you detrend by OLS, and then estimate an $AR(1)$, i.e., fit the model

$$y_t = bt + (1 - \phi L)^{-1}\epsilon_t$$

This model is equivalent to

$$(1 - \phi L)y_t = (1 - \phi L)bt + \epsilon_t = bt - \phi b(t-1) + \epsilon_t = \phi b + b(1-\phi)t + \epsilon_t$$

or

$$y_t = \alpha + \gamma t + \phi y_{t-1} + \epsilon_t,$$

so you could also directly run y on a time trend and lagged y.

It turns out that this case is even worse than the last one, in that $\hat{\phi}$ is biased downward and the OLS standard errors are misleading. Intuitively,

the random walk generates a lot of low-frequency movement. In a relatively small sample, the random walk is likely to drift up or down; that drift could well be (falsely) modeled by a linear (or nonlinear, "breaking" , etc.) trend. Claims to see trends in series that are really generated by random walks are the central fallacy behind much "technical analysis" of asset markets.

Spurious regression

Last, suppose two series are generated by independent random walks,

$$x_t = x_{t-1} + \epsilon_t$$

$$y_t = y_{t-1} + \delta_t \quad E(\epsilon_t \delta_s) = 0 \text{ for all } t, s$$

Now, suppose we run y_t on x_t by OLS,

$$y_t = \alpha + \beta x_t + \nu_t$$

Again, the assumptions behind the usual distribution theory are violated. In this case, you tend to see "significant" β more often than the OLS formulas say you should.

There are an enormous number of this kind of test in the literature. They generalize the random walk to allow serial correlation in the error (unit root processes; we'll study these below) and a wide variety of trends in both the data generating model and the estimated processes. Campbell and Perron (1991) give a good survey of this literature.

10.3 Unit root and stationary processes

The random walk is an extreme process. GNP and stock prices may follow processes that have some of these properties, but are not as extreme. One way to think of a more general process is as a random walk with a serially correlated disturbance

$$(1 - L)y_t = \mu + a(L)\epsilon_t$$

These are called unit root or difference stationary (DS) processes. In the simplest version a(L) = 1 the DS process is a *random walk with drift,*

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

Alternatively, we can consider a process such as log GNP to be stationary around a linear trend:

$$y_t = \mu t + b(L)\epsilon_t.$$

These are sometimes called trend-stationary (TS) processes.

The TS model can be considered as a special case of the DS model. If $a(L)$ contains a unit root, we can write the DS model as

$$y_t = \mu t + b(L)\epsilon_t \Rightarrow (1-L)y_t = \mu + (1-L)b(L)\epsilon_t = \mu + a(L)\epsilon_t$$

$$(a(L) = (1-L)b(L))$$

Thus if the TS model is correct, the DS model is still valid and stationary. However, it has a noninvertible unit MA root.

The DS model is a perfectly normal model for *differences.* We can think of unit roots as the study of the implications for *levels* of a process that is stationary in differences, rather than as a generalized random walk. For this reason, it is very important in what follows to keep track of whether you are thinking about the *level* of the process y_t or its first difference.

Next, we'll characterize some of the ways in which TS and DS processes differ from each other

10.3.1 Response to shocks

The impulse-response function [1] of the TS model is the same as before, $b_j =$ j period ahead response. For the DS model, a_j gives the response of the *difference* $(1-L)y_{t+j}$ to a shock at time t. The response of the *level* of log GNP y_{t+j} is the sum of the response of the differences,

response of $y_{t+j}$to shock at $t = y_t(-y_{t-1}) + (y_{t+1} - y_t) + .. + (y_{t+j} - y_{t+j-1})$

[1]Since these models have means and trends in them, we define the impulse-response function as $E_t(y_{t+j}) - E_{t-1}(y_{t+j})$ when there is a unit shock at time t. It doesn't make much sense to define the response to a unit shock when all previous values are zero!

$$= a_0 + a_1 + a_2 + \ldots + a_j$$

See figure 10.1 for a plot.

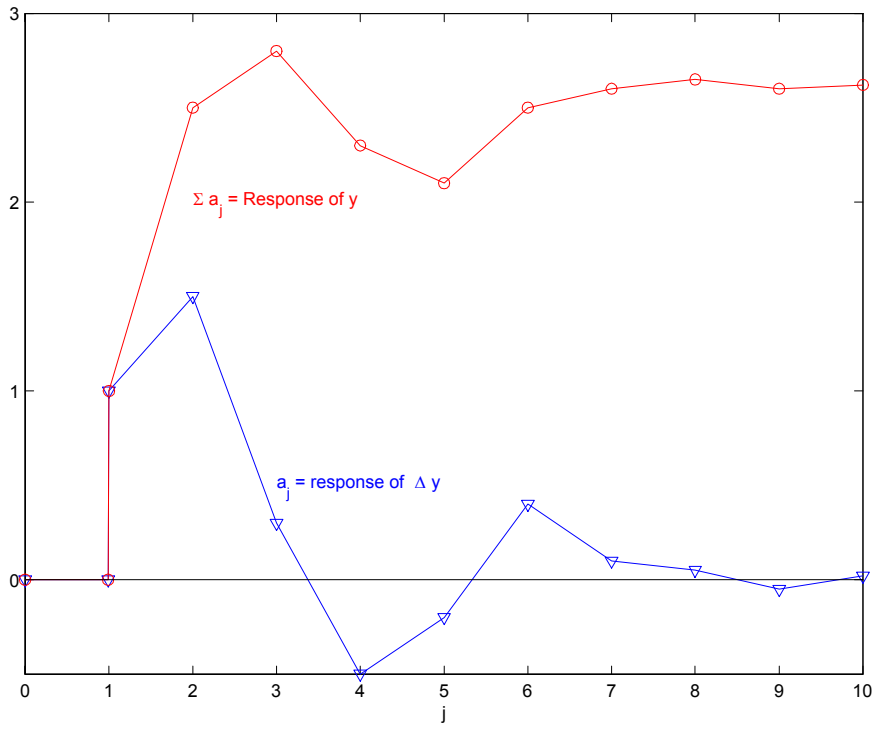

Figure 10.1: Response of differences and level for a series y_t with a unit root.

