Macroeconometrics

Winter 2014, Y1 (Masters students)

Antti Ripatti[1]

version: February 24, 2014

Outline

Contents

1 Introduction **2**

2 Difference equations and lag operators **3**
2.1 Lag operators . . . 3
2.2 Difference equations . . . 5

3 Approximating and Solving DSGE Models **7**
3.1 Examples of dynamic macroeconomic models . . . 7
3.1.1 Real Business Cycle Model . . . 7
3.1.2 The Basic New-Keynesian Model . . . 8
3.2 Approximating . . . 9
3.3 Solving . . . 11
3.3.1 Blanchard and Kahn method . . . 11
3.3.2 Klein method . . . 14
3.3.3 Method of undetermined coefficients . . . 15

4 Moments of Model and Data **16**
4.1 Introduction . . . 16
4.2 Autocovariance functions and alike . . . 18
4.3 Spectral analysis . . . 19
4.3.1 Univariate Frequency-Domain Methods . . . 19
4.3.2 Estimation . . . 23
4.3.3 Linear Filters . . . 24
4.4 Measuring Business Cycles . . . 28
4.5 Multivariate time series models . . . 31
4.6 Filtering the data and model . . . 37

5 Matching moments **38**
5.1 Calibrating steady-state . . . 38
5.2 Comparing model and data moments . . . 39
5.3 GMM . . . 39

6 Filtering and Likelihood approach **41**
6.1 Kalman filter and the maximum likelihood . . . 41
6.2 Filtering and decompositions . . . 47
6.3 Bayesian methods . . . 47
6.4 Simulating posterior . . . 50
6.5 Model comparison . . . 55

1 Introduction

The Goal

The goal of the course is give an overview of the econometric methods to study how to validate macroeconomic models.

If time allows, we study the use dynamic macroeconomics models to address empirical issues like

- Model Estimation
- Model Comparison
- Forecasting
- Measurement
- Shock Identification
- Policy Analysis

Evolution of empirical macro – Part 1: Birth of large models

The representation of theoretical models as complete probability models dates at least to Haavelmo (1944, Econometrica)

Models were loosely related to economic theory:

- Production functions
- National account identities
- Old-Keynesian consumption function

And estimated using OLS or instrumental variables methods (2SLS).
Phillips curve was upward-sloping even in the long-run! Policy tool.

Evolution of empirical macro – Part 2b: Scent of Neo-Classical synthesis

- Developments in time series econometrics: cointegration and error correction models
- Separation of short-run (dynamics) and long-run (cointegration)
- Long-run (cointegration) builds on economic theory
- Not yet general equilibrium
- Ad-hoc dynamics
- No sensible steady-state
- Vulnerable to Lucas critique

AWM, BOF3, BOF4

Evolution of empirical macro – Part 3a: Kydland–Prescott

Methodological revolution
Intertemporally optimizing agents. Budget and technology constraints.

Conceptual revolution

- In a frictionless markets under perfect competition business cycles are efficient: no need for stabilization; stabilization may be counter-productive.
- Economic fluctuations are caused by technology shocks: they are the main source of fluctuation.
- Monetary factors (price level) has a limited (or no) role
- Calibration

Evolution of empirical macro – Part 3a: Kydland–Prescott. . .
Implementation

1. Pose a question
2. Use theory to address the question
3. Construct the model economu
4. Calibrate
5. Run the experiment within the model

Lack of statistical formality in calibration.

Evolution of empirical macro – Part 3b: Bayesian estimation

- Sargent (1989 JPE) presents the mapping of linearized DSGE models into state-space representation
- DeJong, Ingram and Whiteman (2000, JoE, 2000, JAE) developed methods for mapping priors over structural parameters into priors over corresponding parameters of the state-space representation enabling Bayesian inference.
- Smets and Wouters (2003) made a real working application.

This course
This course builds on Part 3 (Kydland and Prescott (1982)) and Sims (2011). The aim is to learn how the model meets data.

We

- stationarize the model: restrictions implied by the balanced growth path
- approximate it with the first order Taylor approximation
- review (some of) the solution methods
- refresh the basic statistical tools to compute moments
- filter the model and the data
- introduce the Bayesian methods to study model parameters

2 Difference equations and lag operators

This section builds mostly on Sargent (1987).

2.1 Lag operators

Lag operators
The *backward shift* or *lag operator* is defined as

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

$$L^n x_t = x_{t-n}, \quad \text{for } n = \ldots, -2, -1, 0, 1, 2, \ldots$$

Note that if $n < 0$ the operator shifts x_t *forward*.
This language is loose, since we start with the sequence

$$\{x_t\}_{t=-\infty}^{\infty},$$

where x_t is a real number. We *operate on* $\{x_t\}_{t=-\infty}^{\infty}$ *by* L.
This section builds mostly on Sargent (1987).

Operations with lag operator

Lag operator and multiplication operator are *commutative*

$$L(\beta x_t) = \beta L x_t.$$

It is *distributive* over the addition operator

$$L(x_t + w_t) = Lx_t + Lw_t.$$

Hence, we are free to use the standard commutative, associative, and distributive algebraic laws for multiplication and addition to express the compound operator in an alternative form.

Examples

$$y_t = (a + bL)Lx_t = (aL + bL^2)x_t = ax_{t-1} + bx_{t-2}$$

or

$$\begin{aligned}(1 - \lambda_1 L)(1 - \lambda_2 L)x_t &= (1 - \lambda_1 L - \lambda_2 L + \lambda_1\lambda_2 L^2)x_t \\ &= [1 - (\lambda_1 + \lambda_2)L + \lambda_1\lambda_2 L^2]x_t \\ &= x_t - (\lambda_1 + \lambda_2)x_{t-1} + \lambda_1\lambda_2 x_{t-2}.\end{aligned}$$

and

$$Lc = c$$

and

$$L^0 = 1$$

Polynomials in the lag operator

Polynomial

$$A(L) = a_0 + a_1 L + a_2 L^2 + \cdots = \sum_{j=0}^{\infty} a_j L^j,$$

where a_j's are constants.

$$\begin{aligned}A(L)x_t &= (a_0 + a_1 L + a_2 L^2 + \cdots)x_t \\ &= a_0 x_t + a_1 x_{t-1} + a_2 x_{t-2} + \cdots = \sum_{j=0}^{\infty} a_j x_{t-j}.\end{aligned}$$

Rational $A(L)$

$$(L) = \frac{B(L)}{C(L)},$$

where

$$B(L) = \sum_{j=0}^{m} b_j L^j, \qquad C(L) = \sum_{j=0}^{n} c_j L^j,$$

where b_j, c_j, m and n are constants.

Simple example

$$A(L) = \frac{1}{1 - \lambda L} = 1 + \lambda L + \lambda^2 L^2 + \cdots,$$

which is "useful" only if $|\lambda| < 1$. *Why?*

Then

$$\frac{1}{1 - \lambda L} x_t = (1 + \lambda L + \lambda^2 L^2 + \cdots)x_t = \sum_{i=0}^{\infty} \lambda^i x_{t-i}.$$

Consider the case $|\lambda| > 1$, the following "trick" is useful:

$$\begin{aligned}\frac{1}{1-\lambda L} &= \frac{-(\lambda L)^{-1}}{1-(\lambda L)^{-1}} \\ &= -\frac{1}{\lambda L}\left[1 + \frac{1}{\lambda}L^{-1} + \left(\frac{1}{\lambda}\right)^2 L^{-2} + \cdots\right] \\ &= -\frac{1}{\lambda}L^{-1} - \left(\frac{1}{\lambda}\right)^2 L^{-2} - \left(\frac{1}{\lambda}\right)^3 L^{-3} + \cdots.\end{aligned}$$

Then

$$\frac{1}{1-\lambda L}x_t = -\frac{1}{\lambda}x_{t+1} - \left(\frac{1}{\lambda}\right)^2 x_{t+2} + \cdots = -\sum_{i=1}^{\infty}\lambda^{-i}x_{t+i},$$

which is geometrically declining weighted sum of *future* values of x_t.

2.2 Difference equations

Difference equation

Consider the following difference equation

$$y_t = \lambda y_{t-1} + bx_t + a, \quad t = \ldots, -1, 0, 1, \ldots, \tag{1}$$

where y_t is an *endogenous* variable and x_t is an *exogenous* variable and sequences of real number and $\lambda \neq 1$. It can be written as

$$(1-\lambda L)y_t = a + bx_t.$$

Multiply both sides by $(1-\lambda L)^{-1}$ to obtain

$$\begin{aligned}y_t &= \frac{a}{1-\lambda L} + \frac{b}{1-\lambda L}x_t + c\lambda^t \\ &= \frac{a}{1-\lambda} + b\sum_{i=0}^{\infty}\lambda^i x_{t-i} + c\lambda^t.\end{aligned} \tag{2}$$

The reason why we have the last term is that for any constant c

$$(1-\lambda L)c\lambda^t = c\lambda^t - c\lambda\lambda^{t-1} = 0.$$

By multiplying both sides by $1-\lambda L$, gives the original difference equation! *This the "general solution"!*

To get *the "particular solution"* we must be able to tie down the constant *c with an additional bit of information*.

An example of an additional bit of information is

$$\lim_{n\to\infty}\sum_{i=n}^{\infty}\lambda^i x_{t-i} = 0 \quad \text{for all } t.$$

It simply says that $\lambda^i x_{t-i}$ must be "small" for large i. If x_t were constant, say $\bar{x}$, all time, this requires $|\lambda| < 1$. This also results bounded constant term in the solution.

Assume $t > 0$ to see the impact of an arbitrary initial condition. Solution may be written as

$$\begin{aligned}y_t &= a\sum_{i=0}^{t-1}\lambda^i + a\sum_{i=t}^{\infty}\lambda^i + b\sum_{i=0}^{t-1}\lambda^i x_{t-i} + b\sum_{i=t}^{\infty}\lambda^i x_{t-i} + c\lambda^t \\ &= a\frac{1-\lambda^t}{1-\lambda} + a\frac{\lambda^t}{1-\lambda} + b\sum_{i=0}^{t-1}\lambda^i x_{t-i} + b\lambda^t\sum_{i=0}^{\infty}\lambda^i x_{0-i} + c\lambda^t,\end{aligned}$$

$$y_t = a\frac{1-\lambda^t}{1-\lambda} + b\sum_{i=0}^{t-1}\lambda^i x_{t-i} + \lambda^t\left(\frac{a}{1-\lambda} + b\sum_{i=t}^{\infty}\lambda^i x_{0-i} + c\right), \quad t \geq 1.$$

The term in braces is equals y_0! Hence

$$\begin{aligned} y_t &= a\frac{1-\lambda^t}{1-\lambda} + b\sum_{i=0}^{t-1}\lambda^i x_{t-i} + \lambda^t y_0 \\ &= \frac{a}{1-\lambda} + \lambda^t\left(y_0 - \frac{a}{1-\lambda}\right) + b\sum_{i=0}^{t-1}\lambda^i x_{t-i}, \quad t \geq 1. \end{aligned}$$

The original difference equation (1) may be solved *"forward"* by applying the "forward inverse".

$$\begin{aligned} y_t &= \frac{-(\lambda L)^{-1}}{1-(\lambda L)^{-1}}a + b\frac{-(\lambda L)^{-1}}{1-(\lambda L)^{-1}}x_t + d\lambda^t, \\ &= \frac{a}{1-\lambda} - b\sum_{i=0}^{\infty}\left(\frac{1}{\lambda}\right)^{i+1} x_{t+i+1} + d\lambda^t, \end{aligned} \tag{3}$$

where d is constant to be determined by some side condition. An example of such side condition is

$$\lim_{n\to\infty}\sum_{i=n}^{\infty}\left(\frac{1}{\lambda}\right)^{i} x_{t+i} = 0.$$

Equivalence of (2) **and** (3)
If $a = 0$, then for any value of $\lambda \neq 1$ both (2) and (3) represent solution to the difference equation (1). They are simply alternative representation of the solution! The equivalence will hold whenever

$$\frac{b}{1-\lambda L}x_t \tag{4}$$

and

$$b\frac{-(\lambda L)^{-1}}{1-(\lambda L)^{-1}}x_t \tag{5}$$

are both finite for all t.

It often happens that one of them fails to be finite.

- If the sequence $\{x_t\}$ is bounded and $|\lambda| < 1$, the (4) is a convergent sum for all t.
- If the sequence $\{x_t\}$ is bounded and $|\lambda| > 1$, the (5) is a convergent sum for all t.

Since our desire is to impose that the sequence $\{y_t\}$ is bounded, and we do not have sufficient side conditions, then we must set $c = 0$ in (2) or $d = 0$ (3). To see that

- If $\lambda > 1$ and $c > 0$, then
$$\lim_{t\to\infty} c\lambda^t = \infty.$$
- If $\lambda < 1$ and $d < 0$, then
$$\lim_{t\to\infty} d\lambda^t = \infty.$$

All of the above stuff means that we need to solve "stable roots" $|\lambda| < 1$ backward and "unstable roots" $|\lambda| > 1$ forward.

Second-order difference equations

3 Approximating and Solving DSGE Models

3.1 Examples of dynamic macroeconomic models

3.1.1 Real Business Cycle Model

Households

Households chooses consumption labour and capital stock by maximizing their discounted present value of utility stream from consumption and leisure

$$\max \mathrm{E}_0 \sum_{t=0}^{\infty} \beta^t \left\{ \frac{c_t^{1-\sigma}}{1-\sigma} - \phi \frac{l_t^{1+\gamma}}{1+\gamma} \right\}$$

subject to the budget constraint

$$c_t + k_{t+1} = w_t l_t + r_t k_t + (1-\delta) k_t$$

Markets are complete and Arrow-Debreu securities.

Setting up Lagrangean

The Lagrangean of the above problem can be written as

$$\mathcal{L} = \mathrm{E}_0 \sum_{t=0}^{\infty} \Big\{ \beta^t \frac{c_t^{1-\sigma}}{1-\sigma} - \phi \frac{l_t^{1+\gamma}}{1+\gamma} - \lambda_t \left[c_t + k_{t+1} - w_t l_t - r_t k_t - (1-\delta) k_t \right] \Big\}$$

The first order conditions with respect to c_t, c_{t+1}, l_t and k_{t+1}:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial c_t} &= c_t^{-\sigma} - \lambda_t = 0 \\
\frac{\partial \mathcal{L}}{\partial c_{t+1}} &= \beta c_{t+1}^{-\sigma} - \lambda_t = 0 \\
\frac{\partial \mathcal{L}}{\partial l_t} &= -\phi l_t^{\gamma} + \lambda_t w_t = 0 \\
\frac{\partial \mathcal{L}}{\partial k_{t+1}} &= -\lambda_t + \lambda_{t+1} r_{t+1} + \lambda_{t+1}(1-\delta) = 0.
\end{aligned}$$

their optimality conditions

$$\begin{aligned}
c_t^{-\sigma} &= \lambda_t \\
\beta \, \mathrm{E}_t \, c_{t+1}^{-\sigma} &= \lambda_{t+1} \\
\phi l_t^{\gamma} &= \lambda_t w_t \\
\lambda_t &= \lambda_{t+1}(r_{t+1} + 1 - \delta)
\end{aligned}$$

Firms

Maximize profits

$$\max \Pi_t = y_t - w_t l_t - r_t k_t$$

subject to the production technology

$$y_t = A_t k_t^{\alpha} l_t^{1-\alpha}.$$

Optimality conditions

$$\begin{aligned} \alpha A_t k_t^{\alpha-1} l_t^{1-\alpha} &= r_t \\ (1-\alpha) A_t k_t^{\alpha} l_t^{1-\alpha} l_t^{-1} &= w_t. \end{aligned}$$

Investments i_t solves from the law of motion of capital:

$$k_{t+1} = (1-\delta)k_t + i_t.$$

Technical development

Techological development A_t follows AR(1) process:

$$\log A_t = \rho \log A_{t-1} + z_t,$$

where

$$z_t \sim N(0, \sigma_z^2).$$

Parameter ρ determines the persistence of the process.

A competitive equilibrium

$$\begin{aligned} c_t^{-\sigma} &= \beta \, \mathrm{E}_t \, c_{t+1}^{-\sigma} (r_{t+1} + 1 - \delta) \\ \phi l_t^{\gamma} &= \lambda_t w_t \\ r_t &= \alpha A_t k_t^{\alpha-1} l_t^{1-\alpha} \\ w_t &= (1-\alpha) A_t k_t^{\alpha} l_t^{1-\alpha} l_t^{-1} \\ c_t + i_t &= y_t \\ \log A_t &= \rho \log A_{t-1} + z_t. \end{aligned}$$

Solving the model

First, we need to specify the functional forms (or make some assumptions regarding their first and second derivatives).

Second, in this course we focus on algorithms of *linear* rational expectation models: *perturbation methods*. Usual steps involve

- Solving deterministic steady state.
- Linearization: approximating the nonlinear model with the first order Taylor approximation.

3.1.2 The Basic New-Keynesian Model

The basic New-Keynesian model

Let

y_t output gap, $y_t = x_t - x_t^n$, where x_t is actual output and x_t^n is the *natural* level of output.

π_t inflation rate (period-by-period), ie $\log(P_t / P_{t-1})$, (multiply by 400 to get annualized %-change)

$\sigma > 0$ intertemporal elasticity of substitution,

i_t nominal interest rate,

$\rho \equiv -\log \beta$ discount rate (degree of time preference)

E_t conditional expectation operator, $\mathrm{E}(\cdot|\mathcal{F}_t)$, where $\mathcal{F}_t$ contais all information available at (and up-to) period t.

The expectation-augmented IS curve is given by

$$y_t = \mathrm{E}_t\, y_{t+1} - \frac{1}{\sigma}\,(i_t - \mathrm{E}_t\, \pi_{t+1} - \rho) + d_t,$$

where d_t is an *exogenously* given stochastic process, whose properties we specify later. We call it *demand shock*

New-Keynesian Phillips Curve

The New-Keynesian Phillips Curve is given by the following equation

$$\pi_t = \beta\, \mathrm{E}_t\, \pi_{t+1} + \lambda(y_t + m_t),$$

where λ is a parameter and m_t is an *exogenously* given stochastic process, whose other properties we specify later. We call it *mark-up* shock.

