Bayesian Methods for DSGE models

Fabio Canova
EUI and CEPR
January 2014

Outline

- Bayes Theorem.
- Prior Selection.
- Posterior Simulators.
- Robustness.
- Estimation of DSGE.

References

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer and Verlag.

Bauwens, L., M. Lubrano and J.F. Richard (1999) *Bayesian Inference in Dynamics Econometric Models*, Oxford University Press.

Bi, X. and Traum, N. (2013) Estimating Fiscal limits: the case of Greece, forthcoming, Journal of Applied Econometrics.

Canova, F. (2010) Bridging DSGE models and the raw data, manuscript.

Canova, F. and Sala, L. (2009) Back to square one, identification issues in DSGE models, Journal of Monetary Economics.

Canova, F., F. Ferroni and C. Matthes (2013) Choosing the variables to estimate DSGE models, forthcoming, Journal of Applied Economics.

Cahri, V., Kehoe, P. and McGrattan, E. (2009) New Keynesian models, not yet useful for policy analysis, AEJ: Macroeconomics,1, 242-266.

Faust, J. and Gupta, A. (2012) Posterior Predictive Analysis for Evaluating DSGE Models, NBER working paper 17906.

Gelman, A., J. B. Carlin, H.S. Stern and D.B. Rubin (1995), *Bayesian Data Analysis*, Chapman and Hall, London.

Herbst, E. and Schorfheide, F. (2013) Sequential Monte Carlo Sampling for DSGE models, forthcoming, Journal of Applied Econometrics.

Iskrev, N. (2010), Local identification of DSGE models Journal of Monetary Economics, 57, 189-202

Komunjer, I. and Ng, S. (2012), Dynamic Identification of DSGE models, Econometrica, 79, 1995-2032.

Poirier, D. (1995) Intermediate Statistics and Econometrics, MIT Press.

Qu, and D. Thachenko, (2012) Identification and frequency domain ML estimation of linearized dynamic stochastic general equilibrium models, Quantitative Economics, 3, 95-132

Casella, G. and George, E. (1992) Explaining the Gibbs Sampler *American Statistician*, 46, 167-174.

Chib, S. and Greenberg, E. (1995) Understanding the Hasting-Metropolis Algorithm, *The American Statistician*, 49, 327-335.

Chib, S. and Greenberg, E. (1996) Markov chain Monte Carlo Simulation methods in Econometrics, *Econometric Theory*, 12, 409-431.

Geweke, J. (1995) Monte Carlo Simulation and Numerical Integration in Amman, H., Kendrick, D. and Rust, J. (eds.) *Handbook of Computational Economics* Amsterdaam, North Holland, 731-800.

Kass, R. and Raftery, A (1995), Empirical Bayes Factors, *Journal of the American Statistical Association*, 90, 773-795.

Mueller, U. (2012): Measuring prior sensitivity and prior informativeness in Large Bayesian models, Journal of Monetary Economics, 59, 581-597.

Schmitt- Grohe, S. and Uribe, M. (2012) What is news in business cycles?, Econometrica, 80, 2733-2764.

Smets, F. and Wouters, R. (2007) Shocks and Frictions in US Business cycles, American Economic Review, 97, 586-606.

Smets, F. and Wouters, R. (2003) An estimated DSGE model of the Euro area, Journal of the European Economic Association,1, 1123-1175.

Tierney, L (1994) Markov Chains for Exploring Posterior Distributions (with discussion), *Annals of Statistics*, 22, 1701-1762.

1 Preliminaries

Classical and Bayesian analysis differ on a number of issues

Classical analysis:

• Probabilities = limit of the relative frequency of the event.

• Parameters are fixed, unknown quantities.

• Unbiased estimators useful because average value of sample estimator converge to true value via some LLN. Efficient estimators preferable because they yield values closer to true parameter.

• Estimators and tests are evaluated in repeated samples (to give correct result with high probability).

Bayesian analysis:

• Probabilities = degree of (typically subjective) beliefs of a researcher in an event.