The limiting value of the impulse-response of the DS model is

$$\sum_{j=0}^{\infty} a_j = a(1).$$

Since the TS model is a special case in which $a(L) = (1 - L)b(L)$, $a(1) = 0$ if the TS model is true. As we will see this (and only this) is the feature that distinguishes TS from DS models once we allow arbitrary serial correlation, i.e. arbitrary a(L) structure.

What the DS model allows, which the random walk does not, is cases intermediate between stationary and random walk. Following a shock, the series could come back towards, but not all the way back to, its initial value. This behavior is sometimes called "long-horizon mean-reverting". For example, if stock prices are not pure random walks, they should decay over a period of years following a shock, or their behavior would suggest unexploited profit opportunities. Similarly, GNP might revert back to trend following a shock over the course of a business cycle, say a few years, rather than never (random walk) or in a quarter or two.

Figure 10.2 shows some possibilities.

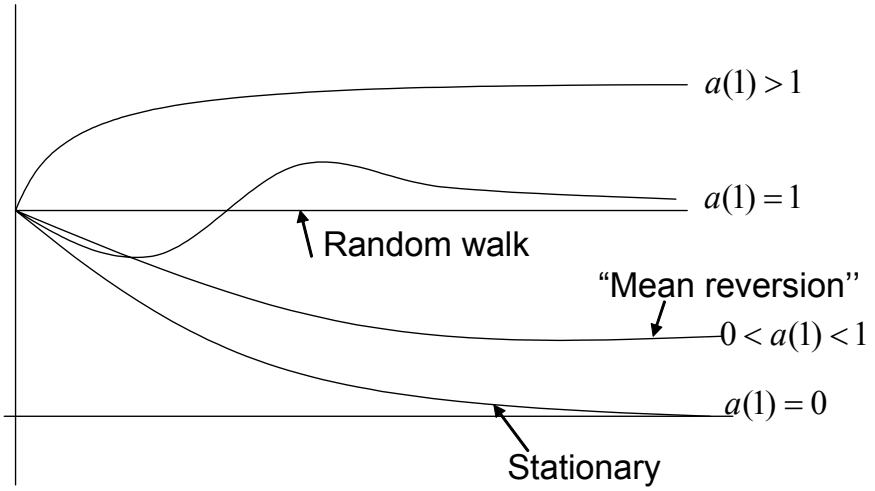

Figure 10.2: Impluse-response functions for different values of $a(1)$.

10.3.2 Spectral density

The spectral density of the DS process is

$$S_{(1-L)y_t}(\omega) = \mid a(e^{-i\omega}) \mid^2 \sigma_\epsilon^2.$$

The spectral density at frequency zero is $S_{(1-L)y_t}(0) = a(1)^2\sigma_\epsilon^2$ Thus, if $\mid a(1) \mid > 0$, then the spectral density at zero of Δy_t is greater than zero. If

a(1) = 0 (TS), then the spectral density of Δy_t at zero is equal to zero. Figure 10.3 shows some intermediate possibilities.

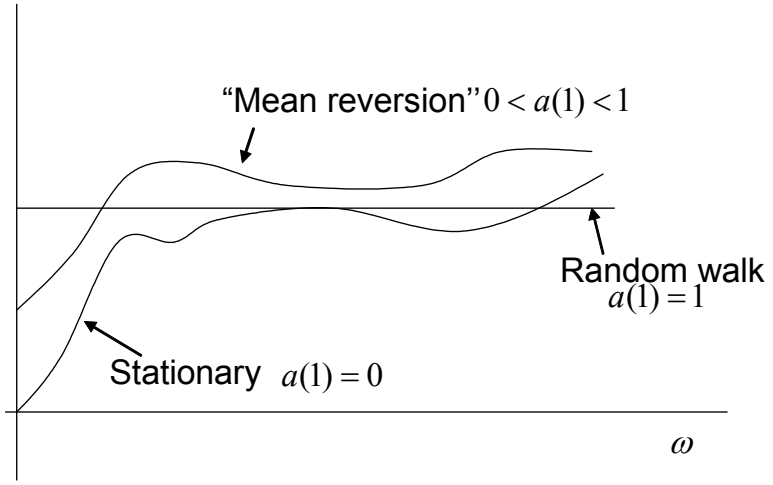

Figure 10.3: Spectral densities for different values of $a(1)$.

Here "long horizon mean reversion" shows up if the spectral density is high quite near zero, and then drops quickly.

10.3.3 Autocorrelation

The spectral density at zero is

$$S_{(1-L)y_t}(0) = \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j = \left(1 + 2\sum_{j=1}^{\infty}\rho_j\right)\gamma_0 = a(1)^2\sigma_\epsilon^2$$

Thus, if the process is DS, $\mid a(1) \mid > 0$, the sum of the autocorrelations is non-zero. If it is TS, $a(1) = 0$, the sum of the autocorrelations is zero. If it is a random walk, the sum of the autocorrelations is one; if it is mean reverting, then the sum of the autocorrelations is less than one.

The "long horizon" alternative shows up if there are many small negative autocorrelations at high lags that draw the process back towards its initial value following a shock.

10.3.4 Random walk components and stochastic trends

A second way to think of processes more general than a random walk, or intermediate between random walk and stationary is to think of combinations of random walks and stationary components. This is entirely equivalent to our definition of a DS process, as I'll show.

Fact: *every DS process can be written as a sum of a random walk and a stationary component.*

I need only exhibit one way of doing it. A decomposition with particularly nice properties is the

Beveridge-Nelson decomposition: If $(1-L)y_t = \mu + a(L)\epsilon_t$ then we can write

$$y_t = c_t + z_t$$

where

$$z_t = \mu + z_{t-1} + a(1)\epsilon_t$$

$$c_t = a^*(L)\epsilon_t; \;\; a_j^* = -\sum_{k=j+1}^{\infty} a_k.$$

Proof: The decomposition follows immediately from the algebraic fact that any lag polynomial $a(L)$ can be written as

$$a(L) = a(1) + (1-L)a^*(L); \;\; a_j^* = -\sum_{k=j+1}^{\infty} a_k.$$

Given this fact, we have $(1-L)y_t = \mu + a(1)\epsilon_t + (1-L)a^*(L)\epsilon_t$ $= (1-L)z_t + (1-L)c_t$, so we're done. To show the fact, just write it out:

$$\begin{array}{rccccc} a(1): & a_0 & +a_1 & +a_2 & +a_3 & .. \\ (1-L)a^*(L): & & -a_1 & -a_2 & -a_3 & .. \\ & & +a_1L & +a_2L & +a_3L & ... \\ & & & -a_2L & -a_3L & ... \\ & & & & ... & \end{array}$$

When you cancel terms, nothing but $a(L)$ remains.