Monetary policy rule

The third unknown endogenous variable i_t is determined by the Taylor rule

$$i_t = \rho + \phi_\pi \pi_t + \phi_y y_t.$$

According to the *Taylor principle*, the monetary policy (here Taylor rule) should respond to inflation (deviation from the target — here zero) more than one-to-one, ie $\phi_\pi > 1$.

3.2 Approximating

(Log)linearizing

The *stationarized* decision rules, budget constraints and equilibrium conditions can typically be written in the following form

$$\mathrm{E}_t\, \Psi(Z_{t+1}, Z_t) = 0, \tag{6}$$

where Z_t and 0 are $n \times 1$ vectors. $\Psi(\cdot)$ is a matrix valued function that is continuous and differentiable. The conditional expectation operator E_t uses information up to and including period t.

Note, that it is not restrictive to use the first order system, i.e. having only one lead. The higher order leads/lags may be introduced by augmenting the state vector Z_t (google "companion form").

The goal is to approximate (6) with a linear system, which can then be solved using the methods that will be described in the following chapters.

Notation

Denote the steady state by a variable without time index, Z. Small letter denotes log of original capital letter variable, $z_t \equiv \log(Z_t)$.

The *deterministic* steady state of (6) is

$$\Psi(Z, Z) = 0,$$

Note, that Z, the deterministic steady state of the model is a nonlinear function of the model's parameters μ.

Compute the first-order Taylor approximation around the steady-state Z.

$$0 \approx \Psi(Z, Z) + \underbrace{\frac{\partial \Psi}{\partial Z_t}(Z, Z)}_{\equiv A(\mu)} \times (Z_t - Z) + \underbrace{\frac{\partial \Psi}{\partial Z_{t+1}}(Z, Z)}_{B(\mu)} \times (Z_{t+1} - Z),$$

where $Z_t - Z$ is $n \times 1$, $\partial\Psi(Z,Z)/\partial Z_t$ denotes the Jacobian of $\Psi(Z_{t+1}, Z_t)$ wrt Z_{t+1} evaluated at (Z, Z). To shorten

$$A(\mu)(Z_t - Z) + B(\mu)\,\mathrm{E}_t(Z_{t+1} - Z) = 0.$$

The coefficient matrices are function of deep (model's) parameters μ.

Note that this is a mechanical step that is typically done by the software (e.g. Dynare, Iris). Often the Jacobian may be computed analytically.

Logarithmic approximation

Logarithmic approximation is a special case of above. Note that $Z_t = \exp(\log(Z_t))$, and, as denoted above, $Z_t = \exp(z_t)$.

Suppose we have

$$f(X_t, Y_t) = g(Z_t), \tag{7}$$

with *strictly positive* X, Y, Z (ie the linearization point). The steady state counterpart is $f(X, Y) = g(Z)$.

This simple summarization is, for example, in the slides by Jürg Adamek. (`http://www.vwl.unibe.ch/studies/3076_e/linearisation_slides.pdf`)

Start from replacing $X_t = \exp(\log(X_t))$ in (7),

$$f\left(e^{\log(X_t)}, e^{\log(Y_t)}\right) = g\left(e^{\log(Z_t)}\right),$$

i.e.

$$f\left(e^{x_t}, e^{y_t}\right) = g\left(e^{z_t}\right),$$

Taking first-order Taylor approximations from both sides:

$$f(X,Y) + f_1'(X,Y)X(x_t - x) + f_2'(X,Y)Y(y_t - y) = g(Z) + g'(Z)Z(z_t - z) \tag{8}$$

Often we denote $\hat{x}_t \equiv x_t - x$.

Divide both sides of (8) by $f(X, Y) = g(Z)$ to obtain

$$1 + f_1'(X,Y)X(x_t - x)/f(X,Y) + f_2'(X,Y)Y(y_t - y)/f(X,Y) = 1 + g'(Z)Z(z_t - z)/g(Z) \tag{9}$$

Note that

$$f_1'(X,Y)X/f(X,Y)$$

is the *elasticity* of $f(X_t, Y_t)$ with respect to X_t at the steady-state point.

Also note, that $100 \times (x_t - x)$ tells X_t's relative deviation from the steady-state point.

Warning!

You may only loglinearize *strictly positive* variable. Typical example of a variable that may obtain negative values is the net foreign asset. This needs to be linearized!

Useful log-linearization rules

Denote $\hat{x}_t \equiv \log(X_t) - \log(X)$

$$X_t \approx X(1 + \hat{x}_t)$$
$$X_t^\rho \approx X^\rho(1 + \rho\hat{x}_t)$$
$$aX_t \approx X(1 + \hat{x}_t)$$
$$X_t Y_t \approx XY(1 + \hat{x}_t + \hat{y}_t)$$
$$Y_t(a + bX_t) \approx Y(1a + bX) + aYy_t + bXY(\hat{x}_t + \hat{y}_t)$$
$$Y_t(a + bX_t + cZ_t) \approx Y(a + bX + cZ) + Y(a + bX + cZ)\hat{y}_t + bXY\hat{x}_t + cZY\hat{z}_t$$
$$\frac{X_t}{aY_t} \approx \frac{X}{aY}(1 + \hat{x}_t - \hat{y}_t)$$
$$\frac{X_t}{Y_t + aZ_t} \approx \frac{X}{Y + aZ}\left[1 + \hat{x}_t - \frac{Y}{Y + aZ}\hat{y}_t - \frac{aZ}{Y + aZ}\hat{z}_t\right]$$

3.3 Solving

3.3.1 Blanchard and Kahn method

Blanchard and Kahn method

Blanchard and Kahn (1980) develop a solution method based on the following setup of a model

$$\begin{bmatrix} x_{1t+1} \\ \mathrm{E}_t\, x_{2t+1} \end{bmatrix} = \tilde{A} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} + E f_t, \tag{10}$$

where

- x_{1t} is $n_1 \times 1$ vector of endogenous predetermined variables = variables for which $\mathrm{E}_t\, x_{1t+1} = x_{1t+1}$.
  - These are typically backward-looking variables.
  - (Do not exist with $t+1$.)
  - For example k_{t+1} in the standard RBC model.
- x_{2t} is $n_2 \times 1$ vector of endogenous nonpredetermined variables = for which $x_{2t+1} = \mathrm{E}_t\, x_{2t+1} + \eta_{t+1}$, where η_{t+1} represents an expectational error.
  - forward-looking variables
  - jump-variables
- f_t contains $k \times 1$ vector of exogenous forcing variables: e.g. shock innovations.
- $\tilde{A}$ is full rank.

Use *Jordan normal form*[1] of the matrix $\tilde{A}$ as follows

$$\tilde{A} = \Lambda^{-1} J \Lambda,$$

where J is a *diagonal matrix* consisting of *eigenvalues* of $\tilde{A}$ that are ordered from in increasing value. It is partitioned as follows

$$J = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix},$$

where

- the eigenvalues in J_1 lie on or within the unit circle (*stable* eigenvalues)
- the eigenvalues in J_2 lie outside the unit circle (*unstable* eigenvalues)

Matrix Λ contains the corresponding *eigenvectors*. It is partitioned accordingly (and E too)

$$\Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \quad E = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}$$

Stability condition

Saddle-path stability

If the number of unstable eigenvalues is equal to the number of nonpredetermined variables, the system is said to be *saddle-path stable* and an unique solution exists.

Other cases

1. If the number of unstable eigenvalues exceeds the number of nonpredetermined variables, no solution exists.
2. If the number of unstable eigenvalues is smaller than the number of nonpredetermined variables, there are infinite solutions

[1] See also Spectral decomposition or Eigendecomposition or canonical form

Saddle-path case

Rewrite (10) as

$$\begin{bmatrix} x_{1t+1} \\ \mathrm{E}_t\, x_{2t+1} \end{bmatrix} = \Lambda^{-1} J \Lambda \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} f_t \tag{11}$$

and premultiply by Λ to obtain

$$\begin{bmatrix} \acute{x}_{1t+1} \\ \mathrm{E}_t\, \acute{x}_{2t+1} \end{bmatrix} = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix} \begin{bmatrix} \acute{x}_{1t} \\ \acute{x}_{2t} \end{bmatrix} + \begin{bmatrix} \acute{E}_1 \\ \acute{E}_2 \end{bmatrix} f_t,$$

where

$$\begin{bmatrix} \acute{x}_{1t} \\ \acute{x}_{2t} \end{bmatrix} \equiv \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix}$$
$$\begin{bmatrix} \acute{E}_1 \\ \acute{E}_2 \end{bmatrix} \equiv \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}.$$

- The system is now "de-coupled" in the sense that the nonpredetermined variables are related only to the unstable eigenvalues J_2 of $\tilde{A}$.
- *Hence, we have two "seemingly unrelated" set of equations.*
- As in the univariate case, we derive the solution of the nonpredetermined variables by forward iteration and predetermined variables by backward iteration.
- We start by analysing the lower block, ie the system of the nonpredetermined variables by performing the forward interation.
- Denote f_{2t} of those f_ts that are conformable with $\acute{E}_2$.
- The lower part of (11) is as follows

$$\acute{x}_{2t} = J_2^{-1}\, \mathrm{E}_t\, \acute{x}_{2t+1} - J_2^{-1} \acute{E}_2 f_{2t} \tag{12}$$

- Shift it one period and use the law-of-iterated-expectations ($\mathrm{E}_t(\mathrm{E}_{t+1}\, x_t) = \mathrm{E}_t\, x_t$)

$$\mathrm{E}_t\, \acute{x}_{2t+1} = J_2^{-1}\, \mathrm{E}_t\, \acute{x}_{2t+2} - J_2^{-1} \acute{E}_2\, \mathrm{E}_t\, f_{2t+1}$$

  and substitute it back to (12) to obtain

$$\acute{x}_{2t} = J_2^{-2}\, \mathrm{E}_t\, \acute{x}_{2t+2} - J_2^{-2} \acute{E}_2\, \mathrm{E}_t\, f_{2t+1} - J_2^{-1} \acute{E}_2 f_{2t} \tag{13}$$

- Because J_2 contains the eigenvalues above the unit disc, $\left(J_2^{-1}\right)^n$ will asymptotically vanish. The *iteration results*

$$\acute{x}_{2t} = -\sum_{i=0}^{\infty} J_2^{-(i+1)} \acute{E}_2\, \mathrm{E}_t\, f_{2t+i}$$

- Using the definition $\acute{x}_{2t}$, we may write in to the form

$$x_{2t} = -\Lambda_{22}^{-1}\Lambda_{21} x_{1t} - \Lambda_{22}^{-1} \sum_{i=0}^{\infty} J_2^{-(i+1)} \acute{E}_2\, \mathrm{E}_t\, f_{2t+i} \tag{14}$$

Finally, the upper part of (11) is given by

$$x_{1t+1} = \tilde{A}_{11} x_{1t} + \tilde{A}_{22} x_{2t} + E_1 f_t, \tag{15}$$

where $\tilde{A}_{11}$ and $\tilde{A}_{22}$ are reshuffled $\tilde{A}$ according to the above ordering.

An example with AR(1) shock

Suppose

$$f_{2t+1} = \Phi f_{2t} + \varepsilon_{t+1}, \quad \varepsilon_t \sim \text{IID}(0, \Sigma)$$

and Φ is full rank and its roots are within the unit disc (ie stationary VAR(1)).

Then

$$\mathrm{E}_t\, f_{2t+i} = \Phi^i f_{2t}, \quad i \geq 0$$

and (14) becomes

$$x_{2t} = -\Lambda_{22}^{-1}\Lambda_{21} x_{1t} - \Lambda_{22}^{-1}(I - \Phi J_2^{-1})^{-1}\acute{E}_2 f_{2t} \tag{16}$$

"VAR" form

Stacking sets of equations (14) and (15) gives the following

$$\underbrace{\begin{bmatrix} x_{1t+1} \\ x_{2t+1} \end{bmatrix}}_{\equiv x_{t+1}} = \underbrace{\begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{22} \\ -\Lambda_{22}^{-1}\Lambda_{21} & 0 \end{bmatrix}}_{\equiv F_1} \underbrace{\begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix}}_{\equiv x_t} + \underbrace{\begin{bmatrix} E_1 & 0 \\ 0 & -\Lambda_{22}^{-1}(I - \Phi J_2^{-1})^{-1}\acute{E}_2 \end{bmatrix}}_{\equiv F_\Gamma} \underbrace{\begin{bmatrix} f_{1t} \\ f_{2t} \end{bmatrix}}_{\equiv f_t}. \tag{17}$$

ie

$$x_{t+1} = F_1 x_t + F_\Gamma f_t$$

Issues

- The model variables has to be classified either predetermined or nonpredetermined. $\longrightarrow$ model-specific system reduction may be required.
- $\tilde{A}$ has to be a full rank matrix. Hence, identities are not allowed!

Example: NK model in BK form

Recall the basic New-Keynesian model:

- Expectation-augmented IS curve

$$y_t = \mathrm{E}_t\, y_{t+1} - \frac{1}{\sigma}(i_t - \mathrm{E}_t\, \pi_{t+1} - \rho) + d_t \tag{18}$$

- Phillips curve

$$\pi_t = \beta\, \mathrm{E}_t\, \pi_{t+1} + \lambda(y_t + m_t), \tag{19}$$

- and the monetary policy rule

$$i_t = \rho + \phi_\pi \pi_t + \phi_y y_t. \tag{20}$$

And assume that d_t and m_t are *white noise*, ie independent with zero mean, and a constant variance.

We aim finding the representation (10). Substitute (20) into (18) to obtain the following system of equations. After substitution, all variables y_t and π_t are non-predetermined.

$$\begin{aligned} y_t &= \mathrm{E}_t\, y_{t+1} - \frac{1}{\sigma}\left(\phi_\pi \pi_t + \phi_y y_t - \mathrm{E}_t\, \pi_{t+1}\right) + d_t \\ \pi_t &= \beta\, \mathrm{E}_t\, \pi_{t+1} + \lambda(y_t + m_t). \end{aligned}$$

or

$$\begin{aligned} \left(1 + \frac{\phi_\pi}{\sigma}\right) y_t &= \mathrm{E}_t\, y_{t+1} - \frac{1}{\sigma}(\phi_\pi \pi_t - \mathrm{E}_t\, \pi_{t+1}) + d_t \\ \pi_t &= \beta\, \mathrm{E}_t\, \pi_{t+1} + \lambda(y_t + m_t). \end{aligned}$$

Moving expectation-related terms to the left:

$$\mathrm{E}_t\, y_{t+1} + \frac{1}{\sigma}\,\mathrm{E}_t\,\pi_{t+1} = \left(1 + \frac{\phi_\pi}{\sigma}\right) y_t + \frac{\phi_\pi}{\sigma}\pi_t - d_t$$
$$\beta\,\mathrm{E}_t\,\pi_{t+1} = -\lambda y_t + \pi_t - \lambda m_t.$$

Same in the matrix form

$$\underbrace{\begin{bmatrix}1 & 1/\sigma\\ 0 & \beta\end{bmatrix}}_{A_0}\underbrace{\mathrm{E}_t\begin{bmatrix}y_{t+1}\\ \pi_{t+1}\end{bmatrix}}_{\mathrm{E}_t\, x_{t+1}} = \underbrace{\begin{bmatrix}1+\phi_y/\sigma & \phi_\pi/\sigma\\ -\lambda & 1\end{bmatrix}}_{A_1}\underbrace{\begin{bmatrix}y_t\\ \pi_t\end{bmatrix}}_{x_t} + \underbrace{\begin{bmatrix}-1 & 0\\ 0 & -\lambda\end{bmatrix}}_{E^\star}\underbrace{\begin{bmatrix}d_t\\ m_t\end{bmatrix}}_{f_t}$$

ie

$$A_0\,\mathrm{E}_t\, x_{t+1} = A_1 x_t + E^\star f_t. \tag{21}$$

To obtain form (10), we need to pre-multiply the previous equation by A_0^{-1}. *Naturally, the parameter values need to be determined such that A_0 is invertible.* Then

$$\mathrm{E}_t\, x_{t+1} = \underbrace{A_0^{-1}A_1}_{\equiv\tilde{A}}\, x_t + \underbrace{A_0^{-1}E^\star}_{\equiv E} f_t.$$

The matrices $\tilde{A}$ and E are the following

$$\tilde{A} = \begin{bmatrix}1+\phi_y/\sigma+\lambda/(\beta\sigma) & \phi_\pi/\sigma - 1/(\beta\sigma)\\ -\lambda/\beta & 1/\beta\end{bmatrix}$$
$$E = \begin{bmatrix}-1 & \lambda/(\beta\sigma)\\ 0 & -\lambda/\beta\end{bmatrix}.$$

3.3.2 Klein method

Klein's method

Klein (2000) proposes a method that overcome some of the drawbacks of Blanchard and Kahn (1980) by allowing singular $\tilde{A}$. It is also computationally fast. The system has to be in the form

$$\tilde{A}\,\mathrm{E}_t\, x_{t+1} = \tilde{B}x_t + Ef_t, \tag{22}$$

where f_t ($n_z \times 1$ vector) follows the VAR(1)[2]. $\tilde{A}$ and $\tilde{B}$ are $n \times n$ matrices, and E $n \times n_z$ matrix.

$$f_t = \Phi f_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{IID}(0, \Sigma)$$

and $\tilde{A}$ may be singular (reduced rank). This mean that *we may have static equilibrium conditions (like identities)* in the system.

