Econ 232B Anton Cheremukhin

4 December 2007

Part 1 (Jesus Fernandez-Villaverde)

Introduction

This set of lectures discusses methods of estimating structural economic models with applications to macroeconomic data. Let's start with a roadmap. In order to estimate a structural economic model given some data a researcher should take four general steps. First the model should be clearly specified including assumptions about the distributions of all the exogenous shocks hitting the economy. Usually a macroeconomic model consists of the choice functions of the agents which are optimized given some constraints, the aggregate resource constraints on the economy, as well as assumptions about expectation formation and stochastic processes driving the shocks.

The second step includes finding the solution of the model, which amounts to describing the state vector and writing policy functions for all the endogenous variables. This solution can usually be expressed in terms of a state-space representation, which consists of a transition equation and a measurement equation. The transition equation describes the evolution of the state vector subject to random shocks, while the measurement equation determines how the observed data is connected to the state vector.

After the model is specified and solved, and a state-space representation is obtained, filtering theory provides tools to compute the likelihood function of the data given a parameter vector. If the model is linear and the shocks are gaussian, then the Kalman filter provides an exact analytical solution for the likelihood and the evolution of the state vector can be recovered using standard techniques. If the model is nonlinear and/or non gaussian then more complicated numerical procedures such as sequential Monte-Carlo methods and the particle filter are required to compute the likelihood.

The fourth step is to compute the likelihood function of the observations for different values of the parameter vector and use it as a measure of goodness of fit in order to find some optimal parameter values or optimal distribution of parameter values. Finding the mode of the likelihood is a consistent way of getting a classical point estimate of the parameters, while a Bayesian econometician would obtain a posterior distribution of parameters using some economically sound prior combined with the likelihood function.

Solving DSGE Models

Let's look at an example - a standard real business cycle model:

$$\begin{aligned} &\max_{c_t,l_t,k_{t+1}} E\sum_{t=0}^{\infty}\beta^t\left(\ln c_t+\psi\ln\left(1-l_t\right)\right)\\ \text{s.t.}\qquad &c_t+k_{t+1}=k_t^{\alpha}\left(e^{z_t}l_t\right)^{1-\alpha}+\left(1-\delta\right)k_t\\ &z_t=\rho z_{t-1}+\varepsilon_t,\qquad \varepsilon_t\sim N\left(0,\sigma\right)\end{aligned}$$

There are several ways to solve this model. In the special case $\delta = 1$ the model has an analytical solution. If that's not the case only a numerical solution can be obtained. One way to find a global numerical solution is via value function iteration. This method is good for simple models, but it's very slow once the models become more complicated.

A faster way to obtain the global solution would be to approximate the policy function using some set of basis functions: $f(s) = \sum_{i=1}^{n} \theta_i \psi_i(s)$.

The basis functions $\psi_i()$ people often use are Chebyshev polynomials or some local functions (functions with a compact support). The latter method is called the finite elements method. Both value function iteration methods and projection methods suffer from the curse of dimensionality, since the time needed to find the solution and the memory requirements increase exponentially with the number of states.

The only way to avoid the curse of dimensionality is to sacrifice the global properties of the solution and instead to take a local approximation of the solution around some point of interest. Most economic models have at least one equilibrium and if the shocks are not very big, taking a local approximation of the model in the vicinity of the equilibrium point is usually not a very bad idea.

Perturbation methods use a local approximation of the policy functions around the steady state of the model, which makes the model much easier to solve. The disadvantage of this approach is that it's not appliccable to bigger deviations from the steady state as well as to the cases of multiple steady states.

Let's use the RBC model above to illustrate the application of perturbation methods. First the equilibrium (first order) conditions are written down.

FOC_{c_t} : $\frac{1}{c_t} = \lambda_t$

FOC_{l_t} : $\frac{\psi}{1-l_t} = \lambda_t k_t^{\alpha} \left(e^{z_t} l_t\right)^{1-\alpha} \frac{1-\alpha}{l_t}$

$\text{FOC}_{k_{t+1}}$: $\lambda_t = E_t \beta \lambda_{t+1} \left(\frac{\alpha}{k_{t+1}} k_{t+1}^{\alpha} \left(e^{z_{t+1}} l_{t+1}\right)^{1-\alpha} + 1 - \delta \right)$

Simplifying the system we get the intertemporal Euler equation:

$$\frac{1}{c_t} = \beta E_t \left[\frac{1}{c_{t+1}} \left(\alpha \left(e^{z_{t+1}} \frac{l_{t+1}}{k_{t+1}} \right)^{1-\alpha} + 1 - \delta \right) \right]$$

and the consumption-leisure trade-off:

$$\psi \frac{c_t}{1-l_t} = (1-\alpha) \left(\frac{k_t}{l_t} \right)^{\alpha} \left(e^{z_t}\right)^{1-\alpha}$$

Adding the resource constraint and the stochastic process for technology we get a system of four equations:

$$\frac{1}{c_t} = \beta E_t \left[\frac{1}{c_{t+1}} \left(\alpha \left(e^{z_{t+1}} \frac{l_{t+1}}{k_{t+1}} \right)^{1-\alpha} + 1 - \delta \right) \right] \tag{1}$$

$$\psi \frac{c_t}{1-l_t} = (1-\alpha) \left(\frac{k_t}{l_t} \right)^{\alpha} \left(e^{z_t}\right)^{1-\alpha} \tag{2}$$

$$c_t + k_{t+1} = k_t^{\alpha} \left(e^{z_t} l_t\right)^{1-\alpha} + (1-\delta) k_t \tag{3}$$

$$z_t = \rho z_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma) \tag{4}$$

The second step is to find a deterministic steady state, which is achieved by setting variances of all random variables to zero: $\sigma = 0$. Here σ is called the perturbation parameter, because we linearize the model in the vicinity of $\sigma = 0$. Then the equations for the steady-state can be rewritten as:

$$\begin{array}{ll} z = \rho z = 0 & \frac{1}{c} = \beta \frac{1}{c}\left(\alpha\left(\frac{l}{k}\right)^{1-\alpha} + 1 - \delta\right) \\ \psi \frac{c}{1-l} = (1-\alpha)\left(\frac{k}{l}\right)^{\alpha} & c = k^{\alpha} l^{1-\alpha} - \delta k \end{array}$$

These imply the steady-state values of the form:

$$\begin{array}{ll} \left[\frac{\frac{1}{\beta}-1+\delta}{\alpha}\right]^{\frac{1}{1-\alpha}} = \frac{l}{k} = \varphi & \frac{c}{1-l} = \frac{1-\alpha}{\psi}\varphi^{-\alpha} = \mu \\ c = \left(\frac{\frac{1}{\beta}-1+\delta}{\alpha} - \delta\right) k = \left(\varphi^{1-\alpha} - \delta\right) k = \Omega k & \end{array}$$

Hence, $c = \mu(1-l) = \frac{\Omega}{\varphi} l$ and

$$k = \frac{\mu}{\Omega + \varphi\mu}, \qquad c = \Omega k, \qquad l = \varphi k, \qquad y = k^{\alpha} l^{1-\alpha}.$$

One can do linearization or loglinearization around the steady-state. Log-linearization means that we take log-deviations from the steady-state of the form: $x_t = x_{ss} e^{\hat{x}_t}$. If we now replace all the variables in the Euler equations using this formula for deviations from the steady state, it follows, that (omitting the hats):

$$\begin{array}{c} \frac{1}{ce^{c_t}} = \beta E_t\left[\frac{1}{ce^{c_{t+1}}}\left(\alpha\left(e^{z_{t+1}}\frac{le^{l_{t+1}}}{ke^{k_{t+1}}}\right)^{1-\alpha} + 1 - \delta\right)\right] \\ \psi \frac{ce^{c_t}}{1-le^{l_t}} = (1-\alpha)\left(\frac{ke^{k_t}}{le^{l_t}}\right)^{\alpha}\left(e^{z_t}\right)^{1-\alpha} \\ ce^{c_t} + ke^{k_{t+1}} = e^{\alpha k_t} k^{\alpha}\left(e^{z_t} l e^{l_t}\right)^{1-\alpha} + (1-\delta) ke^{k_t} \end{array}$$

This can be simplified and after taking logs is equivalent to:

$$\begin{array}{c} E_t\left(c_{t+1} - c_t\right) = (1-\alpha)\left(1 - \beta\left(1+\delta\right)\right) E_t\left(z_{t+1} + l_{t+1} - k_{t+1}\right) \\ c_t - (1-\alpha) z_t - \alpha\left(k_t - l_t\right) + \frac{\varphi\mu}{\Omega} l_t = 0 \\ \frac{\Omega}{\Omega+\delta} c_t + \frac{1}{\Omega+\delta} k_{t+1} - \frac{1-\delta}{\Omega+\delta} k_t = \alpha k_t + z_t + (1-\alpha) l_t \end{array}$$

After some algebra one can express the model in the following way:

$$\begin{array}{c} Ak_{t+1} + Bk_t + Cl_t + Dz_t = 0 \\ E_{t+1}\left(Gk_{t+1} + Hk_t + Jl_{t+1} + Kl_t + Lz_{t+1} + Mz_t\right) = 0 \\ Ez_{t+1} = \rho z_t = Nz_t \end{array}$$

Perturbation methods are based on the method of undetermined coefficients. We guess the policy functions of the form:

$$\begin{array}{c} k_{t+1} = Pk_t + Qz_t, \\ l_t = Rk_t + Sz_t, \end{array}$$

and plug it into the equations of the model. Since these policy rules need to hold for any k_t and z_t, all the coefficients in the resulting equations have to be zero.

$$A\left(Pk_t + Qz_t\right) + Bk_t + C\left(Rk_t + Sz_t\right) + Dz_t = 0$$
$$G\left(Pk_t + Qz_t\right) + Hk_t + J\left(R\left(Pk_t + Qz_t\right) + SNz_t\right) + K\left(Rk_t + Sz_t\right) + LNz_t + Mz_t = 0$$

This implies:

$$AP + B + CR = 0 \qquad AQ + CS + D = 0$$
$$GP + H + JRP + KR = 0 \qquad GQ + JRQ + JSN + KS + LN + M = 0$$

This system of four equations in four unknowns (P,Q,R,S) simplifies to a quadratic equation. Usually it has two solutions, one with an eigenvalue smaller than one corresponding to a stable path converging back to the steady state, and another explosive solution diverging from the steady state, which violates the transversality condition.