• Parameters are random variables with a probability distributions.

• Properties of estimators and tests in repeated samples uninteresting: beliefs not necessarily related to relative frequency of an event in large number of hypothetical experiments.

• Estimators are chosen to minimize expected loss functions (expectations taken with respect to the posterior distribution), conditional on the data. Use of probability to quantify uncertainty.

In large samples (under appropriate regularity conditions):

- Posterior mode $\alpha^* \xrightarrow{P} \alpha_0$ (Consistency)

- Posterior distribution converges to a normal with mean α_0 and variance $(T \times I(\alpha_0))^{-1}$, where $I(\alpha)$ is Fisher's information matrix (Asymptotic normality).

Classical and Bayesian analyses differ in small samples and for dealing with unit root processes.

Bayesian analysis requires:

- Initial information $\longrightarrow$ Prior distribution.
- Data $\longrightarrow$ Likelihood.
- Prior and Likelihood $\longrightarrow$ Bayes theorem $\longrightarrow$ Posterior distribution.
- Can proceed recursively (mimic economic learning).

2 Bayes Theorem

Parameters of interest $\alpha \in A$, A compact. Prior information $g(\alpha)$. Sample information $f(y|\alpha) \equiv \mathcal{L}(\alpha|y)$.

- Bayes Theorem.

$$g(\alpha|y) = \frac{f(y|\alpha)g(\alpha)}{f(y)} \propto f(y|\alpha)g(\alpha) = \mathcal{L}(\alpha|y)g(\alpha) \equiv \grave{g}(\alpha|y)$$

$f(y) = \int f(y|\alpha)g(\alpha)d\alpha$ is the unconditional sample density (Marginal likelihood), and it is constant from the point of view of $g(\alpha|y)$; $g(\alpha|y)$ is the posterior density, $\grave{g}(\alpha|y)$ is the posterior kernel, $g(\alpha|y) = \frac{\grave{g}(\alpha|y)}{\int \grave{g}(\alpha|y)d\alpha}$.

- $f(y)$ it is a measure of fit. It tells us how good the model is in reproducing the data, on average over the parameter space.

• α are regression coefficients, structural parameters, etc.; $g(\alpha|y)$ is the conditional probability of α, given what we observe, y.

• Theorem uses rule: $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$. It tells us how to modify some prior beliefs about α, once we observe y. It does not tells us what the initial beliefs are.

To use Bayes theorem we need:

a) Formulate prior beliefs, i.e. choose $g(\alpha)$.

b) Formulate a model for the data (the conditional probability of $f(y|\alpha)$).

After observing the data, we treat the model as the likelihood of α conditional on y, and update beliefs about α.

• Bayes theorem with nuisance parameters (e.g. α_1 long run coefficients, α_2 short run coefficients; α_1 regression coefficient; α_2 serial correlation coefficient in the errors).

Let $\alpha = [\alpha_1, \alpha_2]$ and suppose interest is in α_1. Then $g(\alpha_1, \alpha_2|y) \propto f(y|\alpha_1, \alpha_2)g(\alpha_1, \alpha_2)$

$$\begin{aligned} g(\alpha_1|y) &= \int g(\alpha_1, \alpha_2|y)d\alpha_2 \\ &= \int g(\alpha_1|\alpha_2, y)g(\alpha_2|y)d\alpha_2 \end{aligned} \tag{1}$$

Posterior of α_1 averages the conditional of α_1 with weights given by the posterior of α_2.

• Bayes Theorem with two (N) samples.

Suppose $y_t = [y_{1t}, y_{2t}]$ and that y_{1t} is independent of y_{2t}. Then

$$\breve{g} \equiv f(y_1, y_2|\alpha)g(\alpha) = f_2(y_2|\alpha)f_1(y_1|\alpha)g(\alpha) \propto f_2(y_2|\alpha)g(\alpha|y_1) \quad (2)$$

Posterior for α is obtained finding first the posterior of using y_{1t} and then, treating it as a prior, finding the posterior using y_{2t}.