Decompose x_t to predetermined[3] x_{1t} and nonpredetermined x_{2t} variables as before. Then

$$\mathrm{E}_t\, x_{t+1} = \begin{bmatrix}x_{1t+1}\\ \mathrm{E}_t\, x_{2t+1}\end{bmatrix}.$$

As before the system will be de-coupled according to x_{1t} ($n_1 \times 1$) and x_{2t} ($n_2 \times 1$). Generalized Schur decomposition, that allows singularity is used instead of standard spectral decomposition. Applying it to $\tilde{A}$ and $\tilde{B}$ gives

$$Q\tilde{A}Z = S \tag{23}$$
$$Q\tilde{B}Z = T, \tag{24}$$

[2] We drop the constant term (and other stationary deterministic stuff) to simplify algebra.

[3] Klein (2000) gives more general definition to predetermidness than BK. He also allows "backward-lookingness", which means that the prediction error is martingale difference process.

where Q, Z are *unitary*[4] and S, T *upper triangular* matrices with diagonal elements containing the generalized eigenvalues of $\tilde{A}$ and $\tilde{B}$. Eigenvalues are ordered as above.

Z is partitioned accordingly

$$Z = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}.$$

Z_{11} is $n_1 \times n_1$ and corresponds the stable eigenvalues of the system and, hence, conforms with x_1, ie predetermined variables.

Next we *triangularize* the system (22) to stable and unstable blocks

$$z_t \equiv \begin{bmatrix} s_t \\ u_t \end{bmatrix} = Z^H x_t,$$

where ^{H} denotes the Hermitian transpose. and (23) and (24) can be written as

$$\tilde{A} = Q' S Z^H$$
$$\tilde{B} = Q' T Z^H.$$

Premultiplying the partitioned system by Q we obtain

$$S \, \mathrm{E}_t \, z_{t+1} = T z_t + Q E f_t \tag{25}$$

and since S and T are upper triangular, (22) may be written as

$$\begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{bmatrix} \mathrm{E}_t \begin{bmatrix} s_{t+1} \\ u_{t+1} \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \begin{bmatrix} s_t \\ u_t \end{bmatrix} + \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} E f_t. \tag{26}$$

Due to block-diagonal (recursive) structure, the process u_t is unrelated s_t. We iterate this forward to obtain[5]

$$u_t = -T_{22}^{-1} \sum_{i=0}^{\infty} [T_{22}^{-1} S_{22}]^i Q_2 E \, \mathrm{E}_t \, f_{t+i}.$$

Since f_t is VAR(1), we obtain

$$u_t = M f_t$$
$$\mathrm{vec}\, M = [(\Phi' \otimes S_{22}) - I_n \otimes T_{22}]^{-1} \mathrm{vec}[Q_2 E].$$

This solution of the unstable component is used to solve the stable block, resulting

$$s_{t+1} = S_{11}^{-1} T_{11} s_t + S_{11}^{-1} [T_{12} M - S_{12} M \Phi + Q_1 E] f_t - Z_{11}^{-1} Z_{12} M \varepsilon_{t+1}.$$

Given the definition of u_t and s_t, we may express the solution in terms of original variables.

3.3.3 Method of undetermined coefficients

Method of undetermined coefficients

Following Anderson and Moore (1985) (AiM) and Zagaglia (2005), DSGE model may be written in the form

$$H_{-1} z_{t-1} + H_0 z_t + H_1 \, \mathrm{E}_t \, z_{t+1} = D \eta_t, \tag{27}$$

where z_t is vector of endogenous variables and η_t are pure innovations with zero mean and unit variance.

The solution to (27) takes the form

$$z_t = B_1 z_{t-1} + B_0 \eta_t,$$

[4] $U'U = I$

[5] See appendix B in Klein (2000).

where

$$B_0 = S_0^{-1} D,$$
$$S_0 = H_0 + H_1 B_1.$$

B_1 satisfies the identity

$$H_{-1} + H_0 B_1 + H_1 B_1^2 = 0.$$

Summarizing

The general feature of the solution methods is that they result a "VAR(1)" representation of the model

$$\xi_t = F_0(\mu) + F_1(\mu)\xi_{t-1} + F_\Gamma(\mu) v_t, \tag{28}$$

where

- Variable vector ξ_t are the variables in the model.
- matrices $F_i(\mu)$ are complicated (and large) matrices whose exact form depends on the solution method (See, eg, previous slide)
- They are also highly nonlinear function of the "deep" parameters μ of the economic model
  - The parameters μ include, among others, the parameters specifying the stochastic processes of the model: shock variances, for example
- We may use this representation to analytically calculate various model moments for given parameter values.
- We may use this also for simulating the model.
- Note that we do not know anything about the data at this stage.

4 Moments of Model and Data

4.1 Introduction

The Idea

The ultimate purpose of bringing model and data together is to *validate* the model. This means

- comparing model and data
- using data to calibrate (some) parameters of the model
- figuring out in where the model fits the data and where not

How to do it:

- computing model moments (statistics), and
- comparing them to data moments (statistics)

More challenging task is to *estimate parameters* (or other objects) of the model.

How to compute model moments

(After all the hazzle of the previous section) the model is in the linear form. Many statistics may be computed analytically using the linear form.

For some statistics this is not possible. The simulation (Monte Carlo) alternative:

```
N = 10000; % large number
for iN = 1:N
  draw_shock_innovation;
  relying_on_the_solution_compute_endog._variables;
  compute_test_statistic;
  save_it_to_vector;
end;
my_test_statistic = meanc(test_statistic)
```

Comparing data and model

Note that the data moments are *estimated*!

$\longrightarrow$ there is uncertainty related to these estimates

$\longrightarrow$ use confidence bounds (of data moments) when comparing data moments and model moments

Very often the comparison involves plotting. Do not forget bounds in plots.

Univariate comparisons are straightforward. Multivariate comparison involves identification problem: *impulse responses*

What moments?

Standard moments are

- means: great ratios
- variances: relative to, for example, output
- correlation: with different lags
- similar moments in *frequency* domain
  - spectral density
  - coherence
  - gain
- impulse responses
- forecast error variance decomposition
- other VAR statistics

People are innovative in comparing, for example, *forecasts*!

Structure of this section

1. Introduce univariate or bivariate moments in time domain: auto and cross correlations
2. Same stuff in frequency domain
3. Have a look at multivariate stuff: VARs

Exclusions

I exclude

- univariate time series model (eg ARMA): you should know these!
- nonlinear time series models (eg ARCH): the current model solution methods (as we use them) do not support nonlinearity

4.2 Autocovariance functions and alike

Stochastic process

Definition 1 (Stochastic process)**.** A set of *random variables*

$$\{X_t = (x_{1t}, \ldots, x_{mt})' | t \in \mathcal{T}\}$$

is called *stochastic process*, where t denotes time and X_t is an observation that is related to time t.

Since this is infinite dimensional, we cannot determine its' joint distribution. According to Kolmogorov (1933), under certain regularity conditions we can characterize its joint distribution by its' finite dimensional marginal distributions

$$F_{X_{t_1},\ldots,X_{t_p}} \quad (t_1, \ldots, t_p \in \mathcal{T}).$$

Large part of material of this section are based on the excellent lecture notes by Markku Rahiala (`http://stat.oulu.fi/rahiala/teachmr.html`)

Stochastic process and its properties

Definition 2 (Stationarity)**.** If

$$F_{X_{t_1},\ldots,X_{t_p}} = F_{X_{t_1+\tau},\ldots,X_{t_p+\tau}}$$

for all $p, \tau, t_1, \ldots, t_p$, the process is *strictly stationary*.

If it applies only to the first and second moments, ie

$$\begin{cases} \mathrm{E}\, X_t \equiv \mu = \text{ constant} & \text{and} \\ \Gamma(\tau) = \mathrm{E}\, X_{t+\tau} X_t' - \mu\mu' = \mathrm{cov}\,(X_{t+\tau}, X_t) \end{cases}$$

the process is *weakly stationary*.

Definition 3 (Autocovariance function)**.**

$$\Gamma(\tau) = \mathrm{cov}\,(X_{t+\tau}, X_t)$$

is called *autocovariance function*.

Note that $\Gamma(\tau)' = \Gamma(-\tau)$. A weakly stationary process $\{X_t\}$ that has $\Gamma(\tau) = 0$ (for all $\tau \neq 0$) is called *white noise*.

Definition 4 (Autocovariance-generating function)**.** For each covariance-stationary (weakly stationary) process X_t we calculate the sequence of autcovariances $\{\Gamma(\tau)\}_{\tau=-\infty}^{\infty}$. If this sequence is *absolutely summable*, we may summarize the autocovariances through a scalar-valued function called *autocovariance-generating function*

$$g_X(z) = \sum_{\tau=-\infty}^{\infty} \Gamma(\tau) z^\tau,$$

where z is complex scalar.

Definition 5 (Autocorrelation)**.** Let $\{x_t\}$ be weakly stationary scalar process and $\gamma_x(\tau)$ its autocovariance function. Then

$$\gamma_x(0) = \mathrm{var}(x_t) = \sigma_x^2 = \text{constant}$$

and

$$\rho_x(\tau) = \frac{\gamma_x(\tau)}{\gamma_x(0)} = \frac{\mathrm{cov}(x_{t+\tau}, x_t)}{\mathrm{var}(x_t)}$$

depends only on τ. The sequency $\rho_x(\tau)$ (for all $\tau = 0, \pm 1, \pm 2, \ldots$).

Definition 6 (Crosscorrelation)**.** If $X_t = (x_{1t} \ \cdots \ x_{mt})'$ is multivariate, then function

$$\rho_{x_i,x_j}(\tau) = \frac{\mathrm{cov}(x_{i,t}, x_{j,t+\tau})}{\sqrt{\mathrm{var}(x_{i,t})\,\mathrm{var}(x_{j,t})}}$$

is called *cross correlation* of components $x_{i,t}$ and $x_{j,t}$.

Sample counterparts and some usefull theorems

Definition 7 (Sample autocorrelation)**.** For observations $(x_1, \ldots, x_n$ the *sample autocorrelation* is the sequence

$$r_x(\tau) = \frac{\frac{1}{n-\tau}\sum_{t=1}^{n-\tau}(x_t - \bar{x})(x_{t+\tau} - \bar{x})}{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2}, \quad \tau = 0, \pm 1, \pm 2, \ldots, \pm(n-1),$$

where $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$.

To make inference about the size of sample autocorrelation, the following theorems can be useful

Theorem 8. *Let K be fixed positive integer and* $\{\varepsilon_t\}$ *white noise. Then*

$$\sqrt{n}R = \sqrt{n}(r_\varepsilon(1) \ \cdots \ r_\varepsilon(K))' \sim^{asympt.} N_K(0, I),$$

when n approaches infinity.

And

$$Q_n = (n+2)\sum_{\tau=1}^{K}\left(1 - \frac{\tau}{n}\right) r_\varepsilon(\tau)^2 \sim^{as.} \chi^2_K.$$

when n approaches infinity.

Similar results can be shown to *sample crosscorrelation*.

4.3 Spectral analysis

4.3.1 Univariate Frequency-Domain Methods

Spectral Analysis

The spectral analysis section is based on the course material by Markku Lanna and Henri Nyberg, see the course "Empirical macroeconomics" `http://blogs.helsinki.fi/lanne/teaching`

- In physics and engineering light is considered as electromagnetic waves that consists of in terms of energy, wavelength, or frequency.
- One may use prism to decompose light to spectrum (probably you did this in the high school).
- Similarly a stationary time series can be considered as constant electromagnetic wave.
- Economic time series can be seen as consisting of components with different periodicity.
  - trend, business cycle, seasonal component etc.
- In *frequency-domain* analysis, the idea is to measure the contributions of different periodic components in a series.
  - Different components neet not be identified with regular (fixed) cycles, but there is only a tendency towards cyclical movements centered around a particular frequency.
    - For instance, the business cycle is typically defined as variation with periodicity of 2–8 years.
- Frequency-domain methods are complementary to time-domain methods in studying the properties of, e.g.,
  - detrending methods,
  - seasonal adjustment methods,
  - data revisions, and
  - interrelations between the business cycle components of various economic variables.

Wave length

wave length $\times$ frequency $=$ constant (light speed in physics)

Preliminaries

From high school algebra:

$$\sin(x+y) = \sin(x)\cdot\cos(y) + \cos(x)\cdot\sin(y)$$

$$\cos(x+y) = \cos(x)\cdot\cos(y) - \sin(x)\cdot\sin(y)$$

$$[\sin(x)]^2 + [\cos(x)]^2 \equiv 1$$

from which follows

$$b\cdot\cos(x) + c\cdot\sin(x) = a\cdot\cos(x+\theta),$$

where

$$a = \sqrt{b^2+c^2} \text{ and } \theta = \arctan\left(\frac{c}{b}\right)$$

θ is called *phase angle*. Complex number

$$z = x + i\cdot y \in \mathcal{C} \text{ and conjugate } \bar{z} = x - i\cdot y \in \mathcal{C}$$

Euler's formula

$$e^z = e^{x+i\cdot y} = e^x e^{i\cdot y} = e^x(\cos(y) + i\sin(y))$$

Spectral representation theorem

Any covariance-stationary process X_t can be expressed as sum of cycles

$$X_t = \mu + \int_0^\pi [\alpha(\omega)\cos(\omega t) + \delta(\omega)\sin(\omega t)]d\omega.$$

The random processes $\alpha(\cdot)$ and $\delta(\cdot)$ have zero mean and they are not correlated across frequencies.

Fixed Cycles

- Although economic time series are not characterized by cycles with fixed periodicity, let us first, for simplicity, consider fixed cycles to introduce the main concepts.
- For example, the cosine function is periodic, i.e., it produces a fixed cycle.
  - $y = \cos(x)$ goes through its full complement of values (one full cycle) as x (measured in radians) moves from 0 to 2π.
  - The same pattern is repeated such that for any integer k, $\cos(x + 2\pi k) = \cos(x)$.
  - By defining $x = \omega t$, where ω is measured in radians and t is time, $y = \cos(\omega t)$ becomes a function of time t with fixed *frequency* ω.
    - By setting ω equal to different values, y can be made to expand or contract in time.
    - For small ω (low frequency) it takes a long time for y to go through the entire cycle.
    - For big ω (high frequency) it takes a short time for y to go through the entire cycle.
    - Example: try to play with command `plot cos(180*(pi/180)*(x-0))` in the site `www.wolframalpha.com`
- Each frequency ω corresponds to the *period* of the cycle (denoted by p), which is the time taken for y to go through its complete sequence of values.
- Given ω, how is the period p computed?
  - $\cos(\omega t) = \cos(\omega t + 2\pi)$, so how many periods does it take for ωt to increase by 2π?
    - $2\pi/\omega$ periods.
  - A cycle with period 4 thus repeats itself every four periods and has frequency $\omega = 2\pi/4 = \pi/2 \approx 1.57$ radians.

Fixed Cycles

- The range of y is controlled by multiplying $\cos(\omega t)$ by the *amplitude*, ρ.
- The location of y along the time axis can be shifted by introducing the *phase*, θ.
- Incorporating the amplitude and phase into y yields

$$y = \rho\cos(\omega t - \theta) = \rho\cos[\omega(t - \xi)],$$

where $\xi = \theta/\omega$ gives the shift in terms of time.

Example 9. Suppose the cosine function has a period of 4 so that it peaks at $t = 0, 4, 8, \ldots$. If we want it to have peaks at $t = 2, 6, 10, \ldots$, i.e. $\xi = 2$, the phase $\theta = \xi\omega = 2(2\pi/4) = \pi$. Note, however, that this phase shift is not unambiguous because the same effect would have been obtained by setting $\xi = -2$, indicating $\theta = \xi\omega = -2(2\pi/4) = -\pi$.

Spectrum

- The *(power) spectrum* (or *spectral density*) gives the decomposition of the variance of a time series process in terms of frequency.
- The spectral density function $s_y(\omega)$ of a weakly stationary process y_t is obtained by evaluating the autocovariance- generating function (AGF) at $e^{-i\omega}$ and standardizing by 2π:

$$s_y(\omega) = \frac{1}{2\pi} g_y\left(e^{-i\omega}\right) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma_j e^{-i\omega j},$$

where $i = \sqrt{-1}$.

- By de Moivre's theorem,

$$e^{-i\omega j} = \cos(\omega j) - i\sin(\omega j),$$

we obtain

$$s_y(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma_j[\cos(\omega j) - i\sin(\omega j)],$$

and making use of the fact that for a weakly stationary process $\gamma_{-j} = \gamma_j$ and well-known results from trigonometry, $s_y(\omega)$ simplifies to

$$s_y(\omega) = \frac{1}{2\pi}\left[\gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j\cos(\omega j)\right].$$

Properties

- Assuming the sequence of autocovariances $\{\gamma_j\}_{j=-\infty}^{\infty}$ is absolutely summable, the spectrum of y_t is a continuous, real-valued function of ω.
- If the process y_t is weakly stationary, the spectrum is nonnegative for all ω.
- It suffices to consider $s_y(\omega)$ in the range $[0, \pi]$.
  - The spectrum is symmetric around zero, because $\cos(-\omega j) = \cos(\omega j)$.
  - The spectrum is a periodic function of ω, i.e., $s_y(\omega + 2\pi k) = s_y(\omega)$, because $\cos[(\omega + 2\pi k)j] = \cos(\omega j)$ for any integers k and j.

Example: White Noise

- To get some intuition of how the spectrum decomposes the variance of a time series in terms frequency, let us first, for simplicity, consider the white noise process.

- For a white noise process $y_t = \varepsilon_t$, $\gamma_0 = \sigma_\varepsilon^2$ and $\gamma_j = 0$ for $j \neq 0$.
  - Hence, the AGF $g_y(z) = \sigma_\varepsilon^2$ and the spectrum is flat:
$$s_y(\omega) = \frac{1}{2\pi}\gamma_0 = \frac{\sigma_\varepsilon^2}{2\pi}.$$
  - The area under the spectrum over the range $[-\pi, \pi]$ equals σ_ε^2:
$$\int_{-\pi}^{\pi} s_y(\omega)\, d\omega = \frac{\sigma_\varepsilon^2}{2\pi}(\pi + \pi) = \sigma_\varepsilon^2.$$

Fact 10. *The area under the spectrum over the range $[-\pi, \pi]$ always equals the variance of y_t. Comparisons of the height of $s_y(\omega)$ for different values of ω indicate the relative importance of fluctuations at the chosen frequencies in influencing variations in y_t.*

- It is useful to relate the radians ω to the associated period (the number of units of time it takes the cyclical component with frequency ω to complete a cycle), $p = 2\pi/\omega$.