The stable solution is equivalent to a system of the form:

$$\begin{bmatrix} k_{t+1} \\ z_{t+1} \\ l_t \end{bmatrix} = \begin{bmatrix} P & Q & 0 \\ 0 & N & 0 \\ R & S & 0 \end{bmatrix} \begin{bmatrix} k_t \\ z_t \\ l_t \end{bmatrix} + \begin{bmatrix} 0 \\ \varepsilon_{t+1} \\ 0 \end{bmatrix}$$

This is a linear (first-order) approximation of the model. When the changes around the steady state are relatively big, or when one attempts to measure welfare implications of different policies, the first-order approximation is not enough. However, in most cases the second order is enough for all interesting questions. Though it might seem, that taking the second order is much more complicated, and will lead to higher order equations, in fact to compute all the higher order approximations beyond the first one the researcher only needs to recursively solve linear systems of equations.

Computing the Likelihood

In the previous section it was demonstrated how one can solve a simple real business cycle model using perturbation methods. The result of that process was a first or second order Markov structure, which can be more generally expressed in the following form:

$$S_t = f\left(S_{t-1}, W_t; \gamma\right),$$

where S_t is the state vector, W_t is the vector of shocks hitting the economy, and γ is the vector of all the structural parameters (e.g. describing preferences, technology and beliefs).

This equation is called the transition equation because it describes how the system goes from one state to another being hit by some exogenous shock. This equation expresses the researcher's assumptions of how the economy moves over time. To compare it to the data one needs to know how the state of the economy affects some variables, that are directly observable. The corresponding equation is called the measurement equation:

$$Y_t = g\left(S_t, V_t; \gamma\right),$$

where Y_t stands for the observed variables, V_t is the measurement error. One can interpret measurement errors as either shocks that hit observables but not the states, or sometimes they can be shocks with clear economic intuition. The more freedom is given to these shocks the better the fit of the model is, but the less useful the estimates of parameters are.

Once the model is expressed in this state-space representation, which includes the transition equation and the measurement equation, standard methods coming from the filtering theory are used to make inference about the states of the economy given the observations. The state-space representation is very flexible, and almost any system can be expressed in this form by either a clever transformation of variables, or by adding lags of the variables as separate states, used to track the history. In the case of the RBC model, when the economist can observe output and labor input, the measurement equation takes the following form:

$$\begin{bmatrix} y_t \\ l_t \end{bmatrix} = \begin{bmatrix} \alpha & \alpha & 1-\alpha \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} k_t \\ z_t \\ l_{t-1} \end{bmatrix} + \begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix}$$

Filtering theory provides the researcher with tools to do filtering (recovering the state of the economy today given all the information up to date), smoothing (recovering the states of the economy in previous periods, given all the information up to date) and forecasting (making projections into the future).

All of the techniques are in fact based on two fundamental results. The first one is the Chapman-Kolmogorov theorem, that is used for predicting the state one step into the future:

$$p\left(S_t|y^{t-1};\gamma\right) = \int p\left(S_t|S_{t-1};\gamma\right) p\left(S_{t-1}|y^{t-1};\gamma\right) dS_{t-1}.$$

The second result is used for updating the conditional distribution of the state vector, given a new observation. It is the Bayes theorem:

$$p\left(S_t|y^{t};\gamma\right) = \frac{p(y_t|S_t;\gamma)p\left(S_t|y^{t-1};\gamma\right)}{\int p(y_t|S_t;\gamma)p(S_t|y^{t-1};\gamma)dS_t}.$$

Using the state-space representations we can invert the transition equation $S_t = f\left(S_{t-1}, W_t;\gamma\right)$ to obtain the conditional distribution $p\left(S_t|S_{t-1};\gamma\right)$, while by using the measurement equation $Y_t = g\left(S_t, V_t;\gamma\right)$ we can sample from $p\left(y_t|S_t;\gamma\right)$.

Therefore, the likelihood can be factorized and computed recursively as:

$$p\left(y^T;\gamma\right) = \prod_{i=1}^{T} p\left(y_t|y^{t-1};\gamma\right) = \int p\left(y_1|S_0;\gamma\right) dS_0 \prod_{i=2}^{T} \int p\left(y_t|S_t;\gamma\right) p\left(S_t|y^{t-1};\gamma\right) dS_t$$

This can be done by:

1. drawing from the distribution of the initial state,
2. computing the probability of the first observation conditional on the initial state
3. sampling from the measurement equation given the initial state and the parameter vector,
4. transiting to a new state using the transition equation and the Chapman-Kolmogorov theorem

5. updating the probabilities of the state given a new observation using the Bayes theorem

6. weighting the results with relative probabilities to obtain the likelihood of the next observation given the new state,

7. going to step 3 and applying steps 3-7 recursively untill the sample ends.

This process computes the likelihood of the observations given the initial state and the parameters of the model, and at the end gives the distribution of the final state S_T conditional on all the data available. States in previous periods were all computed conditional on the data up to that period of time, while it may be interesting to condition on all the available information. This is called smoothing. To find the whole sequence of states conditional on all the information available to the researcher, backward recursion is applied:

$$p\left(S_t|y^T;\gamma\right) = p\left(S_t|y^t;\gamma\right) \int p\left(S_{t+1}|y^T;\gamma\right) p\left(S_{t+1}|S_t;\gamma\right) dS_{t+1}.$$

This allows to correct beliefs about the states, using later observations.

The described general strategy looks simple, but for the fact that it's impossible to compute the above integrals analytically except certain particular cases. One of such cases is when the system is linear (e.g. first order approximation) and the shocks are gaussian. In this case all the conditional distributions in all future periods are also gaussian, and one needs to keep track only of the first two moments of each distribution. The case of a linear model with gaussian shocks is called the Kalman filter.

Kalman Filter

The linear system can be expressed as:

$$\begin{aligned} s_t &= Fs_{t-1} + Gw_t, \qquad w_t \sim N\left(0, Q\right) \\ y_t &= H's_t + v_t, \qquad v_t \sim N\left(0, R\right) \end{aligned}$$

The goal is to find the best linear predictors of the following variables:

$$\begin{gathered} s_{t|t-1} = E\left(s_t|y^{t-1}\right) \qquad y_{t|t-1} = E\left(y_t|y^{t-1}\right) \qquad s_{t|t} = E\left(s_t|y^t\right) \\ \Sigma_{t|t-1} = E\left(\left(s_t - s_{t|t-1}\right)^2|y^{t-1}\right) \qquad \Omega_{t|t-1} = E\left(\left(y_t - y_{t|t-1}\right)^2|y^{t-1}\right) \qquad \Sigma_{t|t} = E\left(\left(s_t - s_{t|t}\right)^2|y^t\right) \end{gathered}$$

In this case the Chapman-Kolmogorov updating rule is equivalent to:

$$\begin{gathered} s_{t+1|t} = Fs_{t|t} \qquad y_{t+1|t} = H's_{t+1|t} \\ \Sigma_{t+1|t} = F\Sigma_{t|t}F' + GQG' \qquad \Omega_{t|t-1} = H'\Sigma_{t|t-1}H + R \end{gathered}$$

The Bayes theorem leads to a linear updating rule:

$$\begin{aligned} s_{t|t} &= s_{t|t-1} + K_t\left(y_t - y_{t|t-1}\right) \\ \Sigma_{t|t} &= \Sigma_{t|t-1} - K_tH'\Sigma_{t|t-1} \end{aligned}$$

with Kalman gain equal to $K_t = \Sigma_{t|t-1} H \Omega^{-1}_{t|t-1}$.

Applying these seven equations recursively follows the general procedure closely and gives the best linear predictors of the mean and variance of the state of the system given all the observations at time T.

The likelihood of the Kalman filter has a closed form, based on the normal distribution:

$$\ln p\left(y^T | F, G, H, Q, R\right) = \Sigma \ln p\left(y_t | y^{t-1}; \gamma\right) = -\Sigma_{t=1}^{T} \left[\frac{N}{2} \ln 2\pi + \frac{1}{2} \ln \left|\Omega_{t|t-1}\right| + \frac{1}{2} \Sigma_{s=1}^{t} v_s' \Omega_{s|s-1} v_s\right],$$

where $v_t = y_t - y_{t|t-1}$.