□

There are many ways to decompose a unit root into stationary and random walk components. The B-N decomposition has a special property: the random walk component is a sensible definition of the "trend" in y_t. z_t is the limiting forecast of future y, or today's y plus all future expected changes in y. If GNP is forecasted to rise, GNP is "below trend" and vice-versa. Precisely,

$$z_t = \lim_{k\to\infty} E_t(y_{t+k} - k\mu) = y_t + \sum_{j=1}^{\infty}(E_t\Delta y_{t+j} - \mu)$$

The first definition is best illustrated graphically, as in figure 10.4

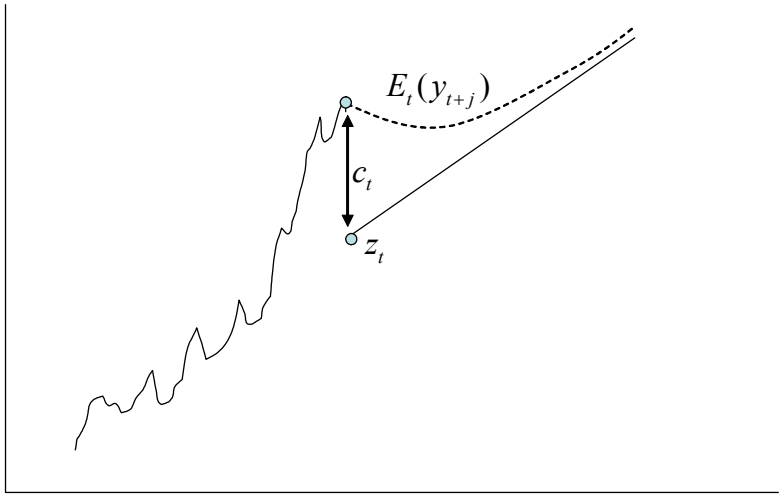

Figure 10.4: Beveridge-Nelson decomposition

Given this definition of z_t, we can show that it's the same z as in the B-N decomposition:

$$z_t - z_{t-1} = \lim_{k\to\infty}(E_t(y_{t+k}) - E_{t-1}(y_{t+k}) + \mu)$$

$E_t(y_{t+k}) - E_{t-1}(y_{t+k})$ is the response of y_{t+k} to the shock at t, $E_t(y_{t+k}) - E_{t-1}(y_{t+k}) = \sum_{j=1}^{k} a_j\epsilon_t$. Thus

$$\lim_{k\to\infty}(E_t(y_{t+k}) - E_{t-1}(y_{t+k})) = a(1)\epsilon_t$$

and

$$z_t = \mu + z_{t-1} + a(1)\epsilon_t.$$

The construction of the B-N trend shows that *the innovation variance of the random walk component is* $a(1)^2\sigma_\epsilon^2$. Thus, if the series is already TS, the innovation variance of the random walk component is *zero.* If the series has a small a(1), and thus is "mean-reverting", it will have a small random walk component. If the series already is a random walk, then $a_0 = 1, a_j = 0$, so $y_t = z_t$.

Beveridge and Nelson defined the trend as above, and then derived the process for c_t as well as z_t. I went about it backward with the advantage of hindsight.

In the Beveridge-Nelson decomposition the innovations to the stationary and random walk components are perfectly correlated. Consider instead an *arbitrary* combination of stationary and random walk components, in which the innovations have some arbitrary correlation.

$$y_t = z_t + c_t$$

$$z_t = \mu + z_{t-1} + \nu_t$$

$$c_t = b(L)\delta_t$$

Fact: *In Every decomposition of* y_t *into stationary and random walk components, the variance of changes to the random walk component is the same,* $a(1)^2\ \sigma_\epsilon^2$.

Proof: Take differences of y_t,

$$(1-L)y_t = (1-L)z_t + (1-L)c_t = \mu + \nu_t + (1-L)b(L)\delta_t$$

$(1-L)y_t$ is stationary, so must have a Wold representation

$$(1-L)y_t = \mu + \nu_t + (1-L)b(L)\delta_t = \mu + a(L)\epsilon_t,$$

Its spectral density at frequency zero is

$$S_{\Delta y}(0) = a(1)^2\sigma_\epsilon^2 = \sigma_\nu^2 help$$

The $(1-L)$ means that the δ term does not affect the spectral density at zero. Thus, *the spectral density at zero of* $(1-L)y_t$ *is the innovation variance of* ANY *random walk component.*

□

The "long-horizon mean reverting" case is one of a small a(1). Thus this case corresponds to a small random walk component and a large and interesting stationary component.

10.3.5 Forecast error variances

Since the unit root process is composed of a stationary plus random walk component, you should not be surprised if the unit root process has the same variance of forecasts behavior as the random walk, once the horizon is long enough that the stationary part has all died down.

To see this, use the Beveridge Nelson decomposition.

$$y_{t+k} = z_{t+k} + c_{t+k}$$

$$= z_t + k\mu + a(1)(\epsilon_{t+1} + \epsilon_{t+2} + \ldots + \epsilon_{t+k}) + a^*(L)\epsilon_{t+k}$$

The variance of the first term is

$$ka(1)^2\sigma_\epsilon^2$$

for large k, the variance of the second term approaches its unconditional variance $\text{var}(c_t) = \text{var}(a^*(L)\epsilon_t)$. Since the $a^*(L)$ die out, the covariance term is also dominated by the first term. r (If $a^*(L)$ only had finitely many terms this would be even more obvious.) Thus, at large k,

$$var_t(y_{t+k}) \to ka(1)^2\sigma_\epsilon^2 + \text{(terms that grow slower than k)}.$$

The basic idea is that the random walk component will eventually dominate, since its variance goes to infinity and the variance of anything stationary is eventually limited.