  Example 11. The period associated with frequency $\omega = \pi/2$ is $2\pi/\omega = 2\pi/(\pi/2) = 4$. With annual data, a cyclical component with frequency $\pi/2$ thus corresponds to a cycle with a periodicity of 4 years. With quarterly data it corresponds to a cycle with a periodicity of 4 quarters or 1 year.

Example: MA(1)

- The AGF of the MA(1) process is
$$g_y(z) = \sigma_\varepsilon^2 (1 + \theta_1 z)\left(1 + \theta_1 z^{-1}\right).$$

  Thus
$$\begin{aligned} s_y(\omega) &= \frac{\sigma_\varepsilon^2}{2\pi} g_y\left(e^{-i\omega}\right) \\ &= \frac{\sigma_\varepsilon^2}{2\pi}\left(1 + \theta_1 e^{-i\omega}\right)\left(1 + \theta_1 e^{i\omega}\right) \\ &= \frac{\sigma_\varepsilon^2}{2\pi}\left(1 + \theta_1 e^{i\omega} + \theta_1 e^{-i\omega} + \theta_1^2 e^{i\omega - i\omega}\right) \\ &= \frac{\sigma_\varepsilon^2}{2\pi}\left(1 + \theta_1^2 + 2\theta_1 \cos(\omega)\right). \end{aligned}$$

Example: AR(1)

- The AGF of the AR(1) process is
$$g_y(z) = \sigma_\varepsilon^2 \frac{1}{\phi(z)\,\phi(z^{-1})} = \frac{1}{(1 - \phi_1 z)(1 - \phi_1 z^{-1})}.$$

  Thus
$$\begin{aligned} s_y(\omega) &= \frac{\sigma_\varepsilon^2}{2\pi} g_y\left(e^{-i\omega}\right) \\ &= \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{\left(1 - \phi_1 e^{-i\omega}\right)\left(1 - \phi_1 e^{i\omega}\right)} \\ &= \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{\left(1 - \phi_1 e^{-i\omega}\right)\left(1 - \phi_1 e^{i\omega}\right)} \\ &= \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{1 - \phi_1 e^{-i\omega} - \phi_1 e^{i\omega} + \phi_1^2} \\ &= \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{1 + \phi_1^2 - 2\phi_1 \cos(\omega)}. \end{aligned}$$

Example: AR(2)

- The spectrum of the AR(2) process need not be monotonically decreasing, but its peak depends on the values of ϕ_1 and ϕ_2.
- A peak at ω^* indicates a *tendency* towards a cycle at a frequency around ω^*.
  - This stochastic cycle is often called a *pseudo cycle* as the cyclical movements are not regular.
- The peak of the spectrum can be found by setting the derivative of $s_y(\omega)$ equal to zero and solving for ω.
- The AGF of an AR(2) process is

$$\begin{aligned} g_y(z) &= \sigma_\varepsilon^2 \frac{1}{\phi(z)\,\phi(z^{-1})} \\ &= \frac{\sigma_\varepsilon^2}{(1-\phi_1 z-\phi_2 z^2)(1-\phi_1 z^{-1}-\phi_2 z^{-2})}. \end{aligned}$$

- Thus

$$\begin{aligned} s_y(\omega) &= \frac{\sigma_\varepsilon^2}{2\pi} g_y\left(e^{-i\omega}\right) \\ &= \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{\left(1-\phi_1 e^{-i\omega}-\phi_2 e^{-i2\omega}\right)\left(1-\phi_1 e^{i\omega}-\phi_2 e^{i2\omega}\right)} \\ &= \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{1+\phi_1^2+\phi_2^2-2\phi_1(1-\phi_2)\cos(\omega)-2\phi_2\cos(2\omega)} \end{aligned}$$

4.3.2 Estimation

Estimation of Spectrum Sample Periodogram

- The spectrum of an observed time series can, in principle, be estimated by replacing the autocovariances γ_j in the theoretical spectrum

$$s_y(\omega) = \frac{1}{2\pi}\left[\gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j\cos(\omega j)\right]$$

  by their estimates $\widehat{\gamma}_j$ to obtain the sample periodogram

$$\widehat{s}_y(\omega) = \frac{1}{2\pi}\left[\widehat{\gamma}_0 + 2\sum_{j=1}^{T-1}\widehat{\gamma}_j\cos(\omega j)\right].$$

- This yields a very jagged and irregular estimate of the spectrum.
  - Because we are, effectively, estimating T parameters with T observations, this pattern persists irrespective of the sample size.

Nonparametric Estimation

- It is reasonable to assume that $s_y(\omega)$ will be close to $s_y(\lambda)$ when ω is close to λ.
- This suggests that $s_y(\omega)$ might be estimated by a weighted average of the values of $\widehat{s}_y(\lambda)$ for values of λ in the neighborhood around ω, with the weights depending on the distance between ω and λ:

$$\widehat{s}_y(\omega_j) = \sum_{m=-h}^{h}\kappa(\omega_{j+m},\omega_j)\,\widehat{s}_y(\omega_{j+m}).$$

- h is a *bandwidth* parameter indicating how many frequencies around ω_j are included in estimating $\widehat{s}_y(\omega_j)$.
- The *kernel* $\kappa(\omega_{j+m}, \omega_{j+m})$ gives the weight of each frequency, and these weights sum to unity,

$$\sum_{m=-h}^{h} \kappa(\omega_{j+m}, \omega_j) = 1.$$

Modified Daniell Kernel

- Several kernels are available for the nonparametric estimation of the spectrum.
- The default kernel in *R* is the modified Daniell kernel, possibly used repeatedly.
  - The kernel puts half weights at end points.
    - For example, with bandwidth parameter $h = 1$, the weights of $\widehat{s}_y(\omega_{j-1})$, $\widehat{s}_y(\omega_j)$ and $\widehat{s}_y(\omega_{j+1})$ are $1/4$, $2/4$ and $1/4$, respectively.
  - Applying the same kernel again yields weights $1/16, 4/16, 6/16, 4/16, 1/16$ for $\widehat{s}_y(\omega_{j-2}), ..., \widehat{s}_y(\omega_{j+2})$, respectively.
- Choosing the bandwidth parameter is arbitrary. One possibility is to plot the estimated spectrum based on several different bandwidths and rely on subjective judgment.

Autoregressive Spectral Estimation

- An estimate of the spectrum can be based on an AR(p) model fitted to the observed time series:
  1. Estimate an adequate AR(p) model for y_t.
  2. Plug the estimated coefficients $\widehat{\phi}_1, \widehat{\phi}_2, ..., \widehat{\phi}_p$ in the expression of $s_y(\omega)$ derived for the AR(p) model.
- The order of the AR model, p, must be carefully chosen.
  - If p is too small, i.e., the AR model does not adequately capture the dynamics of y_t, the ensuing spectrum may be misleading.
  - If p is too large, the spectrum may be inaccurately estimated.

4.3.3 Linear Filters

Linear Filters

- Many commonly used methods of extracting the business cycle component of macroeconomic time series (and seasonal adjustment methods) can be expressed as linear filters,

$$y_t = \sum_{j=-r}^{s} w_j x_{t-j}$$

  where, say, x_t is the original series and y_t its business cycle component.
- A filter changes the relative importance of the various cyclical components. To see how different frequencies are affected, the spectra of the unfiltered and filtered series can be compared.
- A more direct way is to inspect the *squared frequency response function* that gives the factor by which the filter alters the spectrum of y_t.
- The filter may also induce a shift in the series with respect to time, and this is revealed by the *phase diagram*.

Squared Frequency Response Function Assuming the original series x_t has an infinite-order MA representation $x_t = \psi(L)\varepsilon_t$, the filtered series can be written as $y_t = W(L)x_t = W(L)\psi(L)\varepsilon_t$, where $W(L) = w_{-r}L^{-r} + \cdots + w_{-1}L^{-1} + w_0 + w_1L + \cdots w_sL^s$.

The spectrum of y_t thus equals

$$\begin{aligned} s_y(\omega) &= \frac{1}{2\pi}g_y\left(e^{-i\omega}\right) \\ &= \frac{\sigma_\varepsilon^2}{2\pi}W\left(e^{-i\omega}\right)\psi\left(e^{-i\omega}\right)W\left(e^{i\omega}\right)\psi\left(e^{i\omega}\right) \\ &= \frac{\sigma_\varepsilon^2}{2\pi}W\left(e^{-i\omega}\right)W\left(e^{i\omega}\right)\psi\left(e^{-i\omega}\right)\psi\left(e^{i\omega}\right) \\ &= W\left(e^{-i\omega}\right)W\left(e^{i\omega}\right)s_x(\omega). \end{aligned}$$

The term $W\left(e^{-i\omega}\right)W\left(e^{i\omega}\right)$ is called the squared frequency response function and it shows how the filtering changes the spectrum of the original series.

Squared Frequency Response Function:

Example 12. Consider taking first differences of a variable x_t,

$$y_t = x_t - x_{t-1} = (1-L)x_t = W(L)x_t.$$

Now

$$s_y(\omega) = W\left(e^{-i\omega}\right)W\left(e^{i\omega}\right)s_x(\omega),$$

i.e., the squared frequency response function equals

$$\begin{aligned} W\left(e^{-i\omega}\right)W\left(e^{i\omega}\right) &= \left(1-e^{-i\omega}\right)\left(1-e^{i\omega}\right) \\ &= 2 - e^{i\omega} - e^{-i\omega} \\ &= 2 - [\cos(\omega) - i\sin(\omega) + \cos(\omega) + i\sin(\omega)] \\ &= 2[1-\cos(\omega)]. \end{aligned}$$

Linear filters: Gain

In the literature, the properties of filters are often illustrated using the *gain* function, $G(\omega)$, which is just the modulus of the squared frequency response function (squared gain):

$$G(\omega) = \sqrt{W\left(e^{-i\omega}\right)W\left(e^{i\omega}\right)} \equiv \left|W\left(e^{-i\omega}\right)\right|.$$

Example 13. For the first-difference filter $(1-L)$, the gain function is given by

$$\begin{aligned} G(\omega) &= \sqrt{\left(1-e^{-i\omega}\right)\left(1-e^{i\omega}\right)} \\ &= \sqrt{2[1-\cos(\omega)]}. \end{aligned}$$

This carries the same information as the squared frequency response function.

Try `plot sqrt(2*(1-cos(w)))`

- Given the gain function $G(\omega)$, the relationship between $s_y(\omega)$ and $s_x(\omega)$ can be written

$$s_y(\omega) = G(\omega)^2 s_x(\omega).$$

- The gain function expresses how the employed filter serve to isolate cycles.
  - Gain function with the value 0 indicates that the filter eliminates those frequencies.

Linear filters: Phase Diagram

- The squared frequency response function $W\left(e^{-i\omega}\right) W\left(e^{i\omega}\right)$ takes on real values only. However, the frequency response function $W\left(e^{-i\omega}\right)$ is, in general a complex quantity,

$$W\left(e^{-i\omega}\right) = W^{*}(\omega) + iW^{\dagger}(\omega)$$

  where $W^{*}(\omega)$ and $W^{\dagger}(\omega)$ are both real.

- The phase diagram, i.e., the slope of the phase function

$$Ph(\omega) = \tan^{-1}\left[-W^{\dagger}(\omega)/W^{*}(\omega)\right]$$

  gives the delay in time periods.

The phase diagram measures the shift that how much the lead-lag relationships in time series are altered by the filter.

Example 14. Consider a filter that shifts the series back by three time periods,

$$y_t = x_{t-3} = L^3 x_t = W(L)\, x_t.$$

The frequency reponse function equals $W\left(e^{-i\omega}\right) = e^{-3i\omega} = \cos(3\omega) - i\sin(3\omega)$ Thus $Ph(\omega) = \tan^{-1}\left[\sin(3\omega)/\cos(3\omega)\right] = \tan^{-1}\left[\tan(3\omega)\right] = 3\omega$. In other words, the phase diagram is a straight line with slope 3.

Phase Diagram: Symmetric Filter

Most filters used to extract the business cycle component of macroeconomic time series are symmetric.

For a symmetric linear filter, the phase function equals zero, i.e., they exhibit no phase shift.

The frequency response function equals

$$\begin{aligned} W\left(e^{-i\omega}\right) &= w_{-r}e^{-ri\omega} + \cdots + w_{-1}e^{-i\omega} + w_0 \\ &\quad + w_1 e^{i\omega} + \cdots + w_r e^{ri\omega} \\ &= \sum_{j=-r}^{r} w_j \cos(\omega j) - i\sum_{j=-r}^{r} w_j \sin(\omega j) \\ &= w_0 + 2\sum_{j=1}^{r} \cos(\omega j). \end{aligned}$$

Hence, $W^{\dagger}(\omega) = 0$, indicating that $Ph(\omega) = \tan^{-1}\left[-W^{\dagger}(\omega)/W^{*}(\omega)\right] = 0$.

Cross-Spectral Analysis

- In addition to the spectra of single time series, the relationships between pairs of variables can be examined in the frequency domain.

- The spectrum can be generalized for the vector case, and the cross spectrum between y_t and x_t is defined analogously with the spectrum of a single series,

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

- Instead of the cross spectrum, functions derived from it (phase and coherence) allow for convenient interpretation of the relationship between two variables in the frequency domain.

- Given data on y and x, the cross spectrum and the derived functions can be computed analogously to the power spectrum. In particular, some smoothing is required to obtain consistent estimators of the phase and coherence.

The cross-spectral density function can be expressed in terms of its real component $c_{yx}(\omega)$ (the cospectrum) and imaginary component q_{yx} (the quadrature spectrum)

$$s_{yx}(\omega) = c_{yx}(\omega) + i\, q_{yx}(\omega).$$

In polar form, the cross-spectral density can be written

$$s_{yx}(\omega) = R(\omega)\, e^{i\theta(\omega)},$$

where

$$R(\omega) = \sqrt{c_{yx}(\omega)^2 + q_{yx}(\omega)^2}$$

and

$$\theta(\omega) = \arctan\frac{-q_{yx}(\omega)}{c_{yx}(\omega)}.$$

- The function $R(\omega)$ is the gain function.
- $\theta(\omega)$ is called the phase function.

Phase Diagram

- The phase function has the same interpretation as in the case of a linear filter.
  - If in the plot of the phase function against ω (the phase diagram), the phase function is a straight line over some frequency band, the direction of the slope tells which series is leading and the amount of the slope gives the extent of the lag.
    - The lag need not be an integer number.
    - If the slope is zero there is no lead-lag relationship.
    - If the slope is positive, the first variable is lagging.
    - If the slope is negative, the first variable leading.
- A confidence band may be computed around the estimated phase function to evaluate statistical significance.

Coherence

- The coherence measures the strength of the relationship between two variables, y_t and x_t, at different frequencies.
- The (squared) coherence is defined analogously to correlation:

  $$C(\omega) = \frac{\left|s_{yx}(\omega)\right|^2}{s_x(\omega)\, s_y(\omega)},$$

  and it can be interpreted as a measure of the correlation between the series y and x at different frequencies.
- $0 \leq C(\omega) \leq 1$
- If $C(\omega)$ is near one, the ω-frequency components of y and x are highly (linearly) related, and if $C(\omega)$ is near zero, these components are only slightly related.
- A confidence band can be used to evaluate the statistical significance.

4.4 Measuring Business Cycles

(This relies on Lanne's and Nyberg's material).

Methods for Extracting the Cyclical Component

- A number of methods are commonly used to extract the business cycle component of macroeconomic time series.
- In this course, we concentrate on:
  - Linear detrending (fitting a linear trend model)
  - Differencing
  - Beveridge-Nelson decomposition
  - Filters
- As discussed above, a reasonable detrending method must
  1. allow for variation over time in the growth rate and
  2. ensure that short-term fluctuations are correctly categorized as cyclical deviations from trend.
- In addition, the detrended time series must be stationary, i.e., its frequency response must be zero at frequency zero.

Linear Detrending and Differencing

- These two methods are conducted under the implicit assumption that the growth rate of y_t is constant.
- In linear detrending, a constant term and a linear trend is regressed to y_t. In other words, $y_t = \alpha_0 + \alpha_1 t + c_t$).
  - The estimated cyclical component is $\hat{c}_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 t$.
  - Parameter α_1, and its estimate $\hat{\alpha}_1$, can be interpreted as a trend coefficient.

Linear Detrending and Differencing

- The first-order difference leads to the cyclical component

$$c_t = y_t - y_{t-1}.$$

- The choice between detrending and differencing depends on which one provides a more appropriate representation of $y_t = \log(Y_t)$.
- In practice, the choice of either specification is problematic as the data do not appear to follow a constant average growth rate throughout the sample period.

Beveridge-Nelson Decomposition

- Starting point: Any I(1) process y_t consists of the stationarity component η_t and the permanent (trend) component ξ_t, $y_t = \xi_t + \eta_t$.
- Any I(1) process y_t whose first difference Δy_t satisfies

$$\Delta y_t = \psi(L)\,\varepsilon_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

where $\sum_{j=0}^{\infty} j\,|\psi_j| < \infty$, $E(\Delta y_t) = 0$, with ε_t a white noise process, can be written as the sum of

  - a random walk $\psi(1)\varepsilon_t$ or equivalently $\Delta\xi_t = \psi(1)\varepsilon_t$
  - a stationary process η_t, and
  - initial conditions $y_0 - \eta_0$.
- η_t is often interpreted as the cyclical component c_t of the series y_t.

Beveridge-Nelson Decomposition In Practice

The Beveridge-Nelson decomposition can be applied in the following steps

1. Identify an appropriate ARMA model for Δy_t, estimate the parameters $\psi(1)$ and save the residuals $\hat{\varepsilon}_t$.
2. Given the initial value of ξ_0, the trend component can be generated following the presentation $\Delta\widehat{\xi}_t = \hat{\psi}(1)\hat{\varepsilon}_t$.
3. Construct the cyclical component $\hat{c}_t = y_t - \hat{\xi}_t$.

