Bayesian Methodology

Likelihood and Bayesian Estimation

Joe Pearlman
(based in part on notes by Vasco Gabriel and Bo Yang)

City University

April, 2013

Bayesian approach

- Bayesian analysis is based on a few simples rules of probability
- A, B random variables, then

$$p(A, B) = p(A|B)p(B)$$

  - $p(A, B)$: joint probability of A and B
  - $p(A|B)$: conditional probability of A given B
  - $p(B)$: marginal probability of B
- Reversing the roles, we also have $p(A, B) = p(B|A)p(A)$
- Equating these expressions and rearranging, we get Bayes' rule:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)} = \frac{p(A|B)p(B)}{\sum_{all\ B_i} p(A|B_i)p(B_i)}$$

Bayesian approach

- We want to use data (say, y) to learn about the model's parameters (say, θ)
- A Bayesian approach allow us to to just that: replacing B by θ and A by y

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

- Our focus is on $p(\theta|y)$: given the data, what can we tell about θ?
- Main difference: classical (frequentist) econometrics treats θ as some unknown fixed value(s), whereas Bayesian econometrics assumes that, if θ is unknown, then it should be expressed using rules of probability (i.e., θ is effectively a random object)
- Noting that we're interested in θ, we can drop $p(y)$, so

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Bayesian approach

- $p(\theta|y)$: posterior density
  - summarises what we know about θ after (hence *posterior*) seeing the data
- $p(y|\theta)$: likelihood function, the data density given the model parameters - also denoted as $L(y|\theta)$
- $p(\theta)$: prior density
  - contains all relevant information about θ that does not depend on the data, i.e. what we know about θ *prior* to seeing the data
- $p(\theta|y) \propto p(y|\theta)p(\theta)$ is like an updating rule: the data allow us to update our priors about θ, resulting in the posterior, which combines data and non-data information

Computation

- Bayesian estimation does become complex for computational reasons (less of an issue nowadays)
- We are often interested in functions of $p(\theta|y)$ that summarise what we know about θ, such as (posterior) means, medians, modes, etc (and respective standard deviations)
- In general, this can be expressed as $E[g(\theta)|y]$, where $g(\theta)$ is a function of interest:

$$E[g(\theta)|y] = \int g(\theta)p(\theta|y)d\theta$$

- Bar a few exceptions, it is often impossible to to evaluate the integral analytically $\Rightarrow$ simulation methods (Monte Carlo), drawing from the posterior density $p(\theta|y)$
- As the number of draws (N) increases, then we can invoke the Law of Large Numbers and the Central Limit Theorem

Bayesian Maximum Likelihood

- The most conventional approach to estimation is to maximize (in this case, the Bayesian) likelihood $L(y|\theta)$
- If we have a dataset of time series data $y^T = \{y_1, y_2, ..., y_T\}$, then using Bayes Theorem it is straightforward to show that

$$L(y^{t+1}|\theta) = L(y_{t+1}|y^t, \theta)L(y^t|\theta)$$

so that by induction we have

$$L(y^t|\theta) = \prod_{k=2}^{t} L(y_k|y^{k-1}, \theta)L(y_1|\theta)p(\theta)$$

Calculation of the Likelihood Function

For linear models, with normally distributed shocks, the model can in principle be solved for given θ along the saddle path, and written in state space form as

$$x_{k+1} = Ax_k + B\varepsilon_{k+1} \qquad y_k = Cx_k$$

The log-likelihood is then given by

$$lnL(y|\theta) = -\frac{Tr}{2}ln(2\pi) - \frac{1}{2}\sum_{k=1}^{T}(det(F_k) + e_k^T F_k^{-1} e_k) + lnp(\theta)$$

where r is the number of measurements at each period, and e_k, F_k are obtained from the Kalman Filter recursions

$$\begin{aligned} e_k &= y_k - Cx_{k,k-1} \\ F_k &= CP_kC^T \\ x_{k+1,k} &= Ax_{k,k-1} + AP_kC^T F_k^{-1} e_k \\ P_{k+1} &= AP_kA^T - AP_kC^T F_k^{-1} CP_kA^T + Bcov(\varepsilon)B^T \end{aligned}$$

subject to the initial conditions $x_{1,0} = 0$, and P_1 being the solution of the Lyapunov equation $P_1 = AP_1A^T + Bcov(\varepsilon)B^T$.