10.3.6 Summary

In summary, the quantity a(1) controls the extent to which a process looks like a random walk. If

$$(1-L)y_t = \mu + a(L)\epsilon_t$$

then

$$\begin{aligned}
a(1) &= \text{limit of } y_t \text{ impulse-response} \\
a(1)^2\sigma_\epsilon^2 &= S_{(1-L)y_t}(0) \\
a(1)^2\sigma_\epsilon^2 &= \text{var } (1-L) \text{ random walk component} \\
a(1)^2\sigma_\epsilon^2 &= \left(1 + 2\sum_{j=1}^{\infty}\rho_j\right) var(\Delta y_t) \\
a(1)^2\sigma_\epsilon^2 &= 1 + 2\sum_{j=1}^{\infty}\gamma_j \\
a(1)^2\sigma_\epsilon^2 &= \lim_{k\to\infty} var_t(y_{t+k})/k \\
a(1)^2\sigma_\epsilon^2 &=
\end{aligned}$$

In these many ways a(1) quantifies the extent to which a series "looks like" a random walk.

10.4 Summary of $a(1)$ estimates and tests.

Obviously, estimating and testing $a(1)$ is going to be an important task. Before reviewing some approaches, there is a general point that must be made.

10.4.1 Near- observational equivalence of unit roots and stationary processes in finite samples

So far, I've shown that a(1) distinguishes unit root from stationary processes. Furthermore *a(1) is the* only *thing that distinguishes unit roots from stationary processes.* There are several ways to see this point.

1) Given a spectral density, we could change the value at zero to some other value leaving the rest alone. If the spectral density at zero starts positive, we could set it to zero, making a stationary process out of a unit root process and vice versa.

2) Given a Beveridge-Nelson decomposition, we can construct a new series with a different random walk component. Zap out the random walk component, and we've created a new stationary series; add a random walk component to a stationary series and you've created a unit root. The variance of the random walk component was determined only by a(1).

3) And so forth (See Cochrane (1991) for a longer list).

(Actually the statement is only true in a finite sample. In population, we also know that the slope of the spectral density is zero at frequency zero:

$$S(\omega) = \gamma_0 + 2\sum_{j=1}^{\infty} \cos(j\omega)\gamma_j$$

$$\left.\frac{dS(\omega)}{d\omega}\right|_{\omega=0} = \left. -2\sum_{j=1}^{\infty} j\sin(j\omega)\gamma_j \right|_{\omega=0} = 0.$$

Thus, you can't just change the point at zero. Changing the spectral density at a point also doesn't make any difference, since it leaves all integrals unchanged. However, in a finite sample, you can change the *periodogram* ordinate at zero leaving all the others alone.)

Another way of stating the point is that (in a finite sample) *there are unit root processes arbitrarily "close" to any stationary process, and there are stationary processes arbitrarily "close" to any unit root process.* To see the first point, take a stationary process and add an arbitrarily small random walk component. To see the second, take a unit root process and change the unit root to .9999999. "Close" here can mean any statistical measure of distance, such as autocovariance functions, likelihood functions, etc.

Given these points, you can see that testing for a unit root vs. stationary process is hopeless in a finite sample. We could always add a *tiny* random walk component to a stationary process and make it a unit root process; yet in a finite sample we could never tell the two processes apart.

What "unit root tests" do is to restrict the null: they test for a unit root *plus* restrictions on the a(L) polynomial, such as a finite order AR, versus trend stationary *plus* restrictions on the a(L) polynomial. Then, they promise to slowly remove the restrictions as sample size increases.

10.4.2 Empirical work on unit roots/persistence

Empirical work generally falls into three categories:

1) Tests for unit roots (Nelson and Plosser (1982)) The first kind of tests were tests whether series such as GNP contained unit roots. As we have seen, the problem is that such tests must be accompanied by restrictions on a(L) or they have no content. Furthermore, it's not clear that we're that interested in the results. If a series has a unit root but tiny a(1) it behaves almost exactly like a stationary series. For both reasons, unit root tests are in practice just tests for the size of a(1).

Nonetheless, there is an *enormous* literature on testing for unit roots. Most of this literature centers on the asymptotic distribution of various test procedures under a variety of null hypotheses. The problem is econometrically interesting because the asymptotic distribution (though *not* the finite sample distribution) is usually discontinuous as the root goes to 1. If there is even a *tiny* random walk component, it will eventually swamp the rest of the series as the sample grows

2 Parametric Measures of a(1) (Campbell and Mankiw (1988)) In this kind of test, you fit a parametric (ARMA) model for GNP and then find the implied a(1) of this parametric model. This procedure has all the advantages and disadvantages of any spectral density estimate by parametric model. If the parametric model is correct, you gain power by using information at all frequencies to fit it. If it is incorrect, it will happily forsake accuracy in the region you care about (near frequency zero) to gain more accuracy in regions you don't care about. (See the Sims approximation formula above.)

3. "Nonparametric" estimates of $a(1)$. (Cochrane (1988), Lo and MacKinlay (1988), Poterba and Summers (1988)) Last, one can use spectral density estimates or their equivalent weighted covariance estimates to directly estimate the spectral density at zero and thus a(1), ignoring the rest of the process. This is the idea behind "variance ratio" estimates. These estimates have much greater standard errors than parametric estimates, but less bias if the parametric model is in fact incorrect.

Chapter 11

Cointegration

Cointegration is generalization of unit roots to vector systems. As usual, in vector systems there are a few subtleties, but all the formulas look just like obvious generalizations of the scalar formulas.

11.1 Definition

Suppose that two time series are each *integrated,* i.e. have unit roots, and hence moving average representations

$$(1 - L)y_t = a(L)\delta_t$$

$$(1 - L)w_t = b(L)\nu_t$$

In general, linear combinations of y and w also have unit roots. However, if there is some linear combination, say $y_t - \alpha w_t$, that is stationary, y_t and w_t are said to be *cointegrated*, and $[1 - \alpha]$ is their *cointegrating vector.*

Here are some plausible examples. Log GNP and log consumption each probably contain a unit root. However, the consumption/GNP *ratio* is stable over long periods, thus log consumption − log GNP is stationary, and log GNP and consumption are *cointegrated.* The same holds for any two components of GNP (investment, etc). Also, log stock prices certainly contain a unit root; log dividends probably do too; but the dividend/price *ratio* is stationary. Money and prices are another example.