Example 15. Assume that ARMA(1,1) model is an appropriate model for Δy_t. This means that $\psi(1) = \frac{\theta(1)}{\phi(1)} = \frac{1+\theta_1}{1-\phi_1}$. Given the initial value ξ_0, the estimated parameters $\hat{\theta}_1$ and $\hat{\phi}_1$ and residuals $\hat{\varepsilon}_t$, the trend component can be constructed as $\Delta\widehat{\xi}_t = \frac{1+\hat{\theta}_1}{1-\hat{\phi}_1}\hat{\varepsilon}_t$.

Filters Using Filters to Isolate Cycles

- Filters are tools designed to eliminate the influence of cyclical variation at various frequencies.
- Low frequencies are associated with cycles of long periods of oscillation (infrequent shifts from a peak to trough) and vice versa with high frequencies.
  - Slowly evolving trends (i.e. cycles with an infinite periodicity) are associated with very low frequencies.
  - Constant trend is associated with zero frequency $\omega = 0$.

High-Pass and Band-Pass Filters

- Two kinds of filters are used to extract the business cycle component (with period of 2–8 years) of quarterly time series.
  - A *high-pass filter* passes only components with periodicity less than or equal to 32 quarters (or with frequency ω greater than or equal to $2\pi/32 = \pi/16 \approx 0.20$).
    - The frequency response function equals 0 for $\omega < \pi/16$ and 1 for $\omega \geq \pi/16$.
    - A high-pass filter passes also variation associated with seasonal frequencies, so it can only be applied to seasonally adjusted data.
  - A *band-pass filter* passes only components with periodicity between 8 and 32 quarters (or with frequency between $2\pi/8 = \pi/4 \approx 0.79$ and $2\pi/32 = \pi/16 \approx 0.20$).
    - The frequency response function equals 1 for $\pi/16 \leq \omega \leq \pi/4$ and zero otherwise.
- With a finite number of observations it is not possible to filter out exactly the desired frequencies.

Hodrick-Prescott Filter

- The Hodrick-Prescott filtering is probably the most commonly used method of extracting business cycle components in macroeconomics.
- The general idea is to compute the growth (trend) component g_t and cyclical component c_t of y_t by minimizing the magnitude

$$\sum_{t=1}^{T} \underbrace{(y_t - g_t)}_{c_t}^2 + \lambda \sum_{t=1}^{T-1} \left[(g_{t+1} - g_t) - (g_t - g_{t-1})\right]^2.$$

  - The second term minimizes the changes in the growth rate over time while the first term bringing g_t as close as possible to the observed series y_t.
  - The smoothing parameter λ tells how much weight is given to the first objective.
    - If $\lambda = 0$, $g_t = y_t$ (no smoothing).
    - The greater λ is, the smoother the growth component. When $\lambda \to \infty$, g_t is a straight line.
- Using the lag operator L, the cyclical component produced by the Hodrick-Prescott filter can be written as (see, Baxter and King, 1999)

$$c_t = \left[\frac{\lambda (1-L)^2 \left(1-L^{-1}\right)^2}{1 + \lambda (1-L)^2 \left(1-L^{-1}\right)^2}\right] y_t.$$

  - The filter removes unit root components.
  - The filter is a symmetric infinite-order two-sided moving average.
    - c_t depends on past and future values of y_t, $t = 1, 2, ..., T$. Therefore, the first and last values of c_t are inaccurate and may change considerably as new observations become available.
    - The filter introduces no phase shift.

Hodrick-Prescott Filter

- With quarterly data λ is most often set to 1600, which produces a filter reasonably close to the ideal high-pass filter.
- With annual data three values of λ, 400, 100 and 10 are the most common. In this case the filter is not so close to the ideal filter.

Baxter-King Filter

- The cyclical component extracted by all high-pass and band-pass filters are infinite two-sided symmetric moving averages of y_t,

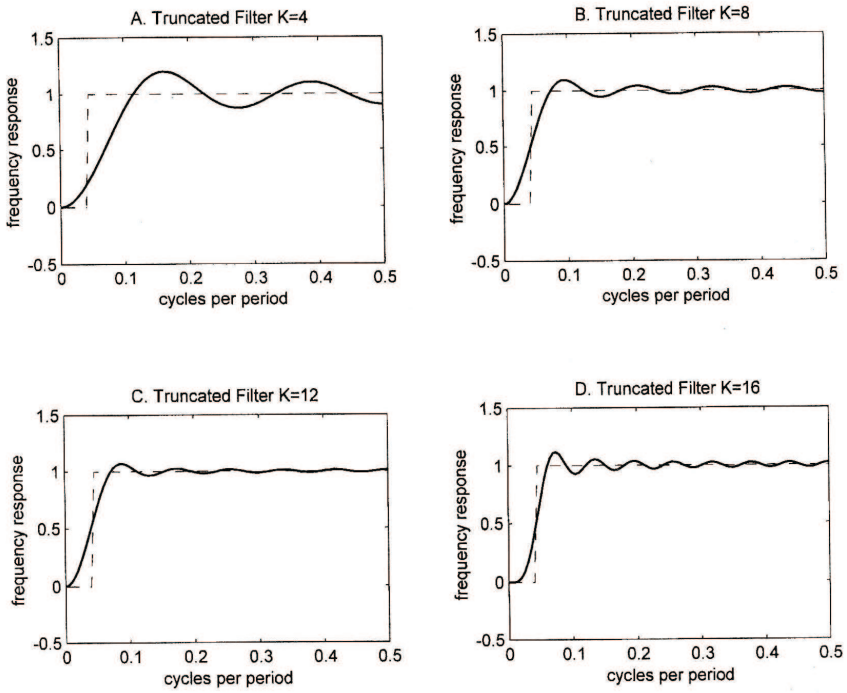
$$c_t = \sum_{j=-\infty}^{\infty} w_j y_{t-j}.$$

- Baxter and King suggest approximating these filters by truncating,

$$c_t = \sum_{j=-K}^{K} w_j y_{t-j},$$

  and selecting the weights w_j based on the frequencies that are filtered out.

  - K observations at the beginning and end of the series are lost.



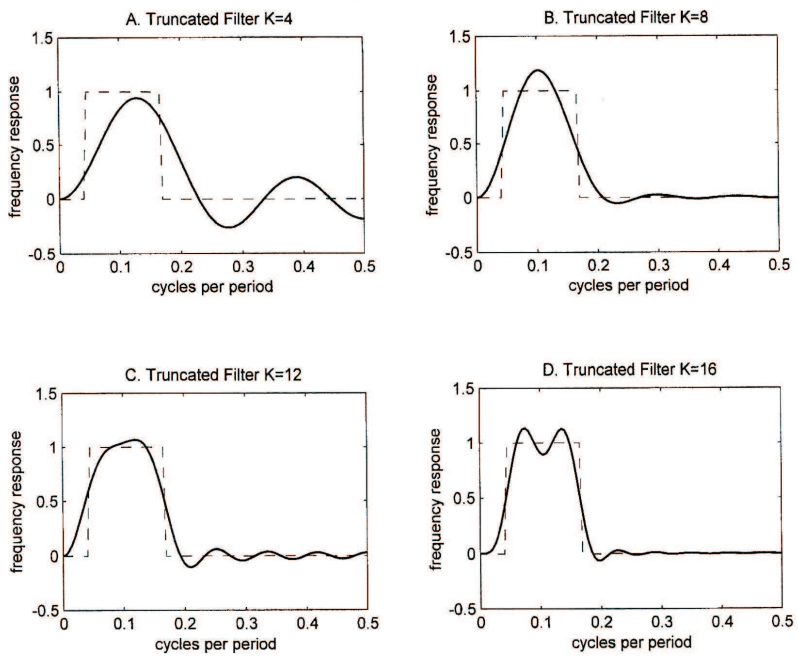
Baxter-King Filter

Baxter-King Filter

4.5 Multivariate time series models

This material is based on professor Lanne's "Macroeconometrics" course.

VAR model



The VAR(p) model is a generalization of the univariate AR(p) model. For example, a two-variable VAR(2) model can be written as

$$\begin{aligned} y_{1t} &= \phi_{1,11}y_{1,t-1} + \phi_{1,12}y_{2,t-1} + \phi_{2,11}y_{1,t-2} + \phi_{2,12}y_{2,t-2} + u_{1t} \\ y_{2t} &= \phi_{1,21}y_{1,t-1} + \phi_{1,22}y_{2,t-1} + \phi_{2,21}y_{1,t-2} + \phi_{2,22}y_{2,t-2} + u_{2t} \end{aligned}$$

or in matrix form,

$$\begin{aligned} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = & \begin{bmatrix} \phi_{1,11} & \phi_{1,12} \\ \phi_{1,21} & \phi_{1,22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \\ & \begin{bmatrix} \phi_{2,11} & \phi_{2,12} \\ \phi_{2,21} & \phi_{2,22} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \end{aligned}$$

or

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \mathbf{u}_t.$$

Let $\mathbf{y}_t = (y_{1t} \ \cdots \ y_{Kt})'$ be a $K \times 1$ vector.

Definition 16 (VAR(p) model)**.** VAR(p) model is

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \cdots + \Phi_p \mathbf{y}_{t-p} + \mathbf{u}_t \quad \mathbf{u}_t \sim NID_m(0, \Sigma).$$

Denote matrix polynomial $\Phi(L) \equiv I - \Phi_1 L - \cdots - \Phi_p L^p$.

$\mathbf{u}_t = (u_{1t}, ..., u_{Kt})'$ is a K-dimensional *white noise* process:

$$\begin{array}{l} E(\mathbf{u}_t) = \mathbf{0}, \\ E(\mathbf{u}_t \mathbf{u}_t') = \Sigma_u, \text{ and} \\ E(\mathbf{u}_t \mathbf{u}_s') = \mathbf{0} \text{ for } s \neq t. \end{array}$$

In subsequent derivations, it is often convenient to write the VAR(p) model using the lag operator L,

$$\mathbf{y}_t = \Phi_1 L \mathbf{y}_t + \Phi_2 L^2 \mathbf{y}_t + \cdots + \Phi_p L^p \mathbf{y}_t + \mathbf{u}_t$$

or

$$\left(I_K - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p\right) \mathbf{y}_t = \mathbf{u}_t$$

or

$$\Phi(L) \mathbf{y}_t = \mathbf{u}_t.$$

- The VAR(p) model may also contain deterministic terms such as an intercept, linear trend or seasonal dummies.
- The VAR(p) model is a simple linear model to describe the joint probability distribution of the data.
- It is hard to interprete and, as such, rarely tells nothing about the cause and effect.

Companion form

Sometimes it is more convenient to use the so-called companion form of the VAR(p) model. The VAR(p) process

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \cdots + \Phi_p \mathbf{y}_{t-p} + \mathbf{u}_t$$

can be written as the following VAR(1) process

$$\xi_t = \mathbf{F} \xi_{t-1} + \mathbf{v}_t,$$

where

$$\xi_t \equiv \begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-p+1} \end{bmatrix},$$

$$\mathbf{F} \equiv \begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p \\ \mathbf{I}_K & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_K & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_K & \mathbf{0} \end{bmatrix},$$

$$\mathbf{v}_t \equiv \begin{bmatrix} \mathbf{u}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

and

$$E\left(\mathbf{v}_t\mathbf{v}_s'\right) = \begin{cases} \mathbf{Q} & t = s \\ \mathbf{0} & \text{otherwise} \end{cases}$$

with

$$\mathbf{Q} \equiv \begin{bmatrix} \Sigma_u & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}.$$

Stationarity

The VAR(p) process $\mathbf{y}_t$ is covariance-stationary (weakly stationary) if $E\left(\mathbf{y}_t\right)$ and $E\left(\mathbf{y}_t\mathbf{y}_{t-j}'\right)$ do not depend on the date t. In this case, the consequences of any given shock $\mathbf{u}_t$ must eventually die out. Weak stationarity imposes the following stability condition: a VAR(p) process $\mathbf{y}_t$ is weakly stationary if the roots of the equation

$$\left|I_K - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p\right| = 0$$

are greater than unity in absolute value. To see how this condition arises, consider for simplicity the VAR(1) process

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \mathbf{u}_t.$$

By *recursive substitution*, this can be written as

$$\mathbf{y}_t = \Phi_1^{j+1}\mathbf{y}_{t-j-1} + \sum_{i=0}^{j} \Phi_1^i \mathbf{u}_{t-i},$$

and the effect of any shock will die out only if $\Phi_1^j \to 0$ as $j \to \infty$. From linear algebra we know that this, in turn, is the case only if all the eigenvalues of Φ_1 are less than unity in absolute value, which is equivalent to the equation

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

having all roots greater than unity in absolute value. Because a VAR(p) process can be written in the companion form, this condition holds for a general VAR process. It can generally be expressed as the matrix $\mathbf{F}$ of the VAR(1) form having all eigenvalues less than unity in absolute value or the equation

$$|I_K - \mathbf{F}z| = \left|I_K - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p\right| = 0$$

having all roots greater than unity in absolute value.

Estimation and testing

- The unrestricted stationary VAR can be estimated *consistently* by OLS, equation by equation.
- If the distribution of $\mathbf{u}_t$ is known, it can be estimated by maximum likelihood.
- Under normality of $\mathbf{u}_t$, ML estimator is numerically equal to OLS. (Otherwise quasi-ML)
- Under regularity conditions (incl. stationarity), the OLS (ML) is asymptotically normal $\longrightarrow$ standard t-tests can be applied

Model checking

The error term $\mathbf{u}_t$ is assumed to be normal white noise. The properties of the *estimated* error term (residuals) may be studied by

- autocorrelation, and cross-correlation of individual residual series
- Pormanteau test and LM-test for (vector) autocorrrelation.
- Autoregressive conditional heteroscedasticity may be studied by *squared* residuals.
- and normality by Jarque-Bera test.

Moving average representation

The roots of the VAR representation corresponds to the mp factors of equation

$$\det\left[\Phi(s^{-1})\right] = 0.$$

If they lie inside of the unit circle, and all elements in X_t are stationary the process X_t is stationary and we have the Wold decomposition

$$\mathbf{y}_t = \Phi(L)^{-1}\mathbf{u}_t = \sum_{j=0}^{\infty} \Psi_j \mathbf{u}_{t-j},$$

where $m \times m$ coefficient matrices Ψ_j $(j = 0, 1, 2, \dots)$ are called *impulse responses*. They can be computed recursively from

$$\Psi_i = \Phi_1\Psi_{i-1} + \Phi_2\Psi_{i-2} + \cdots + \Phi_p\Psi_{i-p} \quad i = 1, 2, \dots,$$

with $\Psi_0 = I_m$ and $\Psi_i = 0$ for $i < 0$. They tell how earlier innovations $\mathbf{u}_{t-j}$ are reflected by $\mathbf{y}_t$ observations.

The MA(∞) representation allows for tracing out the dynamic effects of shocks to the variables. i.e.,

$$\Psi_s = \frac{\partial \mathbf{y}_{t+s}}{\partial \mathbf{u}_t'}.$$

The row i, column j element of Ψ_s gives the effect of a one-unit increase in the error of the jth variable at time t on the ith variable at time $t + s$, holding all other errors constant. For instance, the first column of Ψ_1 gives the effect in period 1 of a unit increase in the error of the first variable in period 0 on each variable in the system.

Structural VAR

The problem is that each component in $\mathbf{u}_t$ may correlate with another component. Identification problem arises. The innovations $\mathbf{u}_t$ may be *orthogonalized* by Cholesky decomposition:

Let ε_t be orthogonalized innovation so that

$$\mathrm{E}\,\varepsilon_t\varepsilon_t' = I.$$

Define a matrix B such that

$$B^{-1}\Sigma B'^{-1} = I.$$

which implies

$$\Sigma = BB'. \tag{29}$$

ε_t may be constructed using

$$\varepsilon_t = B^{-1}\mathbf{u}_t$$

Note, that this is not the only way to make the innovations orthogonal!

The Cholesky decomposition of Σ is a lower-triangular matrix that satisfies (29), with diagonal elements containing the square root of the diagonal elements of Σ . Note, that the ordering of the variables in $\mathbf{y}_t$ affects B. It is not unique in that sense.

The whole *structural VAR* litterature aims finding (economic) theoretically consistent ways to orthogonalize innovations.

Structural VAR: B-model

The recursive structure implied by the Choleski decomposition may not always be reasonable.

Let us keep assuming that the errors of the VAR model are linear combinations of the economic shocks

$$\mathbf{u}_t = \mathbf{B}\varepsilon_t$$

and let us, for simplicity assume that $"_t$ has covariance matrix $\mathbf{I}_K$. As seen above, this implies that

$$\Sigma_u = \mathbf{B}\mathbf{B}'.$$

- Because Σ_u is symmetric, this involves $K(K+1)/2$ different equations.
- However, there are K^2 elements in the matrix $\mathbf{B}$, so that we need $K^2 - K(K+1)/2 = K(K-1)/2$ further restrictions to identify all the K^2 elements.
- The $K(K-1)/2$ restrictions must be inferred from economic theory or institutional knowledge. This happens in Choleski decomposition.
- Once a sufficient number of restrictions have been found, the structural VAR model imposing these restrictions can be estimated by ML and its dynamic properties examined through impulse response analysis.
- The VAR model based on the assumption that $\mathbf{u}_t = \mathbf{B}''_t$ is called the B-model (Lütkepohl 2005, Chapter 9) and it is probably the most commonly employed structural VAR model.

Structural VAR: AB-model

A natural extension of the B-model is the so-called AB-model, where

$$\mathbf{A}\mathbf{u}_t = \mathbf{B}\varepsilon_t.$$

In this model, the structural shocks are identified by modelling the instantaneous relations between the observable variables directly.