- Sequential learning.

- Can use data from different regimes.

- Can use data from different countries.

2.1 Likelihood Selection

• It should reflect an economic model.

• It must represent well the data. Misspecification problematic since it spills across equations and makes estimates uninterpretable.

• For our purposes the likelihood is simply the theoretical (DSGE) model you write down.

2.2 Prior Selection

• Three methods to choose priors in theory. Two not useful for DSGE models since are designed for models which are linear in the parameters.

1) Non-Informative subjective. Choose **reference priors** because they are invariant to the parametrization.

- Location invariant prior: $g(\alpha) =$constant ($=1$ for convenience).

- Scale invariant prior $g(\sigma) = \sigma^{-1}$.

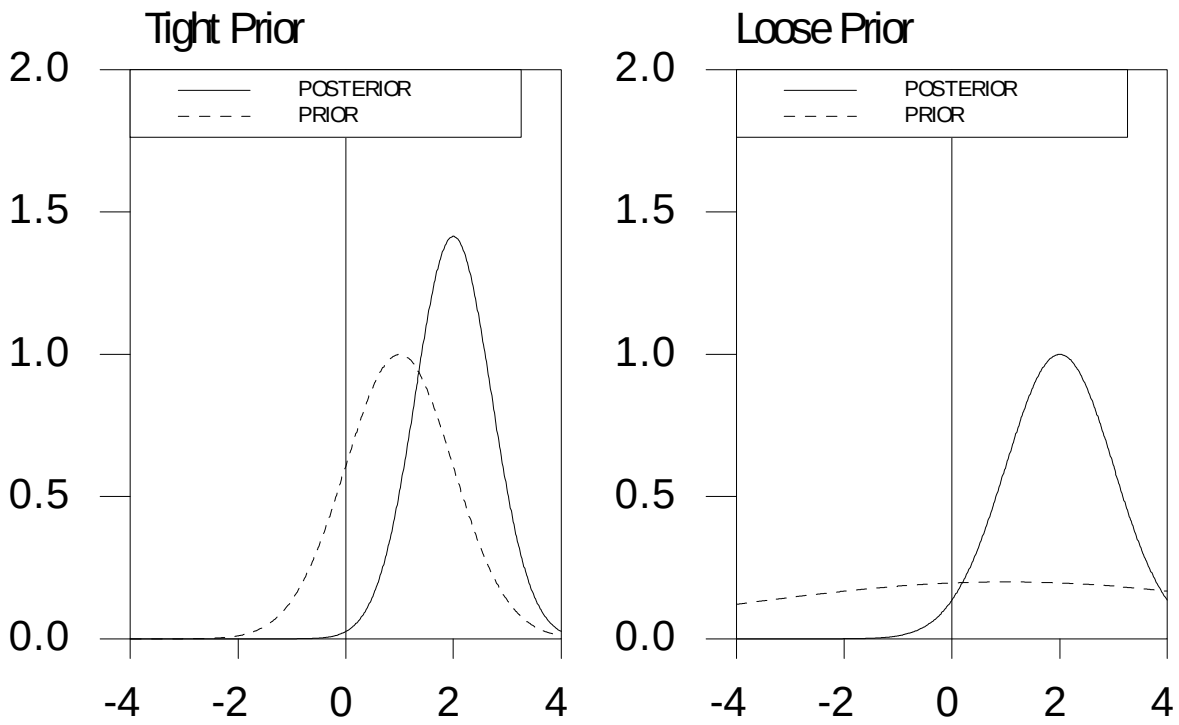
- Location-scale invariant prior : $g(\alpha, \sigma) = \sigma^{-1}$.

• Non-informative priors useful because many classical estimators (OLS, ML) are Bayesian estimators with non-informative priors

2) Conjugate Priors

A prior is conjugate if the posterior has the same form as the prior. Hence, the form posterior will be analytically available, only need to figure out its posterior moments.