11.2 Cointegrating regressions

Like unit roots, cointegration attracts much attention for statistical reasons as well as for the economically interesting time-series behavior that it represents. An example of the statistical fun is that estimates of cointegrating vectors are "superconsistent"—you can estimate them by OLS even when the right hand variables are correlated with the error terms, and the estimates converge at a faster rate than usual.

Suppose y_t and w_t are cointegrated, so that $y_t - \alpha w_t$ is stationary. Now, consider running

$$y_t = \beta w_t + u_t$$

by OLS. OLS estimates of β converge to α, even if the errors u_t are correlated with the right hand variables w_t!

As a simple way to see this point, recall that OLS tries to minimize the variance of the residual. If $y_t - w_t\beta$ is stationary, then for any $\alpha \neq \beta$, $y_t - w_t\alpha$ has a unit root and hence contains a random walk (recall the B-N decomposition above). Thus, the variance of $(y_t - w_t\beta),\ \beta \neq \alpha$ increases to ∞ as the sample size increases; while the variance of $(y_t - w_t\alpha)$ approaches some finite number. Thus, OLS will pick $\hat{\beta} = \alpha$ in large samples.

Here's another way to see the same point: The OLS estimate is

$$\hat{\beta} = (W'W)^{-1}(W'Y) = (\sum_t w_t^2)^{-1}(\sum_t w_t(\alpha w_t + u_t)) =$$

$$\alpha + \sum_t w_t u_t / \sum_t w_t^2 = \alpha + \frac{\frac{1}{T}\sum_t w_t u_t}{\frac{1}{T}\sum_t w_t^2}$$

Normally, the plim of the last term is not zero, so $\hat{\beta}$ is an inconsistent estimate of α. We assume that the denominator converges to a nonzero constant, as does the numerator, since we have not assumed that $E(w_t u_t) = 0$. But, when w_t has a unit root, the *denominator* of the last term goes to ∞, so OLS is consistent, even if the *numerator* does not converge to zero! Furthermore, the denominator goes to ∞ very fast, so $\hat{\beta}$ converges to α at rate T rather than the usual rate $T^{1/2}$.

As an example, consider the textbook simultaneous equations problem:

$$y_t = c_t + a_t$$

$$c_t = \alpha y_t + \epsilon_t$$

α_t is a shock ($a_t = i_t + g_t$); a_t and ϵ_t are iid and independent of each other. If you estimate the c_t equation by OLS you get biased and inconsistent estimates of α. To see this, you first solve the system for its reduced form,

$$y_t = \alpha y_t + \epsilon_t + a_t = \frac{1}{1-\alpha}(\epsilon_t + a_t)$$

$$c_t = \frac{\alpha}{1-\alpha}(\epsilon_t + a_t) + \epsilon_t = \frac{1}{1-\alpha}\epsilon_t + \frac{\alpha}{1-\alpha}a_t$$

Then,

$$\hat{\alpha} \to \frac{plim(\frac{1}{T}\sum c_t y_t)}{plim(\frac{1}{T}\sum y_t^2)} = \frac{plim(\frac{1}{T}\sum(\alpha y_t + \epsilon_t)y_t)}{plim(\frac{1}{T}\sum y_t^2)} = \alpha + \frac{plim(\frac{1}{T}\sum \epsilon_t y_t)}{plim(\frac{1}{T}\sum y_t^2)}$$

$$= \alpha + \frac{\frac{\sigma_\epsilon^2}{1-\alpha}}{\frac{\sigma_\epsilon^2+\sigma_a^2}{(1-\alpha)^2}} = \alpha + (1-\alpha)\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_a^2}$$

As a result of this bias and inconsistency a lot of effort was put into estimating "consumption functions" consistently, i.e. by 2SLS or other techniques.

Now, suppose instead that $a_t = a_{t-1} + \delta_t$. This induces a unit root in y and hence in c as well. But since ϵ is still stationary, $c - \alpha y$ is stationary, so y and c are cointegrated. Now $\sigma_a^2 \to \infty$, so plim$(\frac{1}{T}\sum y_t^2) = \infty$ and $\hat{\alpha} \to \alpha$! Thus, none of the 2SLS etc. corrections are needed!

More generally, estimates of cointegrating relations are robust to many of the errors correlated with right hand variables problems with conventional OLS estimators, including errors in variables as well as simultaneous equations and other problems.

11.3 Representation of cointegrated system.

11.3.1 Definition of cointegration

Consider a first difference stationary vector time series x_t. The elements of x_t are *cointegrated* if there is *at least one* vector α such that $\alpha' x_t$ is stationary in levels. α is known as the *cointegrating vector.*

Since differences of x_t are stationary, x_t has a moving average representation

$$(1-L)x_t = A(L)\epsilon_t.$$

Since $\alpha' x_t$ stationary is an extra restriction, it must imply a restriction on A(L). It shouldn't be too hard to guess that that restriction must involve A(1)!

11.3.2 Multivariate Beveridge-Nelson decomposition

It will be useful to exploit the multivariate Beveridge-Nelson decomposition,

$$x_t = z_t + c_t,$$

$$(1-L)z_t = A(1)\epsilon_t$$

$$c_t = A^*(L)\epsilon_t \; ; A_j^* = -\sum_{k=j+1}^{\infty} A_k$$

This looks just like and is derived exactly as the univariate BN decomposition, except the letters stand for vectors and matrices.

11.3.3 Rank condition on A(1)

Here's the restriction on A(1) implied by cointegration: *The elements of x_t are cointegrated with cointegrating vectors α_i iff $\alpha_i' A(1) = 0$.* This implies that *the rank of A(1) is (number of elements of x_t - number of cointegrating vectors α_i)*

Proof: Using the multivariate Beveridge Nelson decomposition,

$$\alpha' x_t = \alpha' z_t + \alpha' c_t.$$

$\alpha' c_t$ is a linear combination of stationary random variables and is hence stationary. $\alpha' z_t$ is a linear combination of random walks. This is either constant or nonstationary. Thus, it must be constant, i.e. its variance must be zero and, since

$$(1-L)\alpha' z_t = \alpha' A(1)\epsilon_t,$$

we must have

$$\alpha' A(1) = 0$$

□

In analogy to a(1) = 0 or | $a(1)$ |> 0 in univariate time series, we now have three cases:

Case 1 : $A(1) = 0 \Leftrightarrow x_t$ stationary in levels; all linear combinations of x_t stationary in levels.