First Stage Estimation - Maximizing the Likelihood

- The first thing done by Dynare in the estimation stage is to maximize the Bayesian likelihood
- This yields the ML estimates, with parameter standard errors obtained from the information matrix I_N, which corresponds to the Cramer-Rao lower bound
- For a given model M_i, we can write the Bayesian likelihood as $L(y^T|\theta, M_i)$, and the marginal likelihood of model M_i is given by $\int L(y^T|\theta, M_i)d\theta$
- Different models M_i may have some parameters fixed at 0, or estimated under different information sets. The econometrician will prefer the model with lowest marginal likelihood
- The Laplace approximation to the log marginal likelihood is given by $\frac{N}{2}ln(2\pi) + lnL(y^T|\theta^*, M_i) - \frac{1}{2}ln(det(I_T))$ where I_T is the information matrix evaluated at the maximum θ^*.

Problems with Maximizing the Likelihood

- For complex models, with nonlinear effects of parameters, finding the mode is not straightforward
- The main problem is that the algorithm may have converged to a local maximum of the likelihood
- Even changing the initial parameter values is not an assured method of hitting a global maximum
- Instead it is useful to sample the likelihood function over a large range of parameter draws
- The objective when performing this sampling is to ensure that the frequency of sampling a draw should exactly match the probability of that draw
- The most commonly used method is the MCMC algorithm

MCMC Metropolis-Hastings algorithm

Markov Chain Monte Carlo (MCMC) methods

- MCMC methods: samplers wandering over the posterior, taking most draws from high probability areas
- "Markov Chain" bit: a given draw θ^* depends on $\theta^{(s-1)}$
- "Monte Carlo" bit: θ^* is drawn at random from a candidate proposal (or transition) distribution $\alpha(\theta^{(s-1)}, \theta^*)$, and then (see below) $\theta^{(s)}$ is either θ^* or $\theta^{(s-1)}$
- From any starting draw, θ_0, the frequency distribution of the sequence $\{\theta^{(s)}\}$ will hopefully match the posterior distribution
- Usually one discards the first several thousand draws to ensure that the sequence is not dependent on the starting draw

MCMC Metropolis-Hastings algorithm - cont.

Metropolis-Hastings (MH) algorithm

- Intutition: we want to sample from the region with highest posterior probability, but we also want to visit the whole parameter space as much as possible
- given that there is a discrepancy between the candidate and target densities, the MCMC will not take the correct draws $\Rightarrow$ MH algorithm corrects this calculating an acceptance probability and eventually discarding some draws
- Because it is difficult to find a good candidate density, we usually employ a Random Walk Chain MH algorithm

$$\theta^* = \theta^{(s-1)} + z$$

- sampler wanders in random directions, thus visiting most of the parameter space
- z is usually multivariate Normal, key choice is its covariance matrix

MCMC Metropolis-Hastings algorithm - cont.

- For each draw i, $\widehat{\theta_i} = \theta_{i-1}$ with probability $1 - r$;
- $\widehat{\theta_i} = \theta_i^*$ with probability r.
- The acceptance probability of each new draw is defined by:

$$r = min\left[\frac{\alpha(\theta^{(s-1)}, \theta^*)L(y|\theta_i^*)}{\alpha(\theta^*, \theta^{(s-1)})L(y|\theta_{i-1})}, 1\right]$$

  - Thus if the first term in the above expression is > 1, then set $\theta^{(s)} = \theta^*$.
  - If it is smaller, and is equal to $r < 1$, select a number u at random from the uniform $U[0, 1]$ distribution; if $u > r$ then $\theta^{(s)} = \theta^{(s-1)}$, otherwise $\theta^{(s)} = \theta^*$
- The acceptance rate is dependent on $\alpha \Rightarrow$ one chooses α to obtain 'reasonable' acceptance rate by adjusting the covariance matrix of z
  - If $\alpha(\theta^*, \theta_{i-1}) \sim N(\theta_{i-1}, cV)$ then adjust the 'scale' c
  - Ideally acceptance rate is 20-40% $\Rightarrow$ each move goes a reasonable distance in parameter space, but is not rejected too frequently

A Simple Example of MCMC MH

- θ may take one of three values θ^A, θ^B, θ^C
- $r =$ number of 'heads' from tossing a biased coin n times, so that $L(r|\theta) =_n C_r \theta^r (1-\theta)^{n-r}$. Posterior distribution is just three probabilities each proportional to $L(r|\theta^j)$ for $j = A, B, C$.
- Assume $\alpha(\theta_i^*|\theta_{i-1}) = 1/2$ i.e. there is an equal probability of jumping to either one of the other θ values from θ_{i-1}
- $L(r|\theta^A), L(r|\theta^B), L(r|\theta^C)$ are in the ratio $\alpha : \beta : 1-\alpha-\beta$ where $\alpha < \beta < 1-\alpha-\beta$. What we will show is that the frequency of the draws of the $\{\theta^j\}$ has the same distribution.
- If $\theta_{i-1} = \theta^A$, then since $L(r|\theta^A)$ is the lowest, we will always move so that θ_i will be one of the other θ^js, and since we sample them with equal probability, the probability that θ_i is one or the other is $1/2$.