• Important result in linear models with conjugate priors: Posterior moments = weighted average of sample and prior information. Weights = relative precision of sample and prior informations.


Tight Prior
Loose Prior
POSTERIOR
PRIOR
2.0
1.5
1.0
0.5
0.0
-4
-2
0
2
4

3) Objective priors and ML-II approach. Based on:

$$f(y) = \int \mathcal{L}(\alpha|y)g(\alpha)d\alpha \equiv \mathcal{L}(y|g) \tag{3}$$

Given $\mathcal{L}(\alpha|y)$, $\mathcal{L}(y|g)$ reflects the plausibility of g in the data.
If g_1 and g_2 are two priors and $\mathcal{L}(y|g_1) > \mathcal{L}(y|g_2)$, better support for g_1.
Hence, can estimate the "best" g using $\mathcal{L}(y|g)$.

In practice, set $g(\alpha) = g(\alpha|\theta)$, where θ= hyperparameters (e.g. the mean and the variance of the prior). Then $\mathcal{L}(y|g) \equiv \mathcal{L}(y|\theta)$.

The θ that maximizes $\mathcal{L}(y|\theta)$ is called ML-II estimator and $g(\alpha|\theta_{ML})$ is ML-II based prior.

Important:

- $y_1, \ldots y_T$ **should not** be the same sample used for inference.

- $y_1, \ldots y_T$ could represent past time series information, cross sectional/ cross country information.

- Typically $y_1, \ldots y_T$ is called "Training sample".

4) Priors for DSGE - similar to MLII priors.

- Assume that $g(\alpha) = g_1(\alpha_1)g_2(\alpha_2)....g_q(\alpha_q)$.

- Use a conventional format for the distributions: a Normal, Beta and Gamma for individual parameters. Choose moments in a data based fashion: mean = calibrated parameters, variance: subjective.

Problems:

• Independent priors typically inconsistent with any subjective prior beliefs over joint outcomes. In particular, multivariate priors are often too tight!!

• Calibrated value may be different for different purposes. For example, risk aversion mean is 6-10 to fit the equity premium; close to 1-2 if we

want to fit the reaction of consumption to changes in monetary policy; negative values to fit aggregate lottery revenues. Which one do we use? Same for habit parameters (see Faust and Gupta, 2012)

• Circularity: priors based on the same data used to estimate!! Use calibrated values in a " training sample".

Negro and Schorfheide (2008): formally choose data based priors in training samples which are not independent.

Summary

Inputs of the analysis: $g(\alpha),\ f(y|\alpha)$.

Outputs of the analysis:

$g(\alpha|y) \propto f(y|\alpha)g(\alpha)$ (posterior),

$f(y) = \int f(y|\alpha)g(\alpha)$ (marginal likelihood), and

$f(y^{T+\tau}|y^T)$ (predictive density of future observations).

Likelihood should reflect data/ economic theory.

Prior could be non-informative, conjugate, data based (objective).

- In simple examples, $f(y)$ and $g(\alpha|y)$ can be computed analytically.

- In general, they can only be computed numerically by Monte Carlo methods.

- If the likelihood is a (log-linearized) DSGE model: always need numerical computations.

3 Posterior simulators

Objects of interest for Bayesian analysis: $E(h(\alpha)) = \int h(\alpha)g(\alpha|y)d\alpha$. Occasionally, can evaluate the integral analytically. In general, it is impossible.

If $g(\alpha|y)$ were available: we could compute $E(h(\alpha))$ with MC methods:

- Draw α^l from $g(\alpha|y)$. Compute $h(\alpha^l)$

- Repeat draw L times. Average $h(\alpha^l)$ over draws.

Example 3.1 *Suppose we are interested in computing* $Pr(\alpha > 0)$. *Draw* α^l *from* $g(\alpha|y)$. *If* $\alpha^l > 0$, *set* $h(\alpha^l) = 1$, *else set* $h(\alpha^l) = 0$. *Draw L times and average* $h(\alpha^l)$ *over draws. The result is an estimate of* $Pr(\alpha > 0)$.