It is important to mention that both the choice of the measurement equation and of the initial conditions affects the results. The introduction of measurement errors is necessary to avoid stochastic singularity, but in most real world applications it shouldn't account for more then 10-15 percent of the variation in the data. To avoid problems with initial conditions it's useful to either make them part of the parameter vector, or use the steady-state values of the states as initial ones.

Particle Filter

Sometimes a linear gaussian modelling framework cannot address the questions of interest:

1) The shocks are big relative to the steady state values, hence linearization does not do a good job

2) Both the analysis of the risk-premium and welfare implications of policies rely on the curvature of preferences, which requires a second-order approximation of the model

3) Fat tails or skewness of the distribution of shocks may have important economic implications

4) Markov-switching models require computing Lebegue integrals, which the Kalman filter (using Rieman integrals) cannot do.

In this case a more general version of the algorithm above needs to be implemented. It is called the particle filter. The particle filter is based on the idea of sequential Monte-Carlo Importance Sampling. A brief description includes the following steps:

1. drawing from the distribution of the initial state,
2. computing the probabilities of the first observation conditional on the initial state
3. drawing from the distribution of shocks to transit to a new state
4. computing the probability of the next observation conditional on each of the states
5. weighting the conditional probabilities in step 4 by their sum
6. resampling the states from step 3 with weights from step 5 to get the new draw of states
7. summing up all the conditional probabilities in step 4 to obtain the likelihood
8. using the result of step 6 to go to step 3 and transit to the next period
9. repeating steps 3-8 recursively untill the end of the sample.

It is important to mention that the particle filter avoids direct computations of integrals. Instead, it uses the importance sampling approach to resample draws from the conditional distribution of states every period and then uses the law of large numbers to compute the likelihood. It has nice assymptotic properties and avoids the problem of sample depletion. However, it is still very computationally demanding.

Part 2 (Juan Rubio-Ramirez)

Introduction

As discussed in part 1 of these lecture notes, in order to estimate a structural economic model given some data a researcher should take four general steps. First the model should be clearly specified in terms of choice functions of the agents which they optimize given some constraints, the resource constraints of the economy, the assumptions about expectation formation and stochastic processes driving the shocks. Then the model should be solved in terms of policy functions for all the endogenous variables, which allows to write the model in terms of a transition equation and a measurement equation. After the model is specified and solved, and a state-space representation is obtained, filtering theory is used to compute the likelihood function of the data given a parameter vector. This could be done using a Kalman filter if the model is linear and gaussian, or using the particle filter in more comlicated cases.

The fourth step is to use the likelihood of the observations given a vector of parameters to find an optimal parameter value or an optimal distribution of the parameter value. A classical econometrician would find the maximum likelihood estimate, while a Bayesian econometician would obtain a posterior distribution of parameters using some economically sound prior. This part of the lecture notes gives a brief overview of computational techniques used by Bayesian econometricians to find the posterior distribution of the parameters. An example of the standard RBC model is then taken to the data using both classical and Bayesian techniques.

Classical vs Bayesian

In part 1 of the lecture notes we saw how to obtain the likelihood function of the data given specific values of parameters $L\left(Y^T|\theta\right)$. That means that for every value of the parameter vector θ we can come up with a number, characterizing the relative (unnormalized) likelihood of the data. A classical econometrician would then use some robust maximizing algorithm to find the parameter combination which achieves the maximum value of the likelihood, therefore finding the mode of the distribution:

$$\hat{\theta}_{ML} = \arg\max_{\theta} L\left(Y^T|\theta\right).$$

This method leads to a consistent, information-efficient and asymptotically normal estimate of the structural parameters of interest. However, maximization is in general a very difficult task.

A Bayesian econometrician would combine the likelihood of the data with prior knowledge about the parameters of interest to obtain the posterior distribution and the marginal data density:

$$\pi\left(\theta|Y^T\right) = \frac{L\left(Y^T|\theta\right)\pi_0(\theta)}{\int L(Y^T|\theta)\pi_0(\theta)d\theta}$$
$$P\left(Y^T\right) = \int L\left(Y^T|\theta\right)\pi_0\left(\theta\right)d\theta$$

In reality the sample is always finite, so assymptotic properties are not very useful. One implication of the finite sample is that classical estimation assumes the possibility of taking the number of observations to infinity, thus violating the likelihood principle by taking into account the potential outcomes that were never observed. The likelihood principle states that all the information comes from the data, and hence is contained in the likelihood of the data given parameter values. When

finding confidence intervals of parameters a classical econometrician would be projecting the model into the infinite future, thus using observations that never occured.

The difference in the way the two econometricians think about the data is as follows. A classical person assumes that the experiment could be reproduced an infinite number of times and the data is one of the many possible outcomes. A Bayesian experimenter assumes that the data is given, i.e. the experiment has been performed and cannot be reproduced. Therefore, the Bayesian estimates based on the likelihood principle would be consistent ex post, i.e. when experiment has already taken place, while classical estimates would be consistent ex ante, as if the experiment has not taken place yet.

The advantages of classical estimation have already been stated. However it violates the likelihood principle and does not apply to small samples. The Bayesian approach does well both in small samples and asymptotically. Another advantage is that it's convenient to deal with misspecified models. The main disadvantage and the reason it hadn't been used untill recently is its computational intensity.

Metropolis-Hastings

This section discusses the main ideas used to obtain parameters of the posterior distribution. Let's abstract from the likelihood for a moment and assume that we found a way to draw from the posterior distribution. Then computing any moment of the posterior distribution would be a relatively simple task if we use the law of large numbers:

$$E_{\pi(\theta|Y^T)}h(\theta) = \int h(\theta)\,\pi\left(\theta|Y^T\right)d\theta \simeq \tfrac{1}{n}\Sigma_{i=1}^{n}h(\theta_i)$$

Therefore learning about the posterior distribution is equivalent to being able to draw from it, i.e. being able to generate (pseudo) random numbers, which would be distributed as the posterior. To be able to draw from a particular distribution that's hard to characterize analytically, Markov Chains are typically used. Let π_S be the distribution of interest, and let's assume that we found a transition kernel $P(x, B)$ such that the distribution of interest is a stationary distribution for this kernel:

$$\pi_S(B) = \int P(x, B)\,\pi_S(dx)$$

That is $P(x, B)$ defines a Markov Chain, with a fixed point being the distribution of interest. There are two additional conditions: we need the stationary distribution π_S to be unique for $P(\,)$, and we need the Markov Chain to converge assymptotically to the stationary distribution. Then we could start with an arbitrary value of parameters and use this Markov Chain to recursively obtain new values which would then by construction have the distribution of interest. That would allow us to draw from the posterior distribution and, hence fully characterize its properties.

Let's think of a class of transition kernels with a transition function $p(x, y)$ and a rejection probability $r(x)$:

$$P(x, dy) = p(x, y)\,dy + r(x)\,\delta_x(dy)$$

Its most interesting property is that if $p(x, y)$ is time-reversible, i.e. $f(x)\,p(x, y) = f(y)\,p(y, x)$ for any $f(\,), x, y$, then $\int_A f(y)\,dy = \int_X P(x, A)\,f(x)\,dx$ and the resulting Markov Chain converges to a unique stationary distribution and satisfies the properties of irreducibility and aperiodicity.

Different versions of Markov-Chain Monte-Carlo methods are based on a different choice of the transition function $p(x, y)$.

The Metropolis-Hastings algorithm uses a function of the following form:

$$p_{MH}(x, y) = \alpha(x, y)\, q(x, y),$$

where $q(x, y)$ is a known distribution, and

$$\alpha(x, y) = \min\left\{\frac{f(y)q(y,x)}{f(x)q(x,y)}, 1\right\}.$$

It's easy to show, that using a symmetric density function $q(x, y) = q(y, x)$ is very useful and makes life a lot simpler. A classical example of such a distribution is a random walk:

$$y = x + \varepsilon,\ \varepsilon \sim N(0, \sigma^2).$$

In this simple form the Markov Chain Monte-Carlo method the algorithm boils down to wandering along the parameter space $\theta \in \Theta$ using a random walk:

$$\theta^*_{i+1} = \theta_i + \varepsilon,\ \varepsilon \sim N(0, \Sigma)$$

Then the value of the posterior $\pi\left(\theta^*_{i+1}|Y^T\right)$ is computed for the proposed vector of parameters. It's trivial to verify that since the transition function characterizing the random walk is symmetric: $q(x, y) = q(y, x)$, - the expression for the transition probability boils down to: $\alpha(x, y) = \min\left\{\frac{\pi\left(\theta^*_{i+1}|Y^T\right)}{\pi(\theta_i|Y^T)}, 1\right\} = \min\left\{\frac{L\left(Y^T|\theta^*_{i+1}\right)\pi_0\left(\theta^*_{i+1}\right)}{L(Y^T|\theta_i)\pi_0(\theta_i)}, 1\right\}$. Therefore, we don't even need to compute the integral - the posterior could be computed up to an arbitrary constant.