Here

$$\mathbf{u}_t = \mathbf{A}^{-1}\mathbf{B}\varepsilon_t,$$

and, hence,

$$\Sigma_u = \mathbf{A}^{-1}\mathbf{B}\mathbf{B}'\mathbf{A}^{-1}.$$

Thus, again, we have $K(K+1)/2$ restrictions, but now the matrices $\mathbf{A}$ and $\mathbf{B}$ have together $2K^2$ parameters, and, therefore, we need $2K^2 - K(K+1)/2$ additional restrictions.

With enough restrictions, the model can be estimated by restricted ML.

Note that the B-model is a special case of the AB-model, where $\mathbf{A}$ is the identity matrix. By restricting $\mathbf{B}$ to the identity matrix, we obtain the so-called A-model.

Forecast error variance decomposition

In addition to impulse response functions, the effects of the identified structural shocks can be studied by means of their proportional contribution to the forecast error variance of each of the variables at different forecast horizons. Let us first consider the VAR(1) model

$$\mathbf{y}_t = \Phi_1\mathbf{y}_{t-1} + \mathbf{u}_t.$$

Previously it was shown that in this case the forecast error of $\mathbf{y}_{t+h}$ is

$$\mathbf{y}_{t+h} - E_t(\mathbf{y}_{t+h}) = \mathbf{y}_{t+h} - \Phi_1^h\mathbf{y}_t = \sum_{j=0}^{h-1}\Phi_1^j\mathbf{u}_{t+h-i}$$

and, thus, the forecast error covariance matrix is

$$\begin{aligned}\Sigma_y(h) &= E\left[\left(\sum_{j=0}^{h-1}\Phi_1^j\mathbf{u}_{t+h-i}\right)\left(\sum_{j=0}^{h-1}\Phi_1^j\mathbf{u}_{t+h-i}\right)'\right] = \sum_{j=0}^{h-1}\Phi_1^j\Sigma_u\left(\Phi_1^j\right)' \\ &= \Sigma_u + \Phi_1\Sigma_u\Phi_1' + \Phi_1^2\Sigma_u\left(\Phi_1^2\right)' + \cdots + \Phi_1^{h-1}\Sigma_u\left(\Phi_1^{h-1}\right)'\end{aligned}$$

Because any VAR(p) model can be written as a VAR(1) model in companion form, for a general VAR(p) model, this is

$$\Sigma_y(h) = \Sigma_u + \Psi_1\Sigma_u\Psi_1' + \Psi_2\Sigma_u\Psi_2' + \cdots + \Psi_{h-1}\Sigma_u\Psi_{h-1}',$$

where Ψ_js are the coefficient matrices of the MA(∞) representation.

In terms of the structural shocks ε_t with diagonal covariance matrix Σ_ε such that $\mathbf{u}_t = \mathbf{B}\varepsilon_t$, this can be written as

$$\Sigma_y(h) = \Theta_0\Sigma_\varepsilon\Theta_0' + \Theta_1\Sigma_\varepsilon\Theta_1' + \Theta_2\Sigma_\varepsilon\Theta_2' + \cdots + \Theta_{h-1}\Sigma_\varepsilon\Theta_{h-1},$$

where $\Theta_j = \Psi_j\mathbf{B}$.

The forecast error variance of the kth variable is thus the kth diagonal element of $\Sigma_y(h)$,

$$\begin{aligned}\sigma_k^2(h) &= \sum_{n=0}^{h-1}\left(\theta_{n,k1}^2\sigma_1^2 + \theta_{n,k2}^2\sigma_2^2 + \cdots + \theta_{n,kK}^2\sigma_K^2\right)\\ &= \sum_{j=1}^{K}\sigma_j^2\left(\theta_{0,kj}^2 + \theta_{1,kj}^2 + \cdots + \theta_{h-1,kj}^2\right),\end{aligned}$$

and the contribution of the jth shock ε_{kt} at horizon h to this is $\sigma_j^2\left(\theta_{0,kj}^2 + \theta_{1,kj}^2 + \cdots + \theta_{h-1,kj}^2\right)$. Dividing the latter by $\sigma_k^2(h)$ hence gives the proportional contribution of the jth shock to variable k.

By plotting the proportions of the forecast error variance accounted for by each shock at diffent forecast horizons, their importance can be assessed. For instance, it may be concluded that a shock is important only for certain variables or that its contribution for all variables is minor etc.

Some VAR based statistics

Definition 17 (Contemporaneous variance-covariance matrix)**.** Let

$$\Gamma(0) = \mathrm{E}\,\xi_t\xi_t' = \mathrm{E}(\mathbf{F}\xi_{t-1} + \mathbf{v}_t)(\mathbf{F}\xi_{t-1} + \mathbf{v}_t)' = F\Gamma(0)F' + \Sigma,$$

where $\Gamma(0)$ denotes the *contemporaneous variance-covariance matrix* of ξ_t. The solution of above is given by

$$\mathrm{vec}\,(\Gamma(0)) = (I - \mathbf{F}\otimes\mathbf{F})^{-1}\,\mathrm{vec}\,(\Sigma).$$

Definition 18 (τ^{th} order covariance matrix)**.**

$$\Gamma(\tau) = \mathrm{E}\,z_t z_{t-\tau}'$$

Note that

$$\Gamma(1) = \mathrm{E}(\xi_t\xi_{t-1}' = \mathrm{E}(\mathbf{F}\xi_{t-1} + \mathbf{v}_t)\xi_{t-1}' = \mathbf{F}\Gamma(0),$$

and, in general,

$$\Gamma(\tau) = \mathbf{F}^\tau\Gamma(0)$$

4.6 Filtering the data and model

To pre-filter

Often the theoretical model show no growth. The balanced growth path issues are then ignored.

Ironically, the RBC models that build on the neo-classical growth model do not — as a prototype — model the economic growth.

Data portrays growth
$\longrightarrow$ in order to compare data and model moments, one needs to get rid of the growth in data.
$\longrightarrow$ detrend the data!

To pre-filter the data

The idea is to remove certain frequencies of the data

- Seasonal filters remove seasonality
- Hodrick-Prescott filter removes the trend
- Band-pass filter removes chosen frequencies

Univariate filters

- Hodrick-Prescott filter
- Band-pass filter
- Baxter-King filter
- Beverage-Nelson filter
- Seasonal filters

Multivariate filters

To pre-filter or not to pre-filter

Problems in pre-filtering the data

Do not take into account the restriction by the theoretical model: number of trends, style of trends.

Filtering is subsitute of poor modeling of trends/balanced growth path.

Filtering is never perfect. Leakage.

5 Matching moments

Matching moments

Moment is a fancy word to describe the statistics we use to compare model and data.

There is no clear guide which moments to match. It may depend on

- purpose of the modeling: policy vs. forecast, ...
- style of the model: for example growth vs. dynamics or both!
- frequency: short-run vs long-run
- ...

In the previous chapter we had large list of moments. It is easy to invent more!

Benefits of calibration/moment matching

Get aquainted with your model!

5.1 Calibrating steady-state

Three sets of parameters

We may typically partition the parameters of the theoretical model μ into three sets

1. Parameters that affect only the steady-state
2. Parameters that does not affect steady state but affect dynamics (temporal dependency) of the system
3. Parameters that affect both

We may utilize the separation of the sets in calibration.

How to calibrate steady-state: moments of the steady state

Consider the first set of parameters, ie parameters that affect only the steady-state.
They can be chosen such as they exactly match

- great ratios in the data: I/K, C/Y, I/Y, $(X-M)/Y$, ...
- average growth rates in the data: Δy, π_t.

given the chosen parameters of the set 3.

How to do it

You may follow the procedure

1. Choose the data and *the sample*!
2. Compute the great ratios that can be found from the model too. Plot them to check that they really are stationary
3. Compute sample average.
4. Solve the steady-state model
   (a) Compute the same great ratios using steady-state model
   (b) Try another set of parameter values
5. Repeat the above until the distance between great ratios computed from the data and from the model are minimized

Often conflict: Matching certain moment results unmatching other one.
$\longrightarrow$ try GMM and do it formally (statistically consistent way)

5.2 Comparing model and data moments

Calibrating dynamics: moments of temporal dependence

These are difficult to calibrate!

You may try the following moments

- Cross-correlations
- Coherence could be useful.
- Impulse response function
  - look the response at impact,
  - the shape,
  - the persistence
  - Identification problem: how to compute the impulse responses we observe in the data. Structural VAR!

5.3 GMM

The generalized method of moments

Let w_t denote the variables we observe at time t and let Θ be the unknown parameters of interest. They are related by the function $h(\Theta, w_t)$, where the dimension of the function vector is l.

- The orthogonality conditions lies at the hearth of the generalized methods of moments.
- The orthogonality condition says that when evaluated at Θ^* (the true value), the unconditional expections satisfy

$$\mathrm{E}[h(\Theta^*, w_t)] = 0, \tag{30}$$

- The sample mean of these conditions are

$$g(\Theta; w_1, \ldots, w_T) \equiv \frac{1}{T} \sum_{t=1}^{T} h(\Theta, w_t),$$

- The GMM objective function to be minimized is

$$Q(\Theta) = g(\Theta; w_1, \ldots, w_T)' W g(\Theta; w_1, \ldots, w_T), \tag{31}$$

  where W is the weighting matrix of the orthogonality conditions. Any positive definite matrix is ok, but optimal weighting is obtained by using the covariance matrix of the orthogonality conditions S_T.

- The estimator of the covariance matrix is the following

$$S = \lim_{T \to \infty} (1/T) \sum_{t=1}^{T} \sum_{v=-\infty}^{\infty} E\left[h(\Theta^*, w_t) h(\Theta^*, w_{t-v})'\right],$$

  where Θ^* denotes the true value of Θ. I have used

  - the VARHAC estimator by den Haan and Levin (1996) and
  - quadratic the spectral kernel estimator by Newey and West (1994) that allow autocorrelation (and heteroscedasticity) in the sample orthogonality conditions.

- If the number of orthogonality conditions, l, exceeds the number of parameters, j, the model is overidentified. Hansen (1982) shows that it is possible to test the *overidentification restrictions* (*J*-test), since

$$\left[\sqrt{T} g(\hat{\Theta}, w_1, \ldots, w_T)\right]' \hat{S}_T^{-1} \left[\sqrt{T} g(\hat{\Theta}, w_1, \ldots, w_T)\right] \xrightarrow{d} \chi^2(l - j),$$

  where $\xrightarrow{d}$ denotes convergence in distribution.

- Problems and solutions

  - In the GMM estimation, we encounter the problem of defining the orthogonality conditions (instruments in the typical case).
  - The GMM estimator is different for different set of orthogonality conditions.
  - According to the simulation experiments of Tauchen (1986) and Kocherlakota (1990), increasing the number of orthogonality conditions reduces the estimators' variance but increases the bias in small samples.
  - In the iteration of the GMM objective (31), follow the guidence of Hansen et al. (1996): Since the weighting matrix in the objective function is also a function of the parameters, it is useful to iterate that as well. This, of course, increases the computational burden.

- Extensions (that follow GMM spirit)

  - Simulated method of moments: extends the generalized method of moments to cases where theoretical moment functions cannot be evaluated directly, such as when moment functions involve high-dimensional integrals.
  - Indirect inference: introduce auxiliary model and match the simulated and data moments of the auxialiary model. (Use, eg, VAR as the auxiliary model) If the auxiliary model is correctly specified this is *maximum likelihood*.
  - Empirical likelihood
  - ...

6 Filtering and Likelihood approach

6.1 Kalman filter and the maximum likelihood

The state space form

In following I follow the notation of Hamilton (1994) and Yada manual but the logic of Harvey (1989, Structural Time Series Models). Harvey's setup is more general than that of Hamilton.

Definition 19 (Measurement equation)**.** Let y_t be multivariate time series with n elements. They are observed and related to an $r \times 1$ *state* vector ξ_t via *measurement* equation

$$y_t = H_t \xi_t + x_t + w_t, \quad t = 1, \ldots, T \tag{32}$$

where H_t is an $n \times r$ matrix, x_t is $k \times 1$ vector and w_t is ab $n \times 1$ vector of serially uncorrelated disturbances with mean zero and covariance matrix R_t:

$$\mathrm{E}\, w_t = 0 \text{ and } \operatorname{var}(w_t) = R_t.$$

In general the elements of ξ_t are not observable.

Definition 20 (Transition equation, state equation)**.** The state variables ξ_t are generated by a first-order Markov process

$$\xi_t = F_t \xi_{t-1} + c_t + B_t v_t, \tag{33}$$

where F_t is an $r \times r$ matrix, c_t is an $r \times 1$ vector, B_t is an $m \times g$ matrix and v_t is a $g \times 1$ vector of serially uncorrelated disturbances with mean zero and covariance matrix Q_t, that is

$$\mathrm{E}\, v_t = 0 \text{ and } \operatorname{var}(v_t) = Q_t.$$

Assumption 1.

$$\mathrm{E}\, \xi_0 = \hat{\xi}_0 \textit{ and } \operatorname{var}(\xi_0) = P_0$$

Assumption 2.

$$\mathrm{E}(w_t v_k') = 0, \textit{ for all } k, t = 1, \ldots, T$$
$$\mathrm{E}(w_t \xi_0') = 0, \quad \mathrm{E}(v_t \xi_0') = 0 \textit{ for } t = 1, \ldots, T$$

The first line in above assumption may be relaxed.

Matrices F_t, x_t, R_t, H_t, c_t, B_t and Q_t are called *system matrices*. If they do not change over time, the model is said to be *time-invariant* or *time-homogenous*. Stationary models are special case.

State space form and the economic model

Equation (28) on slide 44 looks familiar:

$$\xi_t = F_0(\mu) + F_1(\mu)\xi_{t-1} + F_\Gamma(\mu) v_t, \tag{28}$$

Redefining $c_t = F_0(\mu)$ and $B_t = F_\Gamma(\mu)$, the solution of the economic model is the transition (state) equation of the state space representation of (32) and (33). *This makes the state space representation attractive!*

With the measurement equation, we are able to map the (theoretical) model variables to actual data!

The Kalman filter

Definition 21 (The Kalman filter)**.** The Kalman filter is a recursive procedure for computing the *optimal estimator* of the state vector $\hat{\xi}_{t|t}$ at time t, based on the information available at time t.

Combined with an assumption that the disturbances and the initial state vector are normally distributed, it enables the likelihood function to be calculated via what is known as the prediction error decomposition. The derivation of Kalman filter below rests on that assumption.

Conditional multivariate normal

This a sidestep to statistics. We need this result in constructing the Kalman filter!

If $X \sim N(\mu, \Sigma)$ where μ and Σ are mean and covariance of multivariate normal distribution and are partitioned as follows

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

then the distribution of X_1 conditional on $X_2 = a$ is multivariate normal, i.e. $(X_1|X_2 = a) \sim N(\bar{\mu}, \bar{\Sigma})$, where

$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$$

and covariance matrix

$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Derivation of the Kalman filter

Consider the state space representation of (32) and (33).

Under normality, ξ_0 is multivariate normal with mean $\hat{\xi}_0$ and covariance $P_0 = \mathrm{E}(\xi_0 - \hat{\xi}_0)(\xi_0 - \hat{\xi}_0)'$.

$t = 1$

$$\xi_1 = F_1\xi_0 + c_1 + B_1 v_1$$

$\longrightarrow$ Due to normality, the conditional expectation of ξ_1 is

$$\hat{\xi}_{1|0} = F_1\hat{\xi}_0 + c_1$$

and its conditional covariance matrix

$$\begin{aligned} P_{1|0} &= \mathrm{E}(\xi_1 - \hat{\xi}_{1|0})(\xi_1 - \hat{\xi}_{1|0})' \\ &= \mathrm{E}(F_1\xi_0 + c_1 + B_1 v_1 - F_1\hat{\xi}_0 - c_1)(F_1\xi_0 \\ &\quad + c_1 + B_1 v_1 - F_1\hat{\xi}_0 - c_1)' \\ &= \mathrm{E}[F_1(\xi_0 - \hat{\xi}_0) + B_1 v_1][F_1(\xi_0 - \hat{\xi}_0) + B_1 v_1]' \\ &= \mathrm{E}[F_1(\xi_0 - \hat{\xi}_0) + B_1 v_1][(\xi_0 - \hat{\xi}_0)'F_1' + v_1'B_1'] \\ &= \mathrm{E}[F_1(\xi_0 - \hat{\xi}_0)(\xi_0 - \hat{\xi}_0)'F_1' + F_1(\xi_0 - \hat{\xi}_0)v_1'B_1' \\ &\quad + B_1 v_1(\xi_0 - \hat{\xi}_0)'F_1' + B_1 v_1 v_1'B_1'] \\ &= F_1 P_0 F_1' + B_1 Q_1 B_1', \end{aligned} \tag{34}$$

since — by assumption — $\mathrm{E}(\xi_0 - \hat{\xi}_0)v_1' = 0$.

For the distribution of ξ_1 conditional on y_1 write

$$\begin{aligned} \xi_1 &= & \hat{\xi}_{1|0} + & & (\xi_1 - \hat{\xi}_{1|0}) \\ y_1 &= H_1 & \hat{\xi}_{1|0} + x_1 + H_1 & & (\xi_1 - \hat{\xi}_{1|0}) + w_1. \end{aligned}$$

The second equation is rearranged measurement equation (32).

$\longrightarrow$ vector $[\xi_1' \; y_1']'$ is multivariate normal with mean

$$[\hat{\xi}_{1|0}' \; (H_1\hat{\xi}_{1|0} + x_1)']'$$

and a covariance

$$\begin{bmatrix} P_{1|0} & P_{1|0}H_1' \\ H_1P_{1|0} & H_1P_{1|0}H_1' + R_1 \end{bmatrix}.$$

Applying the conditional distribution of multivariate normal to the case: ξ_1 conditional on particular value of y_1 results the following mean

$$\hat{\xi}_1 = \hat{\xi}_{1|0} + P_{1|0}H_1'S_1^{-1}(y_1 - H_1\hat{\xi}_{1|0} - x_1)$$

and the covariance

$$P_1 = P_{1|0} - P_{1|0}H_1'S_1^{-1}H_1P_{1|0},$$

where

$$S_1 = H_1P_{1|0}H_1' + R_1$$

Iterating this over $t = 2, 3, \ldots, T$ gives us the Kalman filter!