• Approach works because with iid draws the law of large numbers (LLN) insures that sample averages converge to population averages (ergodicity).

• By a central limit theorem (CLT) the difference between sample and population averages has a normal distribution with zero mean and some variance as L grows (numerical standard errors can be used a a measure of accuracy).

- Since $g(\alpha|y)$ is not analytically available, need to use a $g^{AP}(\alpha|y)$, which is similar to $g(\alpha|y)$, and easy to draw from.

• Normal Approximation

• Basic Posterior simulators (Acceptance and Importance sampling).

• Markov Chain Monte Carlo (MCMC) methods

3.1 Normal posterior analysis

If T is large $g(\alpha|y) \approx f(\alpha|y)$. If $f(\alpha|y)$ is unimodal, roughly symmetric, and α^* (the mode) is in the interior of A:

$$\log g(\alpha|y) \approx \log g(\alpha^*|y) + 0.5(\alpha - \alpha^*)'[\frac{\partial^2 \log g(\alpha|y)}{\partial\alpha\partial\alpha'}|_{\alpha=\alpha^*}](\alpha - \alpha^*) \quad (4)$$

Since $g(\alpha^*|y)$ is constant, letting $\Sigma_{\alpha^*} = -[\frac{\partial^2 \log g(\alpha|y)}{\partial\alpha\partial\alpha'}{}^{-1}|_{\alpha=\alpha^*}]$

$$g(\alpha|y) \approx N(\alpha^*, \Sigma_{\alpha^*}) \quad (5)$$

- An approximate 100(1-ρ)% highest credible set is $\alpha^* \pm \Phi(\rho/2)I(\alpha^*)^{-0.5}$ where $\Phi(.)$ the CDF of a standard normal.

• Approximation is valid under regularity conditions when $T \to \infty$ or when the posterior kernel is roughly normal. It is highly inappropriate when:

- Likelihood function flat in some dimension ($I(\alpha^*)$ badly estimated).

- Likelihood function is unbounded (no posterior mode exists).

- Likelihood function has multiple peaks.

- α^* is on the boundary of A (quadratic approximation wrong).

- $g(\alpha) = 0$ in a neighborhood of α^* (quadratic approximation wrong).

How do we construct a normal approximation?

A) Find the mode of the posterior.

$$\max \log g(\alpha|y) = \max(\log L(\alpha|y) + \log g(\alpha))$$

- Problem is identical to the one of finding the maximum of a likelihood. The objective function differs.

Two mode finding algorithms:

i) Newton type of algorithm

- Let $L = \log g(\alpha|y)$ (or $L = \log \breve{g}(\alpha|y)$). Choose α_0.

- Calculate $L' = \frac{\partial L}{\partial \alpha}(\alpha_0)$ $L'' = \frac{\partial^2 L}{\partial \alpha \partial \alpha'}(\alpha_0)$. Approximate L quadratically.

- Set $\alpha^l = \alpha^{l-1} - \gamma(L''(\alpha^{l-1}|y))^{-1}(L'(\alpha^{l-1}|y))$ $\gamma \in (0, 1)$.

- Iterate until convergence i.e. until $||\alpha^l - \alpha^{l-1}|| < \iota$, ι small.

Fast and good if α_0 is good and L close to quadratic. Bad if L'' not positive definite.

ii) Conditional maximization algorithm.

Let $\alpha = (\alpha_1, \alpha_2)$. Start from some $(\alpha_{10}, \alpha_{20})$. Then

- Maximize $L(\alpha_1, \alpha_2)$ with respect to α_1 keeping α_{20} fixed. Let α_1^* the maximizer.

- Maximize $L(\alpha_1, \alpha_2)$ with respect to α_2 keeping $\alpha_1 = \alpha_1^*$ fixed. Let α_2^* the maximizer.

- Iterate on two previous steps until convergence.