Then essentially, if we are travelling up in the distribution, i.e. $\pi\left(\theta^*_{i+1}|Y^T\right) \geq \pi\left(\theta_i|Y^T\right)$, then $\alpha(x, y) = 1$ and we always accept the proposal: $\theta_{i+1} = \theta^*_{i+1}$. Otherwise, if the proposed vector of parameters has a lower probability, i.e. $\pi\left(\theta^*_{i+1}|Y^T\right) < \pi\left(\theta_i|Y^T\right)$ and we are travelling down in the distribution, then $\alpha(x, y) = \frac{\pi\left(\theta^*_{i+1}|Y^T\right)}{\pi(\theta_i|Y^T)} < 1$. In this case we draw a uniform random number $\tilde{z}$ and compare it with α, thus accepting the proposal with probability α and rejecting it with probability $1 - \alpha$. When rejecting the proposal, we keep the original parameter vector and don't transit to the new one: $\theta_{i+1} = \theta_i$.

The choice of the initial value and the number of iterations in the Metropolis-Hastings algorithm are important. If θ_0 is far from the center of the distribution, then the Markov Chain will be travelling towards the center of the distribution for a significant number of periods, and once it comes to that region, it will then stay there forever. Hence, the estimates of all the moments will be biased, if the initial vector is far from the center of the distribution. Therefore, it's useful to start from the mode of the posterior, which is equivalent (in terms of speed) to finding the maximum likelihood estimate first.

To evaluate whether the Markov Chain was long enough one needs to verify that the means and variances of the parameters are not moving. Slow convergence may be a result of a bad initial guess as well as of serial correlation of the draws, which should be avoided if possible. In case the distribution has multiple modes, it's useful to draw random starting points in the vicinity of the mode and then combine the results. Another useful robustness check is the average acceptation rate, which should be in the range $[23\%, 45\%]$ according to Roberts, Gelman and Gilks (1994). It's

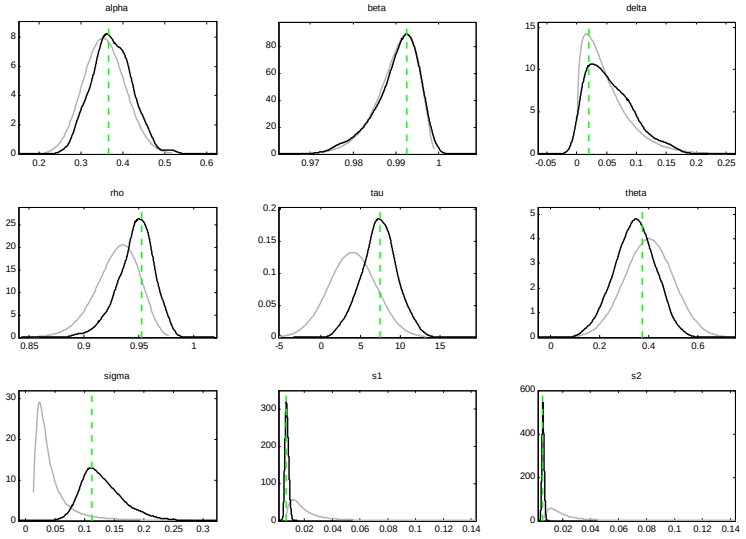
also useful to set the matrix Σ of the random walk proportional to the inverse of the Hessian of the posterior at the mode.

Another important question is how to choose a good prior. The mistake which is often made when a researcher wants to find a "non–informative" prior is to set the prior density to be equal to one. This leads to a wide use of improper priors, which do not have a well-defined pdf, and, hence, violate the likelihood principle. Often the main reason is that an econometrician wants to be classical but to use bayesian methods and introduces a flat prior. Uniform priors are improper most of the time. To discourage the use of flat priors it's important to note, that almost any prior is not invariant to reparametrization. Therefore, even if one uses a uniform distribution as a prior for the parameter of interest, one can reparametrize the model by taking the inverse of the parameter, and then the new prior will not be flat any more, while equivalent to the original one. Hence, almost any prior is informative.

Prior and Posterior Table

	prior dist	mean	s.d.	post mode	s.d.	mean	lower	upper
α	beta	0.356	0.05	0.366	0.046	0.374	0.296	0.448
β	beta	0.990	0.005	0.993	0.004	0.990	0.983	0.997
δ	beta	0.050	0.04	0.020	0.030	0.057	0.001	0.110
ρ	beta	0.930	0.02	0.953	0.014	0.947	0.922	0.974
τ	normal	4.000	3.00	7.506	2.097	7.354	3.666	10.88
θ	normal	0.400	0.10	0.371	0.083	0.341	0.205	0.467
σ_ε	inv gam	0.050	Inf	0.111	0.026	0.130	0.075	0.182
σ_1	inv gam	0.025	Inf	0.007	0.001	0.007	0.005	0.009
σ_2	inv gam	0.025	Inf	0.006	0.001	0.006	0.005	0.007

Prior and Posterior Graph

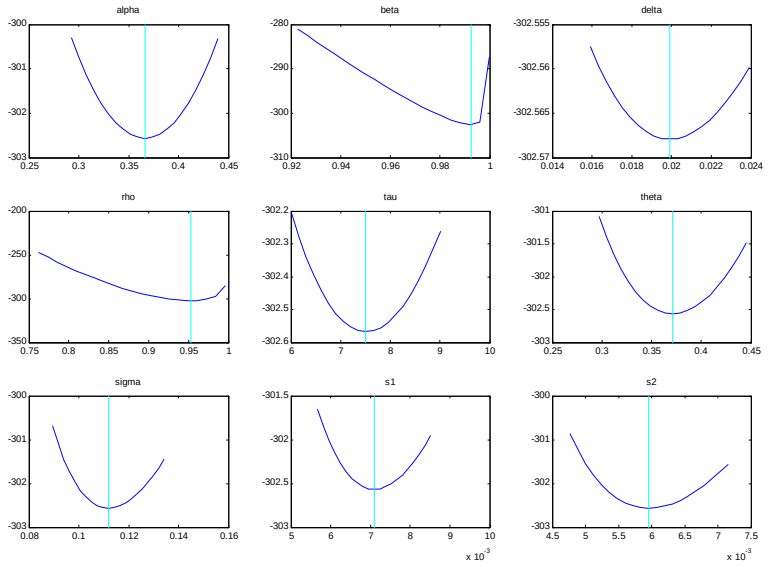


Example

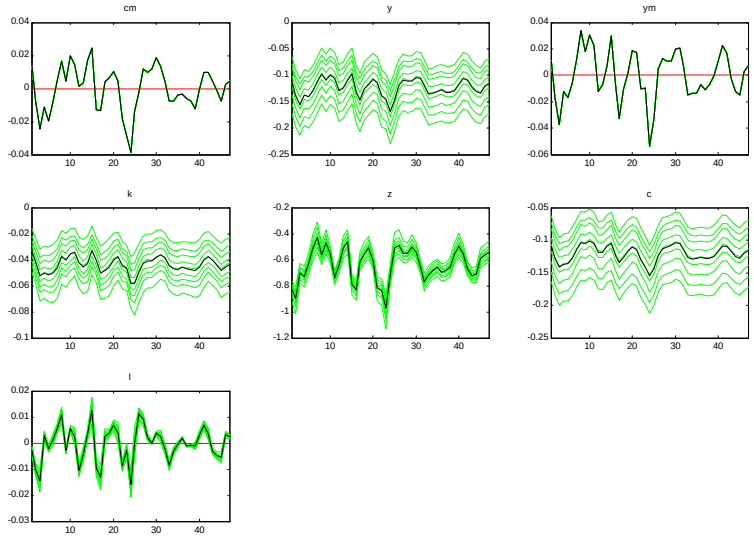
As an example of the use of the McMc algorithm to obtain a posterior I use the standard version of an RBC model, described in part 1 of the lecture notes, and solve it using a first-order perturbation

(linearization) method. I use data on hp-filtered yearly data on real consumption and output in the US in the last 45 years.

Mode Check



The pictures show that the standard RBC model with standard priors returns a reasonable posterior. However, for at least half of the parameters the posteriors almost coincide with the priors, which means that the data is not very informative about those parameters. Five blocks 2000 iterations each were enough to get a reasonable description of the posterior with the acceptance ratio around 30%. The data is not informative about technology coefficients α, β, δ. However, data indicates extremely high risk aversion and persistence of the TFP shock close to a random walk. Measurement errors appear to be small. Dynare also provides tools for smoothing the state variables of the model.



Part 3 (Fabio Canova)

Introduction

The first section of part 3 of the lecture note discusses the various methods of searching for and dealing with identification problems in DSGE models. Briefly the problems can be coming from the model, from the objective function and from the finite sample. If the parameters have no effect on the dynamic properties of the model, the model should be changed. If the moments of the data incorporated into the objective function do not carry information about the parameters of interest, a full information (likelihood) approach is preferrable. If the sample is not big enough to infer the parameters of interest, the dataset should be expanded.

The second section contains an overview of different techniques used to extract the trend and cycle components from non-stationary data. These are the Hodrick-Prescott, the Band-Pass and the Christiano-Fitzgerald filters. The lecture discusses each of the setups, problems associated with their use and spurious cycles they can generate. It argues that one should be very careful when choosing a particular way of pre-filtering the data and one should avoid pre-filtering if possible. The third section briefly discusses model evaluation and selection techniques.

Identification Issues in DSGE models

There are many ways one can estimate DSGE models. The first set of methods called Limited Information methods uses only a particular characterization of the properties of the data. These include generalized method of moments (GMM) which finds the parameters that match a particular set of moments of the data, minimal distance methods which match impulse responses of the model to the data, and sturctural vector autoregressions (SVAR) with magnitude and sign restrictions (Canova (2002)).