Case 2 : $A(1)$ less than full rank $\Leftrightarrow (1 - L)x_t$ stationary, some linear combinations $\alpha' x_t$ stationary.

Case 3 : $A(1)$ full rank $\Leftrightarrow (1 - L)x_t$ stationary, no linear combinations of x_t stationary.

For unit roots, we found that whether a(1) was zero or not controlled the spectral density at zero, the long-run impulse-response function and the innovation variance of random walk components. The same is true here for A(1), as we see next.

11.3.4 Spectral density at zero

The spectral density matrix of $(1 - L)x_t$ at frequency zero is $S_{(1-L)x_t}(0) = \Psi = A(1)\Sigma A(1)'$. Thus, $\alpha' A(1) = 0$ implies $\alpha' A(1)\Sigma A(1)' = 0$, so *the spectral density matrix of* $(1 - L)x_t$ *is also less than full rank, and* $\alpha'\Psi = 0$ *for any cointegrating vector* α.

The fact that the spectral density matrix at zero is less than full rank gives a nice interpretation to cointegration. In the $2x2$ case, the spectral density matrix at zero is less than full rank if its determinant is zero, i.e. if

$$S_{\Delta y}(0)S_{\Delta w}(0) =\mid S_{\Delta y \Delta w}(0) \mid^2$$

This means that *the components at frequency zero are perfectly correlated.*

11.3.5 Common trends representation

Since the zero frequency components are perfectly correlated, there is in a sense only *one* common zero frequency component in a 2-variable cointe-

grated system. The common trends representation formalizes this idea.

$\Psi = A(1)\Sigma A(1)'$ is also the innovation variance-covariance matrix of the Beveridge-Nelson random walk components. When the rank of this matrix is deficient, we obviously need *fewer* than N random walk components to describe N series. This means that there are *common* random walk components. In the $N = 2$ case, in particular, two cointegrated series are described by stationary components around a *single* random walk.

Precisely, we're looking for a representation

$$\begin{bmatrix} y_t \\ w_t \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} z_t + \text{stationary.}$$

Since the spectral density at zero (like any other covariance matrix) is symmetric, it can be decomposed as

$$\Psi = A(1)\Sigma A(1)' = Q\Lambda Q'$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

and $QQ\prime = Q\prime Q = I$. Then λ_i are the eigenvalues and Q is an orthogonal matrix of corresponding eigenvectors. If the system has N series and K cointegrating vectors, the rank of Ψ is $N - K$, so K of the eigenvalues are zero.

Let ν_t define a new error sequence,

$$\nu_t = Q' A(1)\epsilon_t$$

Then $E(\nu_t \nu_t') = Q' A(1)\Sigma A(1)' Q = Q' Q \Lambda Q' Q = \Lambda$ So the variance-covariance matrix of the ν_t shocks is diagonal, and has the eigenvalues of the Ψ matrix on the diagonal.

In terms of the new shocks, the Beveridge-Nelson trend is

$$z_t = z_{t-1} + A(1)\epsilon_t = z_{t-1} + Q\nu_t$$

But components of ν with variance zero might as well not be there. For example, in the $2x2$ case, we might as well write

$$\begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} = \begin{bmatrix} z_{1t-1} \\ z_{2t-1} \end{bmatrix} + Q \begin{bmatrix} \nu_{1t} \\ \nu_{2t} \end{bmatrix}; \quad E(\nu_t v_t') = \begin{bmatrix} \lambda_1 & 0 \\ 0 & 0 \end{bmatrix}$$

as

$$\begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} = \begin{bmatrix} z_{1t-1} \\ z_{2t-1} \end{bmatrix} + \begin{bmatrix} Q_{11} \\ Q_{21} \end{bmatrix} v_{1t}; \;\; E(v_{1t}^2) = \lambda_1.$$

Finally, since $z_1$and z_2 are perfectly correlated, we can write the system with only one random walk as

$$\begin{bmatrix} y_t \\ w_t \end{bmatrix} = \begin{bmatrix} Q_{11} \\ Q_{21} \end{bmatrix} z_t + A^*(L)\epsilon_t.$$

$$(1-L)z_t = \nu_{1t} = [1\ 0]Q'A(1)\epsilon_t$$

This is the *common trends* representation. y_t and w_t share a single *common trend*, or common random walk component z_t.

We might as well order the eigenvalues λ_i from largest to smallest, with the zeros on the bottom right. In this way, we rank the common trends by their importance.

With univariate time series, we thought of a continuous scale of processes between stationary and a pure random walk, by starting with a stationary series and adding random walk components of increasing variance, indexed by the increasing size of $a(1)^2\sigma_\epsilon^2$. Now, start with a stationary series x_t, i.e. let all the eigenvalues be zero. We can then add random walk components one by one, and slowly increase their variance. The analogue to "near-stationary" series we discussed before are "nearly cointegrated" series, with a *very small* eigenvalue or innovation variance of the common trend.

11.3.6 Impulse-response function.

$A(1)$ is the limiting impulse-response of the levels of the *vector* x_t. For example, $A(1)_{yw}$ is the long-run response of y to a unit w shock. To see how cointegration affects $A(1)$, consider a simple (and very common) case, $\alpha = [1\ -1]'$. The reduced rank of A(1) means

$$\alpha' A(1) = 0$$

$$[1\ -1] \begin{bmatrix} A(1)_{yy} & A(1)_{yw} \\ A(1)_{wy} & A(1)_{ww} \end{bmatrix} = 0$$

hence

$$A(1)_{yy} = A(1)_{wy}$$
$$A(1)_{yw} = A(1)_{ww}$$

each variable's long-run response to a shock must be the same. The reason is intuitive: if y and w had different long-run responses to a shock, the difference $y - w$ would not be stationary. Another way to think of this is that the response of the difference $y - w$ must vanish, since the difference must be stationary. A similar relation holds for arbitrary α.

11.4 Useful representations for running cointegrated VAR's

The above are very useful for thinking about cointegration, but you have to run AR's rather than $MA(\infty)$s. Since the Wold $MA(\infty)$ isn't invertible when variables are cointegrated, we have to think a little more carefully about how to run VARs when variables are cointegrated.