- Start from different $(\alpha_{10}, \alpha_{20})$, check if maximum is global.

B) Compute the variance covariance matrix at the mode

- Use the Hessian $\Sigma_{\alpha^*} = -[\frac{\partial^2 \log g(\alpha|y)}{\partial\alpha\partial\alpha'}{}^{-1}|_{\alpha=\alpha^*}]$

C) Approximate the posterior density: $g^{AP}(\alpha|y) = \mathbb{N}(\alpha^*, \Sigma_{\alpha^*})$.

- If multiple modes are present, find an approximation to each mode, and set $g^{AP}(\alpha|y) = \sum_i \varrho_i \mathbb{N}(\alpha_i^*, \Sigma_{\alpha_i^*})$ where $0 \leq \varrho_i \leq 1$. If modes are clearly separated select $\varrho_i = g(\alpha_i^*|y)|\Sigma_{\alpha_i^*}|^{-0.5}$.

- If the sample is small, use a t-approximation i.e. $g^{AP}(\alpha|y) = \sum_i \varrho_i g(\tilde{\alpha}|y)[\nu + (\alpha - \alpha_i^*)'\Sigma_{\alpha_i}(\alpha - \alpha_i^*)]^{-0.5(k+v)}$ with small ν.

(If $\nu = 1$ t-distribution=Cauchy distribution, large overdispersion. Typically $\nu = 4, 5$ appropriate).

D) To conduct inference, draw α^l from $g^{AP}(\alpha|y)$.

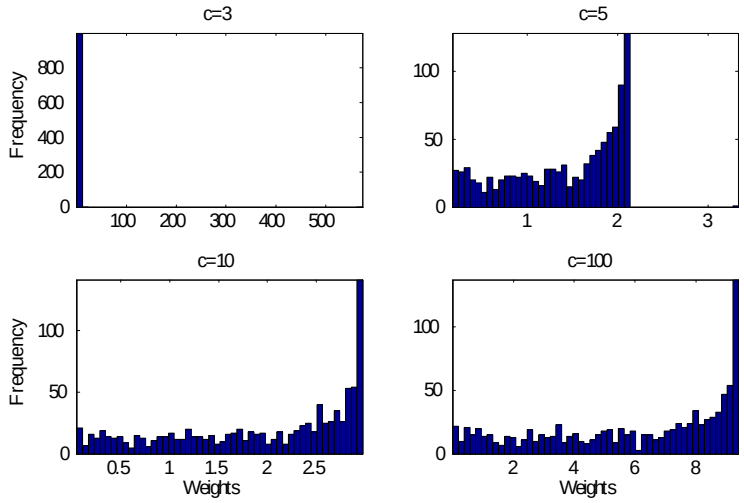
If draws are iid, $E(h(\alpha)) = \frac{1}{L}\sum_l h(\alpha^l)$. Use LLN to approximate any posterior probability contours of $h(\alpha)$, e.g. a 16-84 range is $[h(\alpha^{16}), h(\alpha^{84})]$.

E) Check accuracy of approximation.

Compute *Importance Ratio* $IR^l = \frac{\breve{g}(\alpha^l|y)}{g^{AP}(\alpha^l|y)}$, where $\breve{g}(\alpha^l|y)$ is the kernel of the posterior (which you can always compute). Accuracy is good if IR^l is constant across l. If not, need to use other techniques.

Note: Importance ratios are not automatically computed in Dynare. Need to do it yourself.

Example 3.2 *True:* $g(\alpha|y)$ *is t(0,1,2). Approximation: N(0,c), where* $c = 3, 5, 10, 100$.



Horizontal axis=importance ratio weights, vertical axis= frequency of the weights.

- Posterior has fat tails relative to a normal. Thus, the approximation is poor.

3.2 Basic Posterior Simulators

- Draw from a general $g^{AP}(\alpha|y)$ (not necessarily normal).
- Non-iterative methods - $g^{AP}(\alpha|y)$ is fixed across draws.
- Work well when IR^l is roughly constant across draws.