The second set of methods called Full Information methods uses all the information incorporated in the data by computing its likelihood. These include maximum likelihood (ML) which finds the mode of the likelihood and estimates standard errors using the Hessian at the mode, and the bayesian approach, which finds the posterior as a multiple of the likelihood and prior distribution of the parameters of interest.

The last set of methods can be called calibration, because all the parameters are postulated in order to recover the shocks hitting the system. The business cycle accounting framework also falls into this category.

When matching impulse responses $X_t^M(\theta) = C(\theta)(l)\, e_t^j$ to a model-identified shock e_t^j with data responses $X_t = \hat{W}(l)\, e_t^j$ the distance between the two is minimized to obtain an estimate of the corresponding vector of parameters: $\hat{\theta} = \arg\min_\theta \left\| X_t - X_t^M(\theta) \right\|_{W(T)}$, with $W(T)$ being the weighting matrix. When using the maximum likelihood approach, the vector of parameters maximizes the likelihood of the data given parameters $\hat{\theta} = \arg\max_\theta L(X, \theta)$. When using the bayesian approach, the vector of parameters is the mean of the posterior distribution $\hat{\theta} = \int \theta P(\theta|X)\, d\theta$ or the one that achieves the mode of the marginal density by maximizing a function which is the product of the likelihood and the prior $\hat{\theta} = \arg\max_\theta L(X, \theta)\, \pi_0(\theta)$.

The concept of *identification* is connected with the question whether the mapping from the parameters to the objective function is well behaved. For the parameter to be identified one needs the objective function to have a unique minimum (maximum) at the true parameter with a positive (negative) definite Hessian having a full rank. In finite samples it is also important that the curvature

of the objective function is sufficient to find the optimal value.

In practice these are difficult to verify because the mapping from structural parameters to solution parameters is unknown, the objective function is typically a nonlinear function of solution parameters. Besides, different objective functions may have different "identification power" and the standard rank and order conditions can't be used due to significant non-linearity of the mapping from structural to solution parameters.

There is a solution identification problem if one cannot recover structural parameters from the matrices describing the aggregate decision rule. There is an objective function identification problem if one cannot recover aggregate decision rule matrices from the objective function. There is a population identification problem if either of the two problems exists in which case even using an infinite data sample one cannot recover the parameters of interest using a particular method. In addition there is a sample identification problem if the sample is not big enough to recover the parameters of interest, given that parameters are identified in population. All of these problems can occur separately or in conjunction.

If there is a solution identification problem, that means the problem is due to model specification. In this case in order to make inference about the parameter of interest the researcher has to change the model. If there is an objective function identification problem, one should change the objective function. For example, different moments of the data carry information about different parameters of the model. Full information methods are generally preferrable to limited information approaches because they take into account all of the moments of the data.

The first kind of problem that may arise is observational equivalence of two models. In this case two models would lead to the same minimized value of the objective function for two different models or at two different vectors of parameters of the same model. One example is a New Keynesian model with sticky prices or sticky wages. Another example comes from the paper of Beyer and Farmer (2004). They argue that a completely forward looking model of the form:

$$x_t = \tfrac{1}{\lambda_1} E_t x_{t+1}$$

and a model with both forward-looking and backward-looking effects:

$$x_t = \tfrac{1}{\lambda_1+\lambda_2} E_t x_{t+1} + \tfrac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} x_{t-1} + v_t$$

have the same backward-looking solution:

$$x_t = \lambda_1 x_{t-1} + w_t.$$

This implies that the three models are observationally equivalent and lead to the same minimized value of the objective function. Another similar case is the new-keynesian model with sticky prices, which gives the same law of motion for all the variables for two different values of the Calvo parameter: 0.25 and 0.75.

A similar case of partial or under-identification is when a subset of parameters of the model can't be identified because the objective function uses only a portion of the restrictions of the solution or a subset of the structural parameters enters in a particular functional form in the solution or disappears from the solution. A typical example of this case is the New-Keynesian model with a forward-looking policy rule:

$$x_t = a_1 E_t x_{t+1} + a_2 \left(R_t - E_t \pi_{t+1}\right) + v_{1t}$$
$$\pi_t = a_3 E_t \pi_{t+1} + a_4 x_t + v_{2t}$$
$$R_t = a_5 E_t \pi_{t+1} + v_{3t}$$

This models implies a solution of the form:

$$\begin{bmatrix} x_t \\ \pi_t \\ R_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & a_2 \\ a_4 & 1 & a_2 a_4 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_{1t} \\ v_{2t} \\ v_{3t} \end{bmatrix}$$

Depending on which impulse response the researcher tries to match she can recover some or none of the parameters. However some of the parameters she cannot recover whatever method she uses.

A more involved case of weak/asymmetric identification is when the population mapping is very flat or asymmetric in some dimensions. This could be local or global as well as due to a particular objective function or occur for all objective functions. An example is the classical real business cycle model:

$$\max \Sigma_{t=0}^{\infty} \beta^t \frac{c_t^{1-\sigma}}{1-\sigma}$$
$$s.t. \qquad c_t + k_{t+1} = z_t k_t^{\alpha} + (1-\delta) k_t$$

In this model you cannot separately identify α and σ (weak identification) as well as β and δ which are partially under-identified. In this case changing the method and the objective function won't help because the problem is inherent to the model.

To deal with these problems one should first correctly identify the problem. One major way of helping solve identification problems is by using all the information available, which means using the likelihood function. Therefore use of full information methods is preferrable to limited information methods. When you encounter weak/partial under-identification, fixing some of the parameters may be a solution, but the results may significantly depend on the choice of the exact value. A more robust approach is provided by Bayesian econometrics, which deals with these problems using priors. Fixing a parameter is therefore equivalent to a very tight prior. The Bayesian approach allows the researcher to vary the tightness of prior information.

However, neither full information nor Bayesian approaches can help, if the problem remains in population. You can identify this case by the posterior distribution being identical to the prior, which implies that data carries no information about the parameter of interest. A more complicated combination of the two can arise, when the problem is in population, but a prior on another parameter implicitly restricts the parameter of interest, which leads to the posterior being different from the prior. It is very hard to detect those kind of identification problems.

There are several beliefs which are not true in general. Second-order approximations usually have more curvature and help solving identification problems, but depending on the particular parameter combination, the second order approximation can actually have less curvature around the true parameters. A related problem may arise when a second-order approximation is used to compute welfare costs - very small changes in priors can cause big changes and even reversal of welfare implications.

Ways to diagnose identification problems include:

1. Plots/Preliminary exploration of the objective function

2. Numerical derivatives of the objective function at likely parameter values

3. Condition numbers (ratios of largest to smalles eigenvalues) of the Hessian at the mode
4. Erratic parameter estimates as sample size increases
5. Large or uncomputable standard errors
6. Crazy t-tests

As a rule for identification the model needs the state variables of the model to react to changes in structural parameters. If that is not the case, the only way to solve the identification problem is to respecify the model. Also usually likelihood methods are preferrable to limited information approaches, bigger samples to smaller samples, and full calibration or bayesian calibration preferable to mixed calibration-estimation.

An important recent contribution by Iskrev (2007) proposes a new empirical approach to study parameter identification problems in DSGE models based on analytical evaluation of the Information matrix of such models. The information matrix is decomposed into two parts: one that describes the identification properties of the model, and one that summarizes the properties of the data. This allows the researcher to find out separately, whether the parameters of the model are identified, whether identification is strong or weak, and whether identification problems come from the model or from the data.

Extracting Cyclical Information

Most theories model long-run trends separately from cyclical fluctuations around those trends. Most theories of business cycles are stationary while most data series available are non-stationary. An important practical issue is how to decompose nonstationary time series of interest into a permanent and cyclical component. Any decomposition of data into a cyclical component and a trend embeds an implicit assumption on how they differ from each other.

When the model can generate non-stationary data the way to compare it with real data is generally to apply the same methods/statistics/impulse-responses to both the actual data and data simulated from the model. Comparing theoretical impulse-responses with impulse-responses from the data is usually a mistake.

When the theory does not imply any trends and hence the model cannot generate non-stationary data, some ad hoc assumptions have to be used to filter out the cyclical component. The most commonly used methods are the Hodrick-Prescott filter and the Band-Pass filter.