The Kalman filter

Consider, again, the state space representation of (32) and (33).

Define $\hat{\xi}_{t-1}$ denotes the optimal estimator of ξ_{t-1} based on the observations up to and including y_{t-1}.

P_{t-1} denotes the $r \times r$ covariance matrix of the estimations error

$$P_{t-1} = \mathrm{E}\left[(\xi_{t-1} - \hat{\xi}_{t-1})(\xi_{t-1} - \hat{\xi}_{t-1})'\right]$$

Prediction equations Given $\hat{\xi}_{t-1}$ and P_{t-1}, the optimal estimator of ξ_t is given by

$$\hat{\xi}_{t|t-1} = F_t\hat{\xi}_{t-1} + c_t$$

and the covariance matrix of estimation error is

$$P_{t|t-1} = F_tP_{t-1}F_t' + B_tQ_tB_t', \quad t = 1, \ldots, T$$

Updating equations Once the new observation y_t becomes available, the estimator $\hat{\xi}_{t|t-1}$ of ξ_t can be updated

$$\hat{\xi}_t = \hat{\xi}_{t|t-1} + P_{t|t-1}H_1'S_t^{-1}(y_t - H_t\hat{\xi}_{t|t-1} - x_t)$$

and

$$P_t = P_{t|t-1} - P_{t|t-1}H_t'S_t^{-1}H_tP_{t|t-1},$$

where

$$S_t = H_tP_{t|t-1}H_t' + R_t$$

The equations in these two slides make up the Kalman filter.

They can be written in the single set of recursion from $\hat{\xi}_{t-1}$ to $\hat{\xi}_t$ or from $\hat{\xi}_{t|t-1}$ as follows

$$\hat{\xi}_{t+1|t} = (F_{t+1} - K_tH_t)\hat{\xi}_{t|t-1} + K_ty_t + (c_{t+1} - K_tx_t),$$

where the *Kalman gain* is given by

$$K_t = F_{t+1}P_{t|t-1}H_t'S_t^{-1}, \quad t = 1, \ldots, T.$$

The recursion for the error covariance matrix is

$$P_{t+1|t} = F_{t+1}(P_{t|t-1} - P_{t|t-1}H_t'S_t^{-1}H_tP_{t|t-1})F_{t+1}' + B_{t+1}Q_{t+1}B_{t+1}', \; t = 1, \ldots, T$$

It is know as a *Riccati equation*.

Issues

- The state space representation and Kalman filter tells nothing how to interprete the state variables ξ_t. *Economic model does tell!* Its elements may or may not be identifiable.
- Same economic model may have several state space representations. The one that has smallest state vector is called *minimal realisation*.
- Sometimes transition equation is written in the form where state variables are leaded with one period.
- The system matrices F_t, R_t, H_t, B_t and Q_t may depend on a set of unknown parameters, and one of the main statistical tasks will often be the estimation of these parameters.
- Given initial conditions, the Kalman filter delivers the optimal estimator of the state vector.
- *Information filter* is a version of the Kalman filter, where the recursion involves P^{-1}. This is handy when P_0 is infinite or when system matrices are time-varying.
- *Square root filter* is the one where P is obtained via Cholesky decomposition. It is known to be numerically stable.
- $$\hat{\xi} = \mathrm{E}_t(\xi_t) = \mathrm{E}(\xi|Y_t)$$
  and
  $$P_t = \mathrm{E}_t\left\{\left[\xi_t - \mathrm{E}_t(\xi_t)\right]\left[\xi_t - \mathrm{E}_t(\xi_t)\right]'\right\}$$
  Hence, the conditional mean is the *minimum mean square estimate* of ξ_t.
- The conditional mean can also be regarded as an estimator of ξ_t. Hence it applies to any set of observations. It is also *minimum mean square estimator* of ξ_t.
- In the above case P_t can be considered as the the *mean square error* (MSE) matrix of the estimator.
- For *time invariant* model the necessary and sufficient condition for *stability* is that the characteristic roots of the transition matrix F should have modulus less than one.

Maximum likelihood estimation and the prediction error decomposition

Classical theory of maximun likelihood estimation is based on a situation in which the T sets of observations, $y_1, \ldots, y_T$ are independently and identically distributed. The joint density function is given by

$$L(y;\phi) = \prod_{t=1}^{T} p(y_t),$$

where $p(y_t)$ is the joint probability density function (p.d.f.) of the t-th set of observations. Once the observations have been made, $L(y;\phi)$ is reinterpreted as a likelihood function and the ML estimator is found by maximizing this function wrt ϕ.

In time series or economic models observations are typically not independent. We write instead the conditional probability density function

$$L(y;\phi) = \prod_{t=1}^{T} p(y_t|Y_{t-1}),$$

where $Y_{t-1} = \{y_1, y_2, \ldots, y_{t-1}\}$.

If the disturbances v_t and w_t and initial state vector ξ_0 are multivariate normal, $p(y_t|Y_{t-1})$ is itself normal with the mean and covariance given by Kalman filter.

- Conditionally Y_{t-1}, $\xi_t \sim N(\hat{\xi}_{t|t-1}, P_{t|t-1})$.
- From the measurement equation

$$\tilde{y}_{t|t-1} \equiv \mathrm{E}_{t-1}(y_t) = H_t \hat{\xi}_{t|t-1} + x_t$$

  with covariance matrix S_t.
- Define $\tilde{v}_t = y_t - \mathrm{E}_{t-1}\, y_t = y_t - \tilde{y}_{t|t-1}$

Then the likelihood can be written as

$$\log L = -\frac{nT}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log|S_t| - \frac{1}{2}\sum_{t=1}^{T}\tilde{v}_t' S_t^{-1}\tilde{v}_t.$$

- Since $\tilde{y}_{t|t-1}$ is also the MMSE of y_t, $\tilde{v}_t$ can be interpreted as vector of *prediction errors*. The above likelihood is, therefore, called *prediction error decomposition* form of the likelihood.
- Consequently, the likelihood function may easily computed with Kalman filter.
- And numerically maximized wrt unknown parameter ϕ.

ML issues

- The limiting distribution of the ML estimator of ϕ is normal with covariance matrix obtained as the inverse of the information matrix
  - ϕ has to be interior point
  - derivatives of $\log L$ exists up to order three, and they are continuous in the neighbourhood of true ϕ
  - ϕ is *identifiable*
- In nonstationary models, the influence of the initial point does not vanish. No steady-state point exists. Then, the KF has to be initialized using first observations and "large" matrix P_0. This is called *diffuse prior*.
- Forecasting is easy. Multistep forecast may be obtained by applying KF prediction equations repeteadly. Note: $P_{T+j|T}$ do not take into account the uncertainty related to estimating unknown parameters.

Smoothing

Filtering is prediction/forecasting. In *smoothing* we want to know the distribution of ξ_t conditional on all the smaple, ie $E(\xi_t|Y_T)$.

- it is known as *smoothed estimate*
- and the corresponding estimator as *smoother*
- several algorithms: fixed point, fixed lag, fixed interval

Covariance matrix of $\xi_{t|T}$ *conditional on all T observations*

$$P_{t|T} = \mathrm{E}_T\left[(\xi_t - \hat{\xi}_{t|T})(\xi_t - \hat{\xi}_{t|T})'\right]$$

when $\hat{\xi}_{t|T}$ is viewed as an estimator, $P_{t|T} \leq P_t$ $(t = 1, \ldots, T)$ is its MSE matrix.

Fixed-interval smoothing

Starts with final quantities $\hat{\xi}_T$ and P_T of the Kalman filter and work backwards

$$\hat{\xi}_{t|T} = \hat{\xi}_t + P_t^\star(\hat{\xi}_{t+1|T} - F_{t+1}\hat{\xi}_t)$$

and

$$P_{t|T} = P_t + P_t^\star(P_{t+1|T} - P_{t+1|t})P_t^{\star\prime}$$

where

$$P_t^\star = P_t F_{t+1}' P_{t+1|t}^{-1} \quad t = T-1, \ldots, 1$$

with $\hat{\xi}_{T|T} = \hat{\xi}_T$ and $P_{T|T} = P_T$.

Identification

Based on Andrle (2010)

- Fundamental issue in economic and statistical modeling.
- Easily acute in state space models.
- Y set of observations
- *Model:* distribution for the variable in question.
- *S structure*, ie parameters of that distribution, ie complete probability specification of Y of the form

  $$S = F(Y, \theta),$$

  where $\theta \in \Theta \subset \mathbb{R}^n$ is the vector of parameters that belongs to parameters space Θ.
- *Observational equivalence:* two structures, $S^0 = F(Y, \theta^0)$ and $S^\star = F(Y, \theta^\star)$, having the same joint density function, ie if $F(Y, \theta^0) = F(Y, \theta^\star)$ for almost all Y.
- A structure is *identifiable* if the exists no other observationally equivalent structure, ie $\theta^0 = \theta^\star$.
- A model is identifiable if all its possible structures are identifiable. If no structure is identifiable, the model is said to be *underidentified*.
- *Local identification:* the structure is locally identified if there exists an open neighbourhood of θ^0 containing no other $\theta \in \Theta$, which produces observationally equivalent structure.

Theorem 22 (Rotemberg, 1971)**.** *Subject to some regularity conditions, θ^0 is locally identified for a given structure if and only if the Information matrix evaluated at θ^0 is not singular.*

Equation

$$T(\theta, \tau) = 0$$

defines the mapping between the *reduced form parameters* $\tau \in \mathcal{T} \subset \mathbb{R}^m$ and *structural parameters* θ.

If reduced form parameters are (locally) identified, the necessary and sufficient condition for identification is that

$$T \equiv \frac{\partial T}{\partial \theta'}$$

is of full rank.

Bayesian perspective

"Identification is a property of likelihood" and, hence, it the Bayesian approach makes no difference. Unidentification affect large sample inference and frequency properties of Bayesian inference since likelihood does not dominate prior as the sample size grows.

Data may be marginally informative even for conditionally inidentified parameter and marginal posterior and prior densities may differ.

6.2 Filtering and decompositions

Applications

The KF opens up many applications even within the context of DSGE models

- Missing observations
- Different frequencies for different variables, eg quarterly, annually, live happily together
- Different vintages of data

6.3 Bayesian methods

Features of the maximum likelihood estimation

- Takes probability distribution generated by the DSGE model very seriously
- Maximum likelihood works well if the misspecification of the DSGE model is small. However, likelihood-based analysis potentially very sensitive to misspecification. State space models are not the best setup to recover misspecification.
- As stressed above, it there is a lack of identification, the likelihood function will be flat in certain directions.
- In practice, likelihood function tends to be multi-modal, difficult to maximize, and often peaks in regions of the parameters space in which the model is hard to interpret.

$\longrightarrow$ fix subset of parameters.

Bayesian inference in a nutshell

- *Parameter is a random variable!*
- Combine prior distribution of the structural parameters with likelihood function
  - This prior density can contain information from other sources than the current data.
- Use Bayesian theorem to update the prior density by calculating conditional distribution of the parameters given the data (posterior distribution)
- Inference and decisions are then based on this posterior distribution

Nutshell of a nutshell

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Introduction

- Structural models have interest parameterization in the sense that we may have *a priori* information from
  - microeconometric studies;
  - studies from other fields
  - other countries' data
  - $\vdots$
- Bayesians love this since prior information has an essential role in Bayesian inference.
- Bayesians interprete parameters as random variables.

- Objective is to make conditional probabilistic statements regarding the parameterization of the model:
  1. structure of the model (model specification)
  2. the observed data
  3. prior distribution of the parameters.
- Likelihood is 1. plus 2.
- Combining likelihood and 3. using Bayes' rule yields an associated posterior distribution.
- Data speaks in likelihood and the researcher speaks in prior distribution.

  The incorporation of prior information is not what distinguishes classical ("frequentist") from Bayesian analysis; the distinguishing feature is the probabilistic interpretation assigned to parameters under the Bayesian perspective.
- We may interprete calibration as a Bayesian analysis involving very, very strict prior (probability mass concentrated on a single point).
- By choosing *a diffuse prior* we may let data to speak as much as possible.

Objectives in using Bayesian procedures in DSGEs

1. Implement DSGEs as a source of prior information regarding the parameterization of reduced form models (like VAR).
2. Facilitate direct estimation of the parameters of DSGEs and to implement estimated models to pursuit different tasks.
3. To facilitate model comparisons. Posterior odds analysis. Works also for false and non-nested models.

Preliminaries

Notation:

A structural model

μ vector of parameters, primary focus of analysis

$\Lambda(\mu)$ parameters of state space representation, no (extra) identification problem here. This is the model's solution.

X sample of observations

$L(X|\mu, A)$ likelihood function

$L(X|\mu)$ likelihood function if A is clear/granted

$\pi(\mu)$ prior distribution of μ

$p(X, \mu)$ joint probability of X and μ

$p(X)$ probability of X.

Classical (frequentist) view

- parameters are fixed, but unknown objects
- likelihood function is a sampling distribution for the data
- X is one of many possible realizations from $L(X|\mu)$ that could have obtained.

- Inference regarding μ center on statements regarding probabilities associated with the particular observation of X for given values of μ.

Bayesian view

- X taken as given.
- inference interested in alternative specification of μ conditional on X.
- This probabilistic interpretation of μ gives rise to incorporate prior information on μ.
- This is facilitated by the prior distribution for μ, denoted by $\pi(\mu)$.

How to incorporate prior information

- Calculate joint probability of (X, μ) with the help of conditional and unconditional distribution

$$p(X, \mu) = L(X|\mu)\pi(\mu)$$

  or reversing the role

$$p(X, \mu) = p(\mu|X)p(X)$$

- In the first equation the likelihood function $L(X|\mu)$ has the role of conditional probability and conditioning is made wrt μ.
- In the second equation conditioning is wrt X.
- Get rid of the joint probability by equating those two:

$$p(\mu|X)p(X) = L(X|\mu)\pi(\mu).$$

- Solving for $p(\mu|X)$ gives Bayes' rule:

$$p(\mu|X) = \frac{L(X|\mu)\pi(\mu)}{p(X)} \propto L(X|\mu)\pi(\mu).$$

  - The $\propto$ is because $p(X)$ is a constant from the point of view of the distribution of μ.
  - $p(\mu|X)$ is now the posterior distribution.
  - *Conditional on X and the prior $\pi(\mu)$ it assigns probabilities to alternative values of μ.*

- Bayesian analysis typically involves calcualting the conditional expected value of a function of the parameters $g(\mu)$:

$$\mathrm{E}[g(\mu)] = \frac{\int g(\mu)P(\mu|X)d\mu}{\int P(\mu|X)},$$

  where the denominator $\int P(\mu|X)$ exists to handle the missing $p(X)$ in the Bayes' rule.

- Examples of $g(\mu)$
  - Identitity function would result posterior mean.
  - $g(\mu) = 1$ for $\mu^j \in [\underline{\mu}^j, \bar{\mu}^j)$, and 0 other wise: If repeated to each element in μ would enable to construction of *marginal predictive density functions* (p.d.f.:s) for each μ_i. Here Marginal means that the p.d.f. is *un*conditional on the other elements of μ.
  - marginal p.d.f. of spectra
  - marginal p.d.f. of impulse response functions
  - marginal p.d.f. of predictive densities
  - marginal p.d.f. of unobserved shock processes
  - marginal p.d.f. of forecasts!

- Hence, $\mathrm{E}[g(\mu)]$ is the weighted average of $g(\mu)$ where weights are given by posterior distribution $P(\mu|X)$, ie by data (likelihood function) and the prior.
- Rarely $\mathrm{E}[g(\mu)]$ can be calculated analytically $\longrightarrow$ *numerical integration*.

Numerical integration

- Suppose we may draw directly from $P(\mu|X)$
- we, naturally, know $g(\mu)$
- Let's do Monte Carlo integration:
  `for i=1:N` % where N is a large number like 10000
  1. draw μ_i (one realization) from $P(\mu|X)$
  2. compute $g(\mu_i)$
  3. store it to a vector

  `end`
  compute the average.

- Typically you may do this without loops by using vector/matrix operations, eg `gN = mean(g(randmu(N)))`
- By law of large numbers
$$\bar{g}_N = \frac{1}{N}\sum_i^N g(\mu_i) \xrightarrow{p} \mathrm{E}(g(\mu)],$$
where $\xrightarrow{p}$ means convergence in probability.
- $\mathrm{Std}(\bar{g}_N)$ is given by
$$\mathrm{std}(\bar{g}_N) = \frac{\sigma(g(\mu))}{\sqrt{N}}.$$
and its sample (simulated) counterpart
$$\bar{\sigma}_N[g(\mu)] = \left[\frac{1}{N}\sum_{i=1}^N g(\mu_i)^2 - \bar{g}_N^2\right]^2.$$
- Marginal p.d.f. of $g(\mu)$ may obtained from the draws of $g(\mu)$ by kernel methods.

6.4 Simulating posterior

Implementing DSGEs

- Typically posterior distribution $p(\mu|X)$ is not directly (analytically) available.
  - likelihood is available for $\Lambda(\mu)$, whereas
  - priors are specified in terms μ.
- *We can calculate $p(\mu^\star|X)$ for given value of $\mu^\star$ with the help of the likelihood function $L(X|\mu^\star, A)$ (ie with Kalman filter).*
- The last problem is to define from where to draw μ_i such that the procedure will be
  - accurate
  - efficient
- DeJong and Dave (2007) studies
  - Importance sampling
  - MCMC: Gibbs
  - MCMC: Metropolis-Hastings

Markov chain simulations

A Markov chain is a sequence of random variables, such that the probability distribution of any one, given all preceding realizations, depend on the immediately preceding realization. Hence, $\{x_i\}$ has the property

$$\Pr(x_{i+1}|x_i, x_{i-1}, x_{i-2}, \dots) = \Pr(x_{i+1}|x_i)$$

Here i refers to Monte Carlo replication.