A Simple Example of MCMC MH (cont)

- If $\theta_{i-1} = \theta^B$, then the probability that $\theta_i = \theta^C$ is $1/2$; however since $\alpha/\beta < 1$, the probability that $\theta_i = \theta^A$ is only $\frac{\alpha}{2\beta}$.
- Similar considerations arise when $\theta_{i-1} = \theta^C$
- The Markov chain for the transitions between these 3 choices of θ^j is therefore

$$P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{\alpha}{2\beta} & \frac{1}{2} - \frac{\alpha}{2\beta} & \frac{1}{2} \\ \frac{\alpha}{2(1-\alpha-\beta)} & \frac{\beta}{2(1-\alpha-\beta)} & 1 - \frac{\alpha+\beta}{2(1-\alpha-\beta)} \end{bmatrix}$$

- It is easy to show that the steady state distribution of this Markov chain, which satisfies $\pi^T P = \pi^T$ is given by $\pi^T = [\alpha \quad \beta \quad 1-\alpha-\beta]$.

Priors

- For Bayesian estimation we need parameter 'priors' (location) and their distributions (shape)
- Where do we get the priors from? Micro estimates, calibration, existing studies...
- Typically the prior mean is centered around calibrated value. Std. errors reflect subjective or objective (to cover the range of existing estimates)
- The shape of the distribution
- General guidance - *inverse gamma* distributions are used as priors when non-negativity constraints are necessary, *beta* distributions for fractions or probabilities, *normal* distributions are used when more informative priors seem to be necessary (*uniform* or 'flat' priors if there is little information about the parameter)
- Options in Dynare are normal, gamma, beta, inverse gamma and uniform distribution

A summary of Bayesian estimation procedures in Dynare

- transform the actual data to fit properties of the model (not in Dynare)
- specify prior distributions
- Dynare computes the log-likelihood numerically via the Kalman filter
- finds the maximum of the likelihood and posterior mode
- draws posterior sequences and simulates posterior distribution with Metropolis algorithm
- computes various statistics on the basis of the posterior distribution (post. moments)
- estimates the posterior marginal density (or likelihood) to compare models
- examine sensitivity of the results to choice of priors

Testing for MCMC Convergence

- Dynare utilises some indicative statistics, summarised by diagrams, as recommended by Brooks and Gelman (1998). These are made up of
  - 3 multivariate figures, representing convergence indicators for all parameters considered together
  - 3 figures for each parameter, representing univariate convergence indicators
- Basic univariate test motivated by ANOVA considerations. Generate m MCMC chains, each run for $2n$ iterations; first n are discarded to avoid burn-in period. Let ψ represent one of the parameters, with ψ_{jk}, $j = 1, ..., m$, $k = 1, ..., n$, representing the draws. If the ψ_{jk} were normally distributed with variance σ^2, then an unbiased estimator $\hat{\sigma}^2$ of σ^2 is given by

$$(mn-1)\hat{\sigma}^2 = \sum_{j=1}^{m}\sum_{k=1}^{n}(\psi_{jk}-\psi_{..})^2 \equiv \sum_{j=1}^{m}\sum_{k=1}^{n}(\psi_{jk}-\psi_{j.})^2 + n\sum_{j=1}^{m}(\psi_{j.}-\psi_{..})^2$$

where $\psi_{j.}$ represents the mean for the jth chain, and $\psi_{..}$ is the mean over all chains

Testing for MCMC Convergence (cont)

- One measure of convergence is that the $\psi_{j.}$ are all equal to $\psi_{..}$ i.e. that the initial value of the draw in each chain has not affected the mean. Another test is whether the variance is equal across all the chains.
- We can test these together by checking whether the *Potential Scale Reduction Factor* $R_2 \equiv V/W$ is approaching 1, where

$$V = \frac{1}{mn-1}\sum_{j=1}^{m}\sum_{k=1}^{n}(\psi_{jk}-\psi_{..})^2 \qquad W = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{k=1}^{n}(\psi_{jk}-\psi_{j.})^2$$

- Brook and Gelman recommend that V and W are plotted sequentially for $k = 1, ..., n$; this means that one can check that as n increases, V and W tend individually to a limit, and that this is the same limit as k approaches n.
  - If the posterior distribution is unimodal, this is essentially a check that both means and variances of all chains' estimates of ψ are tending to the same limit.
  - If the posterior distribution is not unimodal, then it makes sense to extend this to other moments, and Dynare does a similar calculation for third moments as well.