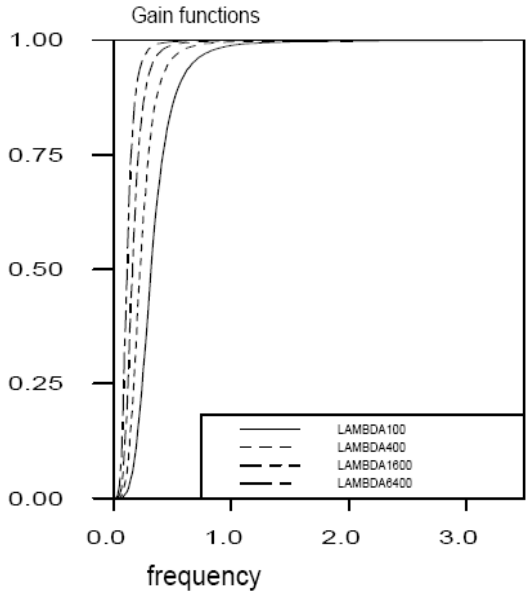
The HP filter minimizes the distance between the trend and the data, given a constraint on the speed of change of the derivative of the trend. The tightness of the constraint is controlled by a smoothnessparameter λ :

$$\min_{y_t^x} \Sigma_{t=0}^{T} \left\{ (y_t - y_t^x)^2 + \lambda \left[\left(y_{t+1}^x - y_t^x \right) - \left(y_t^x - y_{t-1}^x \right) \right]^2 \right\}$$

If λ goes to 0, then the trend is equal to the data: $y_t^x = y_t$. Typically (for US data) quarterly data trends have a parameter around $\lambda = 1600$. Ravn and Uhlig (2002) find that the optimal $\lambda = 129000$ for monthly data, and $\lambda = 6.25$ for yearly data. The idea comes from the curve-fitting literature, which chooses λ to make mean square error of the fitting error minimal, implying $\lambda^* = \frac{\sigma_{cycle}^2}{\sigma_{trend}^2}$. Decomposition of the HP filter shows that it's equivalent to an operator of the form:

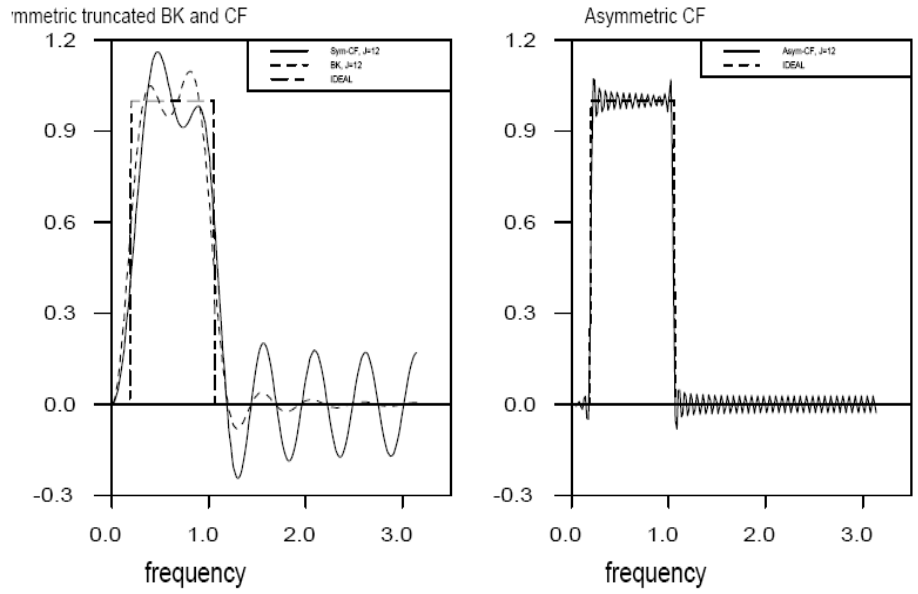
$$B^c(L) = B(L)(1-L)^4 \simeq \frac{(1-L)^2\left(1-L^{-1}\right)^2}{\frac{1}{\lambda}+(1-L)^2(1-L^{-1})^2}.$$

It's phase diagram cuts the low frequency component of the data:



As can be easily seen, it eliminates linear and quadratic trends and makes integrated series of order four (I(4)) stationary. When applied to stationary data, it damps fluctuations with periodicity 24-32 quarters per cycle, and passes short cycles without changes. However, if the series is close to a random walk (I(1)), then HP filter damps long and short run growth cycles, and amplifies growth cycles at the business-cycle frequency. For example, the variance of cycles with average duration of 7.6 years is multiplied by a factor of 13, which generates spurious cycles. The problem is even larger when the time series is integrated of order two (I(2)). Therefore, the HP filter can produce spurious business cycle phenomena from random walk data. Other problems include asymmetric importance of data at the beginning and the end of the dataset, unwanted high frequencies left in the data set, and different cycle lengths making cross-country comparisons difficult. A partial solution to the latter problem has been proposed by Marcet and Ravn (2000), restricting the relative variability of the trend to the cycle across countiries.

Band-pass filters are combinations of high-pass and low-pass moving-average filters, which, following Baxter and King (1999) truncate the infinite dimensional inverse Fourier transform: $B(L) = (1-L)(1-L^{-1})B^*(L)$. This approach makes stationary series with up to quadratic deterministic trends and makes series stationary if they are integrated of order up to two (I(2)). Therefore, it has problems similar to the HP filter. Christiano and Fitzgerald show that non-stationary asymmetric approximation of a band-pass filter is optimal by avoiding important leakage, compression and amplification effects for different frequencies (see picture).



The major problem with this approach is that to apply the BK or CF version of the filter the researcher needs to know the properties of the time series of interest. The asymmetric filter also changes the lead-lag properties of the time-series, creating important phase shifts. Moreover, the CF filter has time-varying parameters, which makes the properties of the resulting series non-comparable for different time intervals.

Given all these problems the main prescription is to avoid pre-filtering the data when possible. Another alternative would be to use model-based filters, which would derive the distinction between the cycle and the trend conditioned on the model. For example when theory implies particular closed forms for the trend, their parameters can be estimated jointly with other parameters of the model.

Model Evaluation and Forecasting

Let the model be described by the equation, $x = m(p, \beta_m, \varepsilon)$, where p is the policy variable, m is the model, β_m are parameters of model m, ε is the random error.

1. If the model parameters and errors are know, then set up a loss function $l(x)$ and find the policy p which mimizes it:

$$\min_p l(x) = \min_p l(m(p, \beta_m, \varepsilon))|_{\beta_m, \varepsilon}.$$

2. If the errors are unknow, but their pdf μ_ε is known, then $l(x)$ is a random variable and for each p the loss function has a distribution. In this case one should evaluate $l_{\mu_\varepsilon}(m(p, \beta_m, \varepsilon))$ and find the p which minimizes it's variability: $\min_p Var\left(l_{\mu_\varepsilon}(m(p, \beta_m, \varepsilon))\right) =$

$$\min_p \left[\int l_{\mu_\varepsilon}(m(p, \beta_m, \varepsilon))^2 d\mu_\varepsilon - E l_{\mu_\varepsilon}(m(p, \beta_m, \varepsilon))^2\right].$$

3. If both the errors and the parameters β_m are unknown, then evaluate policies using $l_{\mu_\varepsilon}\left(m\left(p, \hat{\beta}_m|_d, \varepsilon\right)\right)$ and an estimate $\hat{\beta}_m$ of the parameter given the data d or average over an empirical distribution of parameters

$$\int l_{\mu_\varepsilon}\left(m\left(p,\beta_m,\varepsilon\right)\right)d\mu\left(\beta_m|d\right)d\mu_\varepsilon$$

where $d\mu\left(\beta_m|d\right)$ is the posterior distribution of the parameters conditional on the data.

4. If on top of that the model is unknow, one should use model selection criteria such as AIC and BIC and then estimate $l_{\mu_\varepsilon}\left(\hat{m}\left(p,\hat{\beta}_m|_d,\varepsilon\right)\right)$ or use model averaging:

$$\Sigma_m\mu\left(m|d\right)\int l_{\mu_\varepsilon}\left(m\left(p,\beta_m,\varepsilon\right)\right)d\mu\left(\beta_m|d\right)d\mu_\varepsilon$$

where $\mu\left(m|d\right)$ is the posterior of the model given data.

There are two types of model evaluation excercises: in-sample and out-of-sample. In-sample excercises evaluate the fit of the model. Out-of-sample excercises consider the forecasting properties of the model. The main ways of assessing in-sample fit are comparing root mean square differences between predicted and actual values (RMSE), correlations between predicted and actual values, tests of in-sample performance (when joint and separate estimation of two or more models is performed and χ^2 tests on implied restrictions are performed), unbiasedness regressions, predictive regressions, unobserved factor models.

If models have a Bayesian setup, one can use marginal likelihood, even with non-informative but proper priors. The marginal likelihood of the data given the model is

$$ML\left(y|j\right)=\int f\left(y,\beta\right)g\left(\beta|M_j\right)d\beta.$$

Then comparing alternative models j and k can be done using a posterior odds ratio:

$$PO=\frac{g(M_j)ML(y|j)}{g(M_k)ML(y|k)}.$$

Assessing out-of-sample fit of the model boils down to estimating the model using only a part of the samle and then projecting it into the future and comparign the same statistics to evaluate the difference between actual data and the forecast coming from the model.

A novel method of evaluating DSGE models has been proposed by Del Negro, et. al. (2006). It combines data simulated from a DSGE model with actual data and estimates a VAR using the big sample. This is equivalent to weighting a highly restricted DSGE model with a completely unrestricted VAR with weights λ and 1. $\lambda=0$ is equivalent to not using the model, while $\lambda=\infty$ means the data doesn't contradict the model. The interpretation of this approach is in taking an optimal combination of a completely unrestricted model and a very restricted one.

If the DSGE model is bad, the weight on it will be low, while if it's good, it can improve the fit of the VAR and the weight will be high. The approach can also be interpreted as using a DSGE model as a prior for a VAR. The main advantage of the approach is in joint estimation of the parameters of the DSGE model β and the degree of misspecification of the model λ by maximizing the marginal likelihood.

Part 4 (Tao Zha)

Introduction

The first section of part 4 of these lecture notes discusses identification issues in Structural Vector Autoregressions (SVARs). Sufficient conditions for both local and global identification are described. These conditions extend the existing literature by providing simple means of testing for identificaion in SVARs. The second section contains an overview of methods for estimating Regime-Switching DSGE models using the MCMC algorithm. The last section contains a discusion of efficient ways to compute the marginal data density.