11.4.1 Autoregressive Representations

Start with the autoregressive representation of the levels of x_t, $B(L)x_t = \epsilon_t$ or

$$x_t = -B_1 x_{t-1} - B_2 x_{t-2} + \ldots + \epsilon_t.$$

Applying the B-N decomposition $B(L) = B(1) + (1 - L)B^*(L)$ to the lag polynomial operating on $x_{t-1}(t-1$, not $t)$ on the right hand side, we obtain

$$x_t = (-(B_1+B_2+\ldots))x_{t-1}+(B_2+B_3+..)(x_{t-1}-x_{t-2})+(B_3+B_4+..)(x_{t-2}-x_{t-3})+\ldots+\epsilon_t$$

$$x_t = (-(B_1 + B_2 + \ldots))x_{t-1} + \sum_{j=1}^{\infty} B_j^* \Delta x_{t-j} + \epsilon_t$$

Hence, subtracting x_{t-1} from both sides,

$$\Delta x_t = -B(1)x_{t-1} + \sum_{j=1}^{\infty} B \Delta x_{t-j} + \epsilon_t.$$

As you might have suspected, the matrix $B(1)$ controls the cointegration properties. $\Delta x_t, \sum_{j=1}^{\infty} B\ \Delta x_{t-j}$, and ϵ_t are stationary, so $B(1)x_{t-1}$ had also better be stationary. There are three cases:

Case 1 : $B(1)$ is full rank, and any linear combination of x_{t-1} is stationary. x_{t-1} is stationary. In this case we run a normal VAR in levels.

Case 2 : $B(1)$ has rank between 0 and full rank. There are *some* linear combinations of x_t that are stationary, so x_t is cointegrated. As we will see, the VAR in levels is consistent but inefficient (if you know the cointegrating vector) and the VAR in differences is misspecified in this case.

Case 3 : $B(1)$ has rank zero, so no linear combination of x_{t-1} is stationary. Δx_t is stationary with no cointegration. In this case we run a normal VAR in first differences.

11.4.2 Error Correction representation

As a prelude to a discussion of what to do in the cointegrated case, it will be handy to look at the error correction representation.

If $B(1)$ has less than full rank, we can express it as

$$B(1) = \gamma\alpha';$$

If there are K cointegrating vectors, then the rank of $B(1)$ is K and γ and α each have K columns. Rewriting the system with γ and α, we have the *error correction representation*

$$\Delta x_t = -\gamma\alpha' x_{t-1} + \sum_{j=1}^{\infty} B_j^* \Delta x_{t-j} + \epsilon_t.$$

Note that since γ spreads K variables into N variables, $\alpha' x_{t-1}$ itself must be stationary so that $\gamma\alpha' x_{t-1}$ will be stationary. Thus, α must be the matrix of cointegrating vectors.

This is a very nice representation. If $\alpha' x_{t-1}$ is not 0 (its mean), $\gamma\alpha' x_{t-1}$ puts in motion increases or decreases in Δx_t to restore $\alpha' x_t$ towards its mean. In this sense "errors" – deviations from the cointegrating relation $\alpha' x_t = 0$ – set in motion changes in x_t that "correct" the errors.

11.4.3 Running VAR's

Cases 1 or 3 are easy: run a VAR in levels or first differences. The hard case is case 2, cointegration.

With cointegration, a pure VAR in differences,

$$\Delta y_t = a(L)\Delta y_{t-1} + b(L)\Delta w_{t-1} + \delta_t$$

$$\Delta w_t = c(L)\Delta y_{t-1} + d(L)\Delta w_{t-1} + \nu_t$$

is misspecified. Looking at the error-correction form, there is a missing regressor, $\alpha'[y_{t-1}w_{t-1}]'$. This is a real problem; often the lagged cointegrating vector is the most important variable in the regression. A pure VAR in levels, $y_t = a(L)y_{t-1} + b(L)w_{t-1} + \delta_t w_t = c(L)y_{t-1} + d(L)w_{t-1} + \nu_t$ looks a little unconventional, since both right and left hand variables are nonstationary. Nonetheless, the VAR is not misspecified, and the estimates are consistent, though the coefficients may have non-standard distributions. (Similarly, in the regression $x_t = \phi x_{t-1} + \epsilon_t$; when x_t was generated by a random walk $\hat{\phi} \to 1$, but with a strange distribution.) However, they are not efficient: If there is cointegration, it imposes restrictions on $B(1)$ that are not imposed in a pure VAR in levels.

One way to impose cointegration is to run an error-correction VAR,

$$\Delta y_t = \gamma_y(\alpha_y y_{t-1} + \alpha_w w_{t-1}) + a(L)\Delta y_{t-1} + b(L)\Delta w_{t-1} + \delta_t$$
$$\Delta w_t = \gamma_w(\alpha_y y_{t-1} + \alpha_w w_{t-1}) + c(L)\Delta y_{t-1} + d(L)\Delta w_{t-1} + \nu_t$$

This specification *imposes* that y and w are cointegrated with cointegrating vector α. This form is particularly useful if you know *ex-ante* that the variables are cointegrated, and you know the cointegrating vector, as with consumption and GNP. Otherwise, you have to pre-test for cointegration and estimate the cointegrating vector in a separate step. Advocates of just running it all in levels point to the obvious dangers of such a two step procedure.

A further difficulty with the error-correction form is that it doesn't fit neatly into standard VAR packages. Those packages are designed to regress N variables on their own lags, not on their own lags and a lagged difference of the levels. A way to use standard packages is to estimate instead the *companion form*, one difference and the cointegrating vector.

$$\Delta y_t = a(L)\Delta y_{t-1} + b(L)(\alpha_y y_{t-1} + \alpha_w w_{t-1}) + \delta_t$$

$$(\alpha_y y_t + \alpha_w w_t) = c(L)\Delta y_{t-1} + d(L)(\alpha_y y_{t-1} + \alpha_w w_{t-1}) + \nu_t$$

This is equivalent (except for lag length) and can be estimated with regular VAR packages. Again, it requires a priori knowledge of the cointegrating vector.