Testing for MCMC Convergence - Interval Measures

- Based on the intuitive notion that R_2 also represents a squared ratio of the proportion of draws within a certain confidence interval. To perform this explicitly, Brook and Gelman suggest a measure $R_{interval}$ that uses, as before, the last n of the $2n$ draws of each chain, and then
- From each chain find the empirical 100(1-α)% interval i.e. the number of draws within the empirical $100\frac{\alpha}{2}$% and $100(1-\frac{\alpha}{2})$% points; Dynare sets $\alpha = 0.2$.
- Do the same for all the mn draws from all the m chains
- Evaluate $R_{interval} \equiv V_{interval}/W_{interval}$ where $V_{interval}$=length of total-sequence interval, $W_{interval}$=mean length of within-sequence intervals. As before, it is insightful to plot both $V_{interval}$ and $W_{interval}$.

Testing for MCMC Convergence - Multivariate Measures

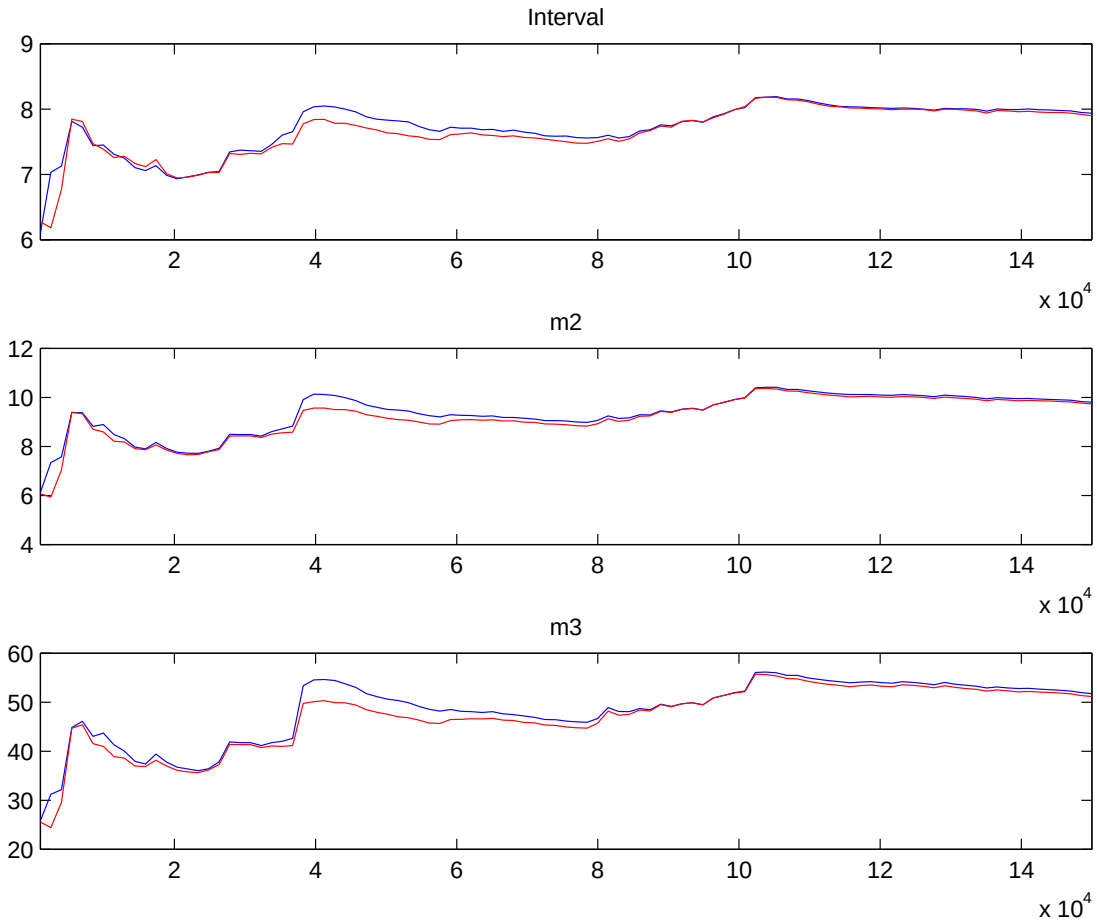
- An unbiased estimate $\hat{\Omega}$ of the covariance matrix of the vector of parameters θ is