Identification Issues in SVARs

Structural vector autoregressions can in general be represented by the following equation:

$$y_t'A_0 = \Sigma_{l=1}^{p} y_{t-l}'A_l + z_t'C_z + \varepsilon_t'$$

The compact form of this expression can be written as

$$y_t'A_0 = x_t'A_+ + \varepsilon_t'$$

The reduced form is then described by:

$$y_t' = x_t'B + u_t',$$

where $B = A_+A_0^{-1}, u_t' = \varepsilon_t'A_0^{-1}, E\left(u_tu_t'\right) = \Sigma = \left(A_0A_0'\right)^{-1}$

Theorem 1 *Two sets of structural parameters,* (A_0, A_+) *and* $\left(\tilde{A}_0, \tilde{A}_+\right)$*, are observationally equivalent, i.e.* $A_+A_0^{-1} = \tilde{A}_+\tilde{A}_0^{-1}$ *and* $A_0A_0' = \tilde{A}_0\tilde{A}_0'$*, if and only if there exists an orthogonal matrix* P *such that* $A_0 = \tilde{A}_0P$ *and* $A_+ = \tilde{A}_+P$*.*

Definition 1 *The set of structural parameters* (A_0, A_+) *is globally identified if and only if there is no other set of parameters* $\left(\tilde{A}_0, \tilde{A}_+\right)$ *that is observationally equivalent to* (A_0, A_+)*.*

Definition 2 *The set of structural parameters* (A_0, A_+) *is locally identified if and only if there exists an* $\varepsilon > 0$ *such that there is no other set of parameters* $\left(\tilde{A}_0, \tilde{A}_+\right)$ *in the open* ε*-ball of* (A_0, A_+) *that is observationally equivalent to* (A_0, A_+)*.*

The class of restrictions of interest can be written as:

$$R = \left\{(A_0, A_+)\,|Q_j f\left(A_0, A_+\right) e_j = 0 \forall j = \overline{1, n}\right\}$$

If restrictions are applied directly to the SVAR representation then the function

$$f\left(A_0, A_+\right) = \left[\begin{array}{c} A_0 \\ A_+ \end{array}\right]$$

The h^{th} impulse responses are defined as:

$$L_h = \left(A_0^{-1} J' \hat{B}^h J\right)'$$

where

$$\hat{B} = \begin{bmatrix} A_1 A_0^{-1} & I_n & ... & 0 \\ ... & ... & ... & ... \\ A_{p-1} A_0^{-1} & 0 & ... & I_n \\ A_p A_0^{-1} & 0 & ... & 0 \end{bmatrix} \text{ and } J = \begin{bmatrix} I_n \\ 0 \\ ... \\ 0 \end{bmatrix}$$

Restrictions on impulse responses can be imposed using the following function:

$$f(A_0, A_+) = \begin{bmatrix} L_0' & ... & L_p' \end{bmatrix}'.$$

The long-run impulse responses are equivalent to:

$$L_\infty = (A_0' - \Sigma_{l=1}^p A_l')^{-1}$$

In this case $f(A_0, A_+) = \begin{bmatrix} L_0' & L_\infty' \end{bmatrix}'$.

Definition 3 *If* $(A_0, A_+) \in R$ *and* $M_j(X) = \begin{bmatrix} Q_j X \\ \begin{bmatrix} I_j & 0 \end{bmatrix} \end{bmatrix}$ *is of rank* n *for all* $j = \overline{1, n}$ *then SVAR is globally identified.*

In this specification k is the number of restrictions, n is the size of the state vector, j takes values from 1 to n, Q_j has size $k \times k$, X has size $k \times n$, M_j has size $(k + j) \times n$, 0 has size $j \times (n - j)$.

Theorem 2 *An SVAR with restrictions* R *is exactly identified if and only if for almost any reduced form* (B, Σ) *there exists a unique structural parameter* $(A_0, A_+) \in R$ *such that* $B = A_+ A_0^{-1}, \Sigma = (A_0 A_0')^{-1}$.

Theorem 3 *An SVAR is exactly identified if and only if for almost every structural parameter* (A_0, A_+) *there exists an orthogonal matrix* P *such that* $(A_0 P, A_+ P) \in R$.

Theorem 4 *An SVAR is exactly identified if and only if the total number of restrictions is equal to* $\frac{n(n-1)}{2}$ *and the rank condition of theorem 1 is satisfied.*

Theorem 5 *An SVAR is exactly identified if and only if rank* $Q_j = n - j$ *for all* $j = \overline{1, n}$.

Let's illustrate these results using a standard New-Keynesian model:

$$\begin{gathered} \pi_t = \beta E_t \pi_{t+1} + \kappa x_t + \sigma_s \varepsilon_{st} \\ x_t = E_t x_{t+1} - \tau (R_t - E_t \pi_{t+1}) + \sigma_d \varepsilon_{dt} \\ R_t = \phi_\pi \pi_t + \sigma_R \varepsilon_{Rt} \\ \text{with } \varepsilon_t \sim N(0, I_3) \end{gathered}$$

The state of the system is a 3-vector $y_t = [\pi_t, x_t, R_t]'$. One can rewrite this system in matrix form as:

$$y_t' \begin{bmatrix} 1/\sigma_s & 0 & -\phi_\pi/\sigma_R \\ -\kappa/\sigma_s & 1/\sigma_d & 0 \\ 0 & \tau/\sigma_d & 1/\sigma_R \end{bmatrix} = E_t y_{t+1}' \begin{bmatrix} \beta/\sigma_s & \tau/\sigma_d & 0 \\ 0 & 1/\sigma_d & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} \varepsilon_{st} \\ \varepsilon_{dt} \\ \varepsilon_{Rt} \end{bmatrix}$$

The solution to this forward-looking model is

$$y_t' = \tfrac{1}{1+\kappa\tau\phi_\pi} \begin{bmatrix} \sigma_s & -\tau\phi_\pi\sigma_d & \phi_\pi\sigma_R \\ \kappa\sigma_s & \sigma_d & \kappa\phi_\pi\sigma_R \\ -\kappa\tau\sigma_s & -\tau\sigma_d & \sigma_R \end{bmatrix} \begin{bmatrix} \varepsilon_{st} \\ \varepsilon_{dt} \\ \varepsilon_{Rt} \end{bmatrix}$$

Since β does not enter the solution, it is not identified. The system can then be rewritten as:

$$y_t' A_0 = \varepsilon_{st} \text{ with } A_0 = \begin{bmatrix} 1/\sigma_s & 0 & -\phi_\pi/\sigma_R \\ -\kappa/\sigma_s & 1/\sigma_d & 0 \\ 0 & \tau/\sigma_d & 1/\sigma_R \end{bmatrix}$$

Clearly any two different points $(\sigma_s, \sigma_d, \sigma_R, \kappa, \tau, \phi_\pi)$ imply a different matrix A_0. Therefore rest of the parameters are locally identified. However there is no global identification. For example, if
$(\sigma_s = 2, \sigma_d = 1, \sigma_R = 0.2, \kappa = 0.58, \tau = 0.54, \phi_\pi = 2)$
imply the same dynamics as
$(\sigma = 2.5, \sigma_d = 1.02, \sigma_R = 0.2, \kappa = 0.9, \tau = 0.566, \phi_\pi = 2.5)$.
Another version of the same model includes a lag of the interest rate:

$$\begin{gathered} \pi_t = \beta E_t \pi_{t+1} + \kappa x_t + \sigma_s \varepsilon_{st} \\ x_t = E_t x_{t+1} - \tau \left(R_t - E_t \pi_{t+1} \right) + \sigma_d \varepsilon_{dt} \\ R_t = \rho_R R_t + (1 - \rho_R) \left(\phi_\pi \pi_t + \phi_x x_t \right) + \sigma_R \varepsilon_{Rt} \\ \text{with } \varepsilon_t \sim N(0, I_3) \end{gathered}$$

In this case the model can be written in the following lead-lag form:

$$y_t' B_0 = y_{t-1}' C + E_t y_{t+1}' D + \varepsilon_t, \qquad \text{where}$$

$$B_0 = \begin{bmatrix} 1/\sigma_s & 0 & -\phi_\pi/\sigma_R \\ -\kappa/\sigma_s & 1/\sigma_d & -\phi_x/\sigma_R \\ 0 & \tau/\sigma_d & 1/\sigma_R \end{bmatrix}, \qquad D = \begin{bmatrix} \beta/\sigma_s & \tau/\sigma_d & 0 \\ 0 & 1/\sigma_d & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \rho_R/\sigma_R \end{bmatrix}$$

The minimal state variable representation of this system satisfies $A_0 = B_0 - A_1 A_0^{-1} D$ and $A_1 = C$, that implies a matrix:

$$A_0 = \begin{bmatrix} 1/\sigma_s & 0 & -\phi_\pi/\sigma_R \\ -\kappa/\sigma_s & 1/\sigma_d & -\phi_x/\sigma_R \\ a_{31}^0 & \tau/\sigma_d & 1/\sigma_R \end{bmatrix} \text{ where } \beta = \frac{\sigma_s\left(1+\phi_\pi\tau(1+\kappa)+a_{31}^0\phi_\pi\sigma_s\right)a_{31}^0}{\rho_R\left(a_{31}^0\sigma_s+\kappa\tau\right)}.$$