There is much controversy over which approach is best. My opinion is that when you really don't know whether there is cointegration or what the vector is, the AR in levels approach is probably better than the approach of a battery of tests for cointegration plus estimates of cointegrating relations followed by a companion or error correction VAR. However, much of the time you *know* the variables are cointegrated, and you *know* the cointegrating vector. Consumption and GNP are certainly cointegrated, and the cointegrating vector is [1 -1]. In these cases, the error-correction and companion form are probably better, since they impose this prior knowledge. (If you run the AR in levels, you will get close to, but not exactly, the pattern of cointegration you know is in the data.) The slight advantage of the error-correction form is that it treats both variables symmetrically. It is also equivalent to a companion form with a longer lag structure in the cointegrating relation. This may be important, as you expect the cointegrating relation to decay very slowly. Also, the error correction form will likely have less correlated errors, since there is a y on both left hand sides of the companion form. However, the companion form is easier to program.

Given any of the above estimates of the AR representation, it's easy to find the $MA(\infty)$ representation of first differences by simply simulating the impulse-response function.

11.5 An Example

Consider a first order VAR

$$y_t = a y_{t-1} + b w_{t-1} + \delta_t$$

$$w_t = c y_{t-1} + d w_{t-1} + \nu_t$$

or

$$\left(I - \begin{bmatrix} a & b \\ c & d \end{bmatrix} L \right) x_t = \epsilon_t; \quad B(1) = \begin{bmatrix} 1-a & -b \\ -c & 1-d \end{bmatrix}$$

An example of a singular $B(1)$ is $b = 1 - a, c = 1 - d$,

$$B(1) = \begin{bmatrix} b & -b \\ -c & c \end{bmatrix}.$$

The original VAR in levels with these restrictions is

$$\begin{aligned} y_t &= (1-b)y_{t-1} + bw_{t-1} + \delta_t \\ w_t &= cy_{t-1} + (1-c)w_{t-1} + \nu_t \end{aligned}$$

or

$$\left(I - \begin{bmatrix} 1-b & b \\ c & 1-c \end{bmatrix} L\right) x_t = \epsilon_t;$$

The error-correction representation is

$$\Delta x_t = -\gamma\alpha' x_{t-1} + \sum_{j=1}^{\infty} B_j^* \Delta x_{t-j} + \epsilon_t.$$

$$B(1) = \begin{bmatrix} b \\ -c \end{bmatrix}[1 \ -1] \ \Rightarrow \ \gamma = \begin{bmatrix} b \\ -c \end{bmatrix}, \ \alpha = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Since B is only first order, B_1^* and higher $= 0$, so the error-correction representation is

$$\begin{aligned} \Delta y_t &= -b(y_{t-1} - w_{t-1}) + \delta_t \\ \Delta w_t &= c(y_{t-1} - w_{t-1}) + \nu_t \end{aligned}$$

As you can see, if $y_{t-1} - w_{t-1} > 0$, this lowers the growth rate of y and raises that of w, restoring the cointegrating relation.

We could also subtract the two equations,

$$\Delta(y_t - w_t) = -(b+c)(y_{t-1} - w_{t-1}) + (\delta_t - \nu_t)$$

$$y_t - w_t = (1 - (b+c))(y_{t-1} - w_{t-1}) + (\delta_t - \nu_t)$$

so $y_t - w_t$ follows an $AR(1)$.(You can show that $2 > (b+c) > 0$, so it's a stationary $AR(1)$.) Paired with either the Δy or Δw equation given above, this would form the companion form.

In this system it's also fairly easy to get to the $MA(\infty)$ representation for differences. From the last equation,

$$y_t - w_t = (1 - (1 - (b + c))L)^{-1}(\delta_t - \nu_t) \equiv (1 - \theta L)^{-1}(\delta_t - \nu_t)$$

so

$$\Delta y_t = -b(y_{t-1} - w_{t-1}) + \delta_t$$
$$\Delta w_t = c(y_{t-1} - w_{t-1}) + v_t$$

becomes

$$\Delta y_t = -b(1 - \theta L)^{-1}(\delta_t - \nu_t) + \delta_t = (1 - b(1 - \theta L)^{-1})\delta_t + b(1 - \theta L)^{-1}v_t$$
$$\Delta w_t = c(1 - \theta L)^{-1}(\delta_t - \nu_t) + v_t = c(1 - \theta L)^{-1}\delta_t + (1 - c(1 - \theta L)^{-1})v_t$$

$$\begin{bmatrix} \Delta y_t \\ \Delta w_t \end{bmatrix} = (1 - \theta L)^{-1} \begin{bmatrix} (1 - \theta L) - b & b \\ c & (1 - \theta L) - c \end{bmatrix} \begin{bmatrix} \delta_t \\ v_t \end{bmatrix}$$

$$\begin{bmatrix} \Delta y_t \\ \Delta w_t \end{bmatrix} = (1-(1-b-c)L)^{-1} \begin{bmatrix} 1 - b - (1 - b - c)L & b \\ c & 1 - c - (1 - b - c)L \end{bmatrix} \begin{bmatrix} \delta_t \\ v_t \end{bmatrix}$$

Evaluating the right hand matrix at $L = 1$,

$$(b + c)^{-1} \begin{bmatrix} c & b \\ c & b \end{bmatrix}$$

Denoting the last matrix $A(L)$, note $\alpha' A(1) = 0$. You can also note $A(1)\gamma = 0$, another property we did not discuss.

11.6 Cointegration with drifts and trends

So far I deliberately left out a drift term or linear trends. Suppose we put them back in, i.e. suppose

$$(1 - L)x_t = \mu + A(L)\epsilon_t.$$

The $B - N$ decomposition was

$$z_t = \mu + z_{t-1} + A(1)\epsilon_t.$$

Now we have two choices. If we stick to the original definition, that $\alpha' x_t$ must be stationary, it must also be the case that $\alpha'\mu = 0$. This is a *separate* restriction. If $\alpha' A(1) = 0$, but $\alpha'\mu \neq 0$, then

$$\alpha' z_t = \alpha'\mu + \alpha' z_{t-1} \Rightarrow \alpha' z_t = \alpha' z_0 + (\alpha'\mu)t.$$

Thus, $\alpha' x_t$ will contain a *time trend* plus a stationary component. Alternatively, we could define cointegration to be $\alpha' x_t$ contains a *time* trend, but no *stochastic* (random walk) trends. Both approaches exist in the literature, as well as generalizations to higher order polynomials. See Campbell and Perron (1991) for details.