- x_i can take one of n possible values, which collectively defines the state space over x
- Movements of x_i over i is characterized via P, an $n \times n$ transition matrix. Hence, a Markov chain can be fully described by its initial state and a rule describing how the chain moves from its state in i to a state in $i+1$.
- p_{qr} is the $(q,r)^{\text{th}}$ element of P:
  - it is the probability that x_i will transition to state r in replication $i+1$ given its current state q.
  - q^{th} row of P is the conditional probability distribution for x_{i+1} given the current state q in replication i

The posterior distribution is not standard. The idea of Markov Chain Monte Carlo is to construct a stochastic process (Markov Chain), such that:

- it has stationary distribution
- it converges to that stationary distribution
- the stationary distribution is the target distribution, ie the posterior distribution $p(\mu|X)$. (ergodic distribution)

$\longrightarrow$ construct a Markov chain in μ, ie the transition matrix P, such that the process converges to the posterior distribution. *The 'process' here is the sequence of draws, ie the simulations.*

Metropolis algorithm creates a sequence of random points $(\mu_1, \mu_2, \dots)$ whose distribution converges to $p(\mu|X)$.

Proposal distribution

Let $\iota(\mu|\mu_{i-1}, \theta)$ be a *stand-in density* (*proposal distribution* or *jumping distribution*), where θ represents the parameterization of this density.

- Ideally $\iota(\mu|\mu_{i-1}, \theta)$ should have fat tails relative to posterior: algorithm should visit all parts of the posterior distribution.
- center μ_{i-1} and "rotation" θ should different from posterior.

Metropolis algorithm

Let $\mu_i^\star$ denote a draw from $\iota(\mu|\mu_{i-1}, \theta)$. It is a *candidate* to become next *successful* μ_i.

Calculate the ratio of the densities

$$r = \frac{p(\mu_i^\star|X)}{p(\mu_{i-1}|X)}.$$

Set

$$\mu_i = \begin{cases} \mu_i^\star & \text{with probability } \min(r, 1) \\ \mu_{i-1} & \text{otherwise.} \end{cases}$$

- $p(\mu_i^\star|X) \geq p(\mu_{i-1}|X)$
  $\longrightarrow$ always include candidate as a new point in the sequence.
- $p(\mu_i^\star|X) < p(\mu_{i-1}|X)$
  $\longrightarrow$ $\mu_i^\star$ not always included; the lower $p(\mu_i^\star|X)$, the lower the chance it is included.

Given the current value μ_{i-1}, the Markov chain transition distribution p_{μ_{i-1},μ_i} is a mixture of the jumping distribution $\iota(\mu|\mu_{i-1},\theta)$ and a point mass at $\mu_i = \mu_{i-1}$.

Note! If $\mu_i = \mu_{i-1}$, ie the jump is not accepted, this counts as an iteration in the algorithm.

The algorithm can be viewed as a stochastic version of a stepwise mode-finding algorithm: always stepping to increase the density but only sometimes stepping to decrease. This is the idea in simulated annealing algorithm.

Why does the Metropolis algorithm work?

The proof of the convergence to the target distribution has two steps:

1. Show that the simulated sequence $(\mu_1, \mu_2, \dots)$ is a Markov chain with a unique stationary distribution: holds if the Markov chain is irreducible, aperiodic, and not transient
   $\longrightarrow$ hold if the jumping distribution is chosen such that it covers all possible states with positive probability $\rightarrow$ has fat tails.
2. Show that the stationary distribution is the target distribution.

To proof the second part:

- Suppose μ_{i-1} a draw from the target distribution $p(\mu|X)$.
- Consider two draws, μ_a and μ_b, from the target distribution such that $p(\mu_b|X) \geq p(\mu_a|X)$.
- The unconditional transitional probability from μ_a to μ_b is

  $$p_{\mu_a\mu_b} = p(\mu_a|X)\iota(\mu_b|\mu_a,\theta) \times 1,$$

  where the acceptance probability is 1 (because $p(\mu_b|X) \geq p(\mu_a|X)$).
- The unconditional transitional probability from μ_b to μ_a

  $$\begin{aligned} p_{\mu_b\mu_a} &= p(\mu_b|X)\iota(\mu_a|\mu_b,\theta)\frac{p(\mu_a|X)}{p(\mu_b|X)} \\ &= p(\mu_a|X)\iota(\mu_a|\mu_b,\theta) \end{aligned}$$

- They are *the same*, because we have assumed that $\iota(\cdot|\cdot,\theta)$ is symmetric.
- Since the joint distribution is symmetric, μ_i and μ_{i-1} have the same marginal distributions, and so
- $p(\mu|X)$ is the stationary distribution of the Markov chain.

Metropolis-Hasting algorithm

Let $\mu_i^\star$ denote a draw from $\iota(\mu|\mu_{i-1},\theta)$. It is a *candidate* to become next *successful* μ_i.
It has the following probability to become successful:

$$q(\mu_i^\star|\mu_{i-1}) = \min\left[1, \frac{p(\mu_i^\star|X)/\iota(\mu_i^\star|\mu_{i-1},\theta)}{p(\mu_{i-1}|X)/\iota(\mu_{i-1}|\mu_{i-1},\theta)}\right]$$

This probability can be simulated as follows

- draw $\mho$ from uniform distribution over $[0,1]$.
- `if` $q(\mu_i^\star|\mu_{i-1}) > \mho$ `then` $\mu_i = \mu_i^\star$ `else` redraw new candidate $\mu_i^\star$ from $\iota(\mu|\mu_{i-1},\theta)$.

Note that if $\iota(\mu|\theta) = p(\mu|X)$ $q(\mu_i^\star|\mu_{i-1}) = 1$.
Finally, $E[g(\mu)]$ can be calculated usual way using sequence of *accepted* draws.
Two variants:

1. independence chain: $\iota(\mu|\mu_{i-1},\theta) = \iota(\mu|\theta)$.

2. random walk MH: candidate draws
$$\mu_i^\star = \mu_{i-1} + \varepsilon_i$$
then stand-in density evolves
$$\iota(\mu|\mu_{i-1}, \theta) = \iota(\mu - \mu_{i-1}|\theta),$$
so that the center (mean) of $\iota(\mu|\mu_{i-1}, \theta)$ evolves over Monte Carlo replications following a random walk.

Choosing the stand-in distribution:

- The "quality" of stand-in distribution helps in convergence. It should be as close as possible to posterior distribution.
- *Natural candidate:* ML estimates of the parameters μ follow *asymptotically* multivariate Normal distribtion $N(\hat{\mu}, \Sigma)$.
- Since we want to be sure to have fatter tails, lets scale the covariance matrix:
$$N(\hat{\mu}, c^2\Sigma)$$
This would be our first draw from stand-in density, ie $\iota(\mu_0^\star|\theta)$.
- Random walk Metropolis-Hastings would involve
$$\iota(\mu - \mu_{i-1}|\theta) = N(\mu_{i-1}, c^2\Sigma).$$

$$\text{Acceptance rate} = \frac{\text{accepted draws}}{\text{all draws}}$$

Optimal acceptance rate is 0.44 for a model with one parameter and 0.23 if there more than *five* parameters.

This guides us in choosing c (the scale factor of the asymptotic covariance matrix):

1. start with $c = 2.4/\sqrt{\text{number of parameters}}$.
2. Increase (decrease) c if the acceptance rate is too high (low).

Markov chain diagnostics

In the ML estimation the problems related to likelihood (eg identification, or the model specification in general) show typically up as numerical problem in maximizing the likelihood.

In Bayesian estimation they show up in the MCMC. Fundamental questions is the convergence of the chain(s):

- Is the target distribution achieved?
- What is the impact of the starting values of the chain?
- Are blocks of draws correlated?

Partial solutions to possible problem

- Remember that this is not standard Monte Carlo where 10 000 replication is a big number. Here it is million (Markov chain)
- Throw away early iterations of each MCMC chain. Even 50 %
- Asses the convergence
  - Plot the raw draws, ie parameter value against draw: if they are trending, there is a trouble.
  - Plot multiple chains to the same graph: are they in the same region.

- Plot:
$$\frac{1}{n_s}\sum_{s=1}^{n_s} g(\mu^{(s)})$$
as a function of n_s.
- start Markov chain at extreme values of μ and check whether different runs of the chain settle to the same distribution

- Simulate multiple chains: Look, among other things, whether the moments of a single parameter $\mu(j)$ are similar across the chains
- Compute test statistics that measure *variation within a chain* and *between the chains*.
- These test statistic builds on the classical test set-ups that try to detect structural breaks.

Testing the structural change, ie convergence.

- Cusum-type of tests of structural change (see Yu and Mykland (1998) or Yada manual)
- Partial means test by Geweke (2005) splits the chain into groups:
  - N is number of draws, $N_p = N/(2p)$, where p is positive integer. Define p separated partial means
$$\hat{S}_{j,p}^{(N)} = \frac{1}{N_p}\sum_{m=1}^{N_p} S(\mu^{(m+N_p(2j-1))}), \quad j = 1,\ldots,p$$
where S is some summary statistic of the parameters μ (eg original parameters). Let $\hat{\tau}_{j,p}$ be the Newey-West (1987) numerical standard error for $j = 1,\ldots,p$. Define the $(p-1)$ vector $\hat{S}_p^{(N)}$ with typical element $\hat{S}_{j+1,p}^{(N)} - \hat{S}_{j,p}^{(N)}$ and the $(p-1)\times(p-1)$ tridiagonal matrix $\hat{V}_p^{(N)}$ where
$$\hat{V}_{j,j}^{(N)} = \hat{\tau}_{j,p}^2 + \hat{\tau}_{j+1,p}^2, \quad j = 1,\ldots,p-1$$
and
$$\hat{V}_{j,j+1}^{(N)} = \hat{V}_{j+1,j}^{(N)} = \hat{\tau}_{j+1,p}^2, \quad j = 1,\ldots,p-1.$$
The statistic
$$G_p^{(N)} = \left(\hat{S}_p^{(N)}\right)' [\hat{V}_p^{(N)}]^{-1}\hat{S}_p^{(N)} \xrightarrow{d} \chi^2(p-1),$$
as $N \to \infty$ under the null that the MCMC chain has converged with the separated partial means being equal.

- Multiple chain diagnostics
  - $i = 1,\ldots,N$ chain length; $j = 1,\ldots,M$ number of chains.
  - *Between chain variance* is
$$B = \frac{N}{M-1}\sum_{j=1}^{M}(\bar{S}_j - \bar{S})^2,$$
where
$$\bar{S}_j = \frac{1}{N}\sum_{i=1}^{N} S_{ij} \quad \bar{S} = \frac{1}{M}\sum_{j=1}^{M}\bar{S}_j.$$
and *within chain variance*
$$W = \frac{1}{M(N-1)}\sum_{j=1}^{M}\sum_{i=1}^{N}(S_{ij} - \bar{S}_j)^2.$$
  - Define variance of S
$$\hat{\Sigma}_S = \frac{N-1}{N}W + \frac{1}{N}B,$$
and
$$\hat{V} = \hat{\Sigma}_S + \frac{B}{MN}.$$

– The *potential scale reduction factor* $\hat{R}$ should decline to 1 if simulation converges

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}} = \sqrt{\frac{N-1}{N} + \frac{(M+1)B}{MNW}}.$$

Choosing prior distribution

Think carefully:

- What is the domain?
- Is it bounded?
- Shape of distribution: symmetric, skewed, which side?

How to choose prior:

- The choice of a prior distribution for a deep parameter often follows directly from assumptions made in the DSGE model.
- Results from other studies (Microstudies or DSGE estimates) are already available.
  - Pre-sample.
  - Other countries and data sets
  - *NOT* the data you are currently using the estimation.
- Should not be too restrictive — allow for a wide support.

Most common distributions employed are:

- Beta distribution; range is $\in (p_1, p_2)$; autoregressive parameters $\in (0, 1)$
- Gamma distribution; range is $\in (p_1, \infty)$;
- Normal distribution
- Uniform distribution; range is $\in (p_1, p_2)$; diffuse priors
- Inverted-gamma distribution; range is $\mathbb{R}^+$; variances

Yada manual has nice plots of the distributions.

Look at the model moments implied by priors!

Sensitivity analysis: how robust are the results to choice of prior?

6.5 Model comparison

Model comparison

Posterior distribution can be used to assess conditional probabilities associated with alternative model specifications. This lies at the heart of the Bayesian approach to model comparison.

Relative conditional probabilities are calculated using *odds ratio*.

- Two models: A and B with corresponding parameter vectors that may differ μ_A and μ_B.
- Study model A (and change the symbols to study the model B).

- Posterior (conditional on the model A)

$$p(\mu_A|X,A) = \frac{L(X|\mu_A,A)\pi(\mu_A|A)}{p(X|A)}$$

Integrate both sides by μ_A (ie calculate the marginal probabilities; ie average over the all possible parameter values) to have

$$p(X|A) = \int L(X|\mu_A,A)\pi(\mu_A|A)d\mu_A.$$

This is the *marginal likelihood* of model A.

- Use Bayes' rule to calculate conditional probability associated with model A:

$$p(A|X) = \frac{p(X|A)\pi(A)}{p(X)}$$

(use the similar logic as in the start of the section, ie define joint probability of data and model (X,A).

- Substitute the marginal likelihood into the above equation

$$p(A|X) = \frac{\left[\int L(X|\mu_A,A)\pi(\mu_A|A)d\mu_A\right]\pi(A)}{p(X)}$$

Look at the ratio of the above condtional density to that of the model B. This is called *posterior odds ratio*:

$$PO_{A,B} = \underbrace{\frac{\left[\int L(X|\mu_A,A)\pi(\mu_A|A)d\mu_A\right]}{\left[\int L(X|\mu_B,B)\pi(\mu_B|B)d\mu_B\right]}}_{\text{Bayes factor}} \underbrace{\frac{\pi(A)}{\pi(B)}}_{\text{prior odds ratio}}$$

Beauty of this approach

- all models are treated symmetrically; no "null hypothesis"
- all models can be false
- works equally well for non-nested models

Hence, favour the model whose *marginal likelihood* is larger:

- Odds $\in (1-3)$ "very slight evidence" in favour of A
- Odds $\in (3-10)$ "slight evidence" in favour of A
- Odds $\in (10-100)$ "strong to very strong evidence" in favour of A
- Odds > 100 "decisive evidence" in favour of A

Implementation is tough, because we need to integrate the parameters out.

Easy solution would be to assume functional form to marginal likelihood. *Laplace approximation* utilizes Gaussian large sample properties

$$\hat{p}(X|A) = (2\pi)^{\frac{T}{2}}|\Sigma_{\mu_A^m}|^{\frac{1}{2}}p(\mu_A^m|X,A)p(\mu_A^m|A),$$

where μ_A^m is the *posterior mode*, ie the initial (ML estimated) value of the parameters in stand-in distribution.

Harmonic mean estimator is based on the following idea:

$$\mathrm{E}\left[\frac{f(\mu_A)}{p(\mu_A|A)p(X|\mu_A,A)}|\mu_A,A\right] = \frac{\int f(\mu_A)d\mu_A}{\int p(\mu_A|A)p(X|\mu_A,A)d\mu_A} = p(X|A)^{-1},$$

where f is a probability density function. This suggest the following estimator of the marginal density

$$\hat{p}(X|A) = \left[\frac{1}{B} \sum_{b=1}^{B} \frac{f(\mu_A^b)}{p(\mu_A^b|A)p(X|\mu_A^b, A)} \right]^{-1},$$

where μ_A^b is a draw of μ_A in Metropolis-Hastings iteration. Typically f is replaced by a truncated normal, this estimator is then called by *modified harmonic mean estimator*.

References

References

Anderson, Gary, and George Moore (1985) 'A linear algebraic procedure for solving linear perfect foresight models.' *Economics Letters* 17(3), 247–252

Andrle, Michal (2010) 'A note on identification patterns in DSGE models.' Working Paper Series 1235, European Central Bank, August

Blanchard, Olivier J., and C. Kahn (1980) 'Solution of linear difference models under rational expectations.' *Econometrica* 38, 1305–1311

DeJong, David N., and Chetan Dave (2007) *Structural Macroeconometrics* (Princeton and Oxford: Princeton University Press)

den Haan, Wouter J., and Andrew Levin (1996) 'A practioner's guide to robust covariance matrix estimation.' Discussion Paper 96-17, University of California, San Diego. to be appear in Handbook of Statistics 15 (Chapter 12, 291–341)

Hansen, Lars Peter (1982) 'Large sample properties of generalized method of moments estimator.' *Econometrica* 50(4), 1029–1054

Hansen, Lars Peter, John Heaton, and Amir Yaron (1996) 'Finite-sample properties of some alternative GMM estimators.' *Journal of Business and Economic Statistics* 14(3), 262–280

Klein, Paul (2000) 'Using the generalized schur form to solve a multivariate linear rational expectations model.' *Journal of Economic Dynamics and Control* 24(10), 1405–1423

Kocherlakota, Narayana R. (1990) 'On tests of representative consumer asset pricing models.' *Journal of Monetary Economics* 26, 385–304

Kydland, Finn E., and Edward C. Prescott (1982) 'Time to build and aggregate fluctuations.' *Econometrica* 50(6), 1345–1370

Newey, Whitney K, and Kenneth West (1994) 'Automatic lag selection in covariance matrix estimation.' *Review of Economic Studies* 61, 631–653

Sargent, Thomas J. (1987) *Macroeconomic Theory*, second edition ed. (Emerald)

Sims, Christopher A. (2011) 'Nobel lecture.' Technical Report

Tauchen, George (1986) 'Statistical properties of generalized method of moments estimators of structural parameters obtained from financial market data.' *Journal of Business Econom. Statistics* 4, 397–425

Zagaglia, Paolo (2005) 'Solving rational-expectations models through the Anderson-Moore algorithm: An introduction to the Matlab implementation.' *Computational Economics* 26(1), 91–106