To test the restrictions let's form all the necessary matrices:

$$f(A_0, A_1) = \begin{bmatrix} 0 & 1/\sigma_s & -\phi_\pi/\sigma_R \\ 1/\sigma_d & -\kappa/\sigma_s & -\phi_x/\sigma_R \\ \tau/\sigma_d & a_{31}^0 & 1/\sigma_R \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \rho_R/\sigma_R \end{bmatrix}, \text{ k=6,n=3,j=1,2,3.}$$

Here the columns of the stacked matrices A_0 and A_1 are sorted by the decreasing number of restrictions (zeros).

$$Q_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \Rightarrow \quad Q_1 f(A_0, A_1) = \begin{bmatrix} 0 & \frac{1}{\sigma_s} & -\frac{\phi_\pi}{\sigma_R} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{\rho_R}{\sigma_R} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$Q_2 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \Rightarrow \quad Q_2 f(A_0, A_1) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{\rho_R}{\sigma_R} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$Q_3 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \Rightarrow \quad Q_3 f(A_0, A_1) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Therefore, matrices $M_j(f(A_0, A_1))$ for $j = 1, 2, 3$ are equal to:

$$M_1 = \begin{bmatrix} 0 & \frac{1}{\sigma_s} & -\frac{\phi_\pi}{\sigma_R} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{\rho_R}{\sigma_R} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad M_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{\rho_R}{\sigma_R} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad M_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Since they all have rank 3, the system is globally identified.

Regime-Switching Models

Regime switches are introduced into DSGE models if there are reasons to assume, that dynamics of the model in different periods of time have different properties. This is typically a result of changes in structural parameters, representing policy, market structure, variances or persistence of shocks. A parsimonious way of capturing such changes is by letting the parameters take a finite set of values and switch between them following a first-order Markov chain. Then in addition to the main parameters of the model the transition matrix has to be estimated.

Let y_t be the vector of endogenous variables, z_t - vector of exogenous variables, $x_t = [y_t', z_t']'$, s_t - state governed by a Markov chain, θ - constant model parameters, $Q = [q_{ij}]$ - transition matrix with q_{ij} being the probability of transiting from state j to state i. Let the solution of the model be summarized by the state-space representation:

$$x_t = c(s_t, \theta) + F(s_t, \theta) x_{t-1} + G(s_t, \theta) \varepsilon_t$$
$$y_t = H(s_t, \theta) x_t + u_t$$

Then the likelihood of the data given the parameters can be computed using the formula:

$$L(Y_T|\theta, Q) = \Pi_{t=1}^{T} \Sigma_{s_t} p(y_t|Y_{t-1}, x_t, s_t; \theta, Q) p(x_t, s_t|Y_{t-1}, x_{t-1}, s_{t-1}; \theta, Q)$$

The posterior density is equal to the likelihood times the prior:

$$p(\theta, Q|Y_T) \propto L(Y_T|\theta, Q) * p(\theta, Q)$$

The empirical posterior density is obtained using the Kalman filter combined with the Gibbs sampler. In practice the number of parameters is too large for simultaneous estimation of θ and Q. The procedure tends to be very time-consuming, and the Hessian - very inaccurate. A block optimization algorithm can be used instead, with optimizing over (subblocks of) θ and Q in turns. In this case $p(s_T|Y_T, x_T, \theta, Q)$ can be evaluated recursively, $p(\theta|Y_T, x_T, s_T, Q)$ is obtained using some form of Metropolis-Hastings or Maximum Likelihood, while $p(Q|Y_T, x_T, s_T, \theta)$ is obtained sampling from a Dirichlet distribution using a Dirichlet prior.

If the model is more complicated and the transition equation coefficients depend both on s_t and s_{t-1} :

$$x_t = c(s_t, s_{t-1}, \theta) + F(s_t, s_{t-1}, \theta) x_{t-1} + G(s_t, s_{t-1}, \theta) \varepsilon_t$$

then the generalized state can be defined as $\tilde{s}_t = (s_t, s_{t-1})$.

For example let s_t only take two values $\{1, 2\}$. Then Q is a 2x2 matrix:

$$Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}$$

Then the generalized state $\tilde{s}_t$ takes four values: $(1,1), (1,2), (2,1), (2,2)$. The corresponding generalized transition matrix is:

$$\tilde{Q} = \begin{bmatrix} q_{11} & q_{11} & 0 & 0 \\ 0 & 0 & q_{12} & q_{12} \\ q_{21} & q_{21} & 0 & 0 \\ 0 & 0 & q_{22} & q_{22} \end{bmatrix}$$

If several parameters change regimes, then each has a transition matrix Q_i. The transition matrix for the generalized state is then equal to $Q = Q_1 \otimes ... \otimes Q_l$. The tensor product representation of Q implies that if $i = (i_1, ..., i_l)$ and $j = (j_1, ..., j_l)$ then $q_{ij} = \Pi_{k=1}^{l} q_{i_k, j_k}^{k}$. Therefore, the linear restriction on the column of Q is of the form:

$$q_j = M_j b_j, \qquad \text{where} \qquad b_{ij} \geq 0, \qquad \Sigma_i b_{ij} = 1.$$

Any linear set of restrictions on Q one can imagine can be expressed in terms of a sequence of M matrices of this form.

Computing the Marginal Data Density

Marginal data density measures the fit of the model and is the main means of model evaluation when the Bayesian approach is used. To compute the marginal data density, defined as

$$\pi\left(Y_T\right) = \int \pi\left(Y_T|\theta\right)\pi\left(\theta\right)d\theta,$$

harmonic means are typically used. Let $h\left(\theta\right)$ be a p.d.f. of some known distribution and denote:

$$m\left(\theta\right) = \frac{h(\theta)}{\pi(Y_T|\theta)\pi(\theta)}$$

Then the marginal data density can be approximated using the output of the Metropolis-Hastings algorithm:

$$\hat{\pi}\left(Y_T\right) = \frac{1}{N}\Sigma_{i=1}^{N} m\left(\theta^{(i)}\right)$$

Simple harmonic mean is computed using a function with a uniform p.d.f (i.e. $h\left(\theta\right) = const$).

A modified harmonic mean uses a truncated normal p.d.f. of the form:

$$\Theta_{N,p} = \left\{\theta : \left(\theta - \bar{\theta}\right)' \bar{\Omega}_N^{-1}\left(\theta - \bar{\theta}\right) < \chi_p^2\left(n\right)\right\} \text{ and } p \in (0,1)$$
$$h\left(\theta\right) = \frac{\chi\left(\theta \in \Theta_{N,p}\right)}{p} \frac{\left|\bar{\Omega}_N\right|^{-1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\left(\theta - \bar{\theta}\right)' \bar{\Omega}_N^{-1}\left(\theta - \bar{\theta}\right)\right]$$

A more robust "New" modified harmonic mean, proposed by Zha et.al.(2007) uses an elliptical distribution around the mode of the posterior density (let hat in $\hat{\theta}$ and $\hat{\Omega}$ represent the mode):

$\hat{\Omega} = \frac{1}{N}\Sigma_{i=1}^{N}\left(\theta^{(i)} - \hat{\theta}\right)\left(\theta^{(i)} - \hat{\theta}\right)'$

An elliptical distribution centered at $\hat{\theta}$ and scaled by $\hat{S} = \hat{\Omega}^{1/2}$ has the form:

$$g\left(\theta\right) = \frac{\Gamma(k/2)}{2\pi^{k/2}\left|\hat{S}\right|} \frac{f(r)}{r^{k-1}},$$

where $f\left(r\right) = \frac{vr^{v-1}}{b^v - a^v} \geq 0$ on $r \in \left[a, b\right]$, $r = \sqrt{\left(\theta^{(i)} - \hat{\theta}\right)' \hat{\Omega}^{-1}\left(\theta^{(i)} - \hat{\theta}\right)}$ and k is the dimension of θ.

A random variable with an elliptical distribution can be obtained by drawing x from a k-dimensional gaussian distribution and then transforming it as follows:

$$\theta = \frac{r}{\|x\|}\hat{S}x + \hat{\theta}$$

The values of v, a and b can be obtained using the 10^{th} and 90^{th} percentiles of the empirical distribution of posterior draws and the following formulas:

$$v = \frac{\log(1/9)}{\log(c_{10}/c_{90})}, \qquad b = \frac{c_{90}}{0.9^{1/v}}, \qquad a = c_{10}.$$

Now to get the constant of proportionality of the truncated distribution right we need to renormalize:

$$\Theta_U = \{\theta : m(\theta) < U\}$$
$$h(\theta) = \frac{\chi(\theta \in \Theta_U)}{q_U} g(\theta)$$

A simpler and more efficient method was proposed by Mueller. If $g(\theta)$ is the target kernel, and $g(\theta) = c^* g^*(\theta)$ then the goal is to find an accurate estimate of c^*. Define a function:

$$f(c) = E_h\left[1 - \frac{cg^*(\theta)}{h(\theta)}\right]_+ - E_g\left[1 - \frac{h(\theta)}{cg^*(\theta)}\right]_+$$

This function is monotonically decreasing and hence the equation $f(c) = 0$ has a unique solution, which could be found by bisection. However the proposed method is not more accurate then the previous one, because the funciton $f(c)$ is typically very flat in the vicinity of the solution.