$$
\begin{aligned}
(mn-1)\hat{\Omega} &= \sum_{j=1}^{m}\sum_{k=1}^{n}(\theta_{jk}-\theta_{..})(\theta_{jk}-\theta_{..})^T \\
&\equiv \sum_{j=1}^{m}\sum_{k=1}^{n}(\theta_{jk}-\theta_{j.})(\theta_{jk}-\theta_{j.})^T + n\sum_{j=1}^{m}(\theta_{j.}-\theta_{..})(\theta_{j.}-\theta_{..})^T
\end{aligned}
$$

  Matrices V and W are then defined analogously to their scalar versions above. One measure closeness is the maximum root statistic - the solution to $max_a(a^T Va)/a^T Wa)$, which is given by the largest eigenvalue of $W^{-\frac{1}{2}}VW^{-\frac{1}{2}}$, which should tend to 1 if the chains are converging to the posterior distribution. The determinants of V and W should also converge.
- A similar approach is taken for third moments
- Interval measure: count the number of draws for which each of the elements θ_i of the vector θ lie within their individual empirical 100(1-α)% intervals. Find the average, and compare with the whole sample taken together.

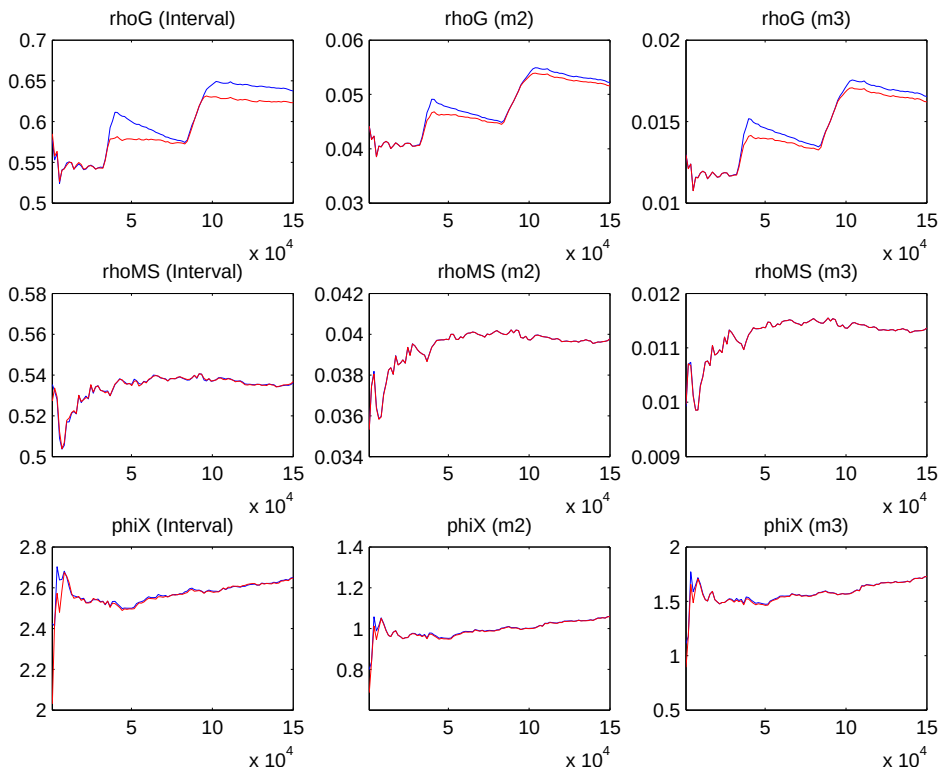
Example of Convergence - Multivariate Measures

As an example, consider the diagnostics for the NK model estimated earlier. The multivariate diagnostics shown below indicate that the chains have converged to similar means and distributions - *Interval* refers to the interval measure, and *m2, m3* refer to second and third order multivariate moment measures.



Example of Convergence - Univariate Measures

For individual parameters, the results are mixed. Only for $\rho_A, \rho_{MS}, \phi_X, \sigma_c, h_c, \gamma_p, \alpha_\pi, \alpha_y$ is there clear evidence of convergence in mean and distribution. Whereas the posterior distribution of ρ_{MS} and ϕ_X is similar for each of the chains, and their interval and moment diagnostics appear to be converging, it is evident this is not the case for ρ_G.



What to do about Lack of Convergence

- Convergence is a notorious problem for MCMC, and the only theorem is that if convergence occurs, it is to the correct distribution.
- Crucially, one would want multivariate convergence
- Improving convergence could be done in one of two ways:
  - Increase the number of draws
  - Increase the 'scale factor' for the Monte Carlo part. This increases the range of search but at the expense of reducing the acceptance ratio

Identification issues in DGSE models

- See Canova and Sala (2009), Komunjer and Ng (2009) and Iskrev (2010)
- DSGE model as a likelihood function $L(y|\theta)$
- Given data y (and the model), what is the most plausible θ (technology, preferences, shocks)?
- θ_1 is identified if $L(y|\theta_1) = L(y|\theta_2)$, for all $y \Rightarrow \theta_1 = \theta_2$
- $\Rightarrow$ no other $\theta_1 \neq \theta_2$ is observationally equivalent to θ_1
- DSGE models: reduced form solution $Z_t = A(\theta)Z_{t-1} + B(\theta)U_t$ for endogenous variables Z and structural shocks U
- Problem: mapping from structural parameters into the above law of motion for Z is highly nonlinear

Identification issues in DGSE models

- Canova and Sala (2009): focus on methods that match model and empirical impulse responses
- distinction between observational equivalence (2 structural models generate the same IRFs), under-identification (structural parameters may disappear after log-lin), weak identification (insufficient curvature of LL) and partial identification (parameters cannot be recovered separately) $\Rightarrow$ shape and rank of the information matrix
- poor identification leads to serious biases; calibration may induce distortions in the distribution of parameter estimates
- main recommendations: plot the objective function fixing parameters in turn; check the rank of the Hessian, using appropriate tests (Cragg and Donald, 1997, Kleibergen and Paap, 2005); work separately with portions/equations of the model to understand sources of identification failures

Identification issues in DGSE models

- Iskrev (2010): Hessian can be decomposed into two terms: the gradient of the mapping between reduced-form and structural parameters + the information matrix of the reduced-from model
- τ: vector collecting all the reduced-form coefficients (the elements in A, $\Omega = BB'$ and the steady-state of Z_t that depend on θ).
- $\mathbf{m_T} := [\mu', \sigma'_T]$: vector collecting the first and second order moments (which includes all covariances and auto-cross-correlations up to $T-1$) of the observable variables
- Local identification can be verified by means of a rank condition of the Jacobian matrix $J_T = \frac{\delta m_T}{\delta \theta'} = \frac{\delta m_T}{\delta \tau'} \frac{\delta \tau'}{\delta \theta'}$
- It follows that a necessary condition for local identifiability of θ is that the rank of $H = \frac{\delta \tau'}{\delta \theta'}$ evaluated at θ is equal to the dimension of θ
- The H term is independent of the data $\rightarrow$ it is possible to detect identification problems that are inherent to the structure of the DSGE model, *before* taking the model to the data!

Identification issues in DGSE models

- This suggests a Monte Carlo approach: draw θ from Θ (ensuring stability and determinacy), compute rank and condition number of HH' and J, and repeat this many times
- if H is rank-deficient at θ_j, this particular point is unidentifiable
- if H has full rank but J_T does not, then θ_j cannot be identified for the particular set of observables and contemporaneous and lagged moments under consideration, i.e. given y^T and T

Identification toolbox in Dynare

- Dynare implementation: use $\mathrm{identification}(< \textit{options} >=< \textit{values} >)$
- options on the number of MC draws, using a previous MC sample, etc. Usually, defaults are OK.
- This can optionally be run without estimating the model. After the MC run, outputs concerning the ranks of H and J will appear
- Strength of identification: weak identification may arise because the moments in the data change little for a particular θ_i ('sensitivity') or because collinearity dampens the effect of θ_i ('correlation'). For parameter θ_i it is given by $\sqrt{\theta_i^2/(I_T^{-1})_{ii}}$ (logged) where I_T is the information matrix from the maximum likelihood estimation
- Sensitivity for parameter θ_i is $\sqrt{\theta_i^2(I_T)_{i,i}}$ (logged as well) - represents the curvature of the 'marginal' plot of the likelihood vs the parameter ignoring correlation with other parameters
- Note that for one parameter only, strength=sensitivity

Identification toolbox in Dynare