A) Acceptance sampling

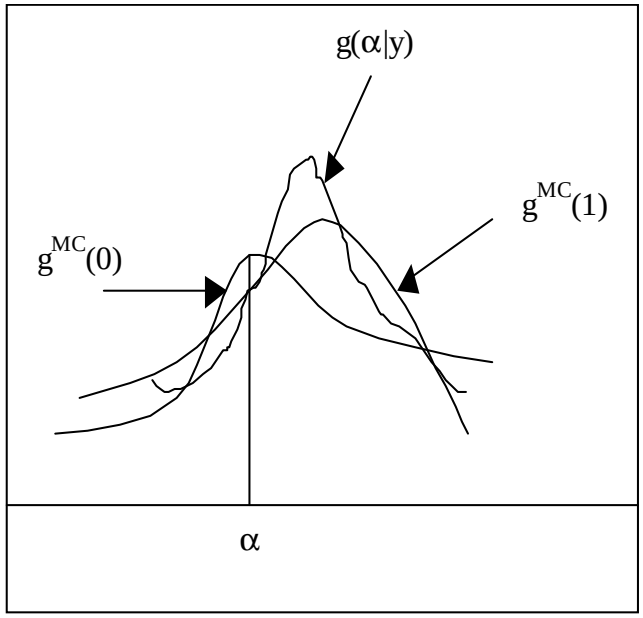
B) Importance sampling

3.3 Markov Chain Monte Carlo Methods

• Problem with basic simulators: approximating density is selected once and for all. If mistakes are made, they stay. With MCMC location of approximating density changes as iterations progress.

• Idea: Suppose n states $(x_1, \ldots x_n)$. Let $P(i,j) = Pr(x_{t+1} = x_j | x_t = x_i)$ and let $\mu(t) = (\mu_{1t}, \ldots \mu_{nt})$ be the unconditional probability at t of each state n. Then $\mu(t+1) = P\mu(t) = P^t\mu(0)$ and μ is an equilibrium (ergodic, steady state, invariant) distribution if $\mu = \mu P$.

Set $\mu = g(\alpha|y)$, choose some initial density $\mu(0)$ and some transition P across states. If conditions are right, iterate from $\mu(0)$ and limiting distribution is $g(\alpha|y)$, the unknown posterior.



• Under general conditions, the ergodicity of P insures consistency and asymptotic normality of estimates of any $h(\alpha)$.

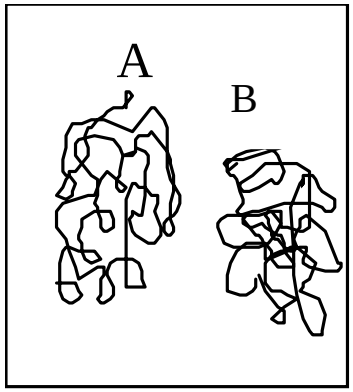
Need a transition $P(\alpha, A)$, where A is some set, such that $||P(\alpha, A) - \mu(\alpha)|| \rightarrow 0$ in the limit. For this need that the chain associated with P :

- is irreducible, i.e. it has no absorbing state.

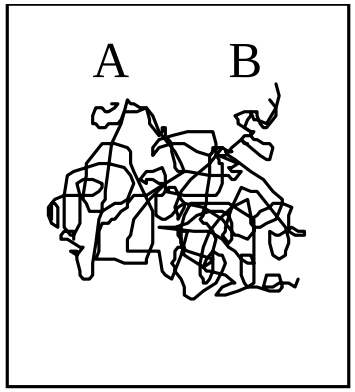
- is aperiodic, i.e. it does not cycle across a finite number of states.

- it is Harris recurrent, i.e. each cell is visited an infinite number of times with probability one.

Bad draws



Good draws



Result 1: A reversible Markov chain, has an ergodic distribution (existence). (if $\mu_i P_{i,j} = \mu_j P_{j,i}$ then $(\mu P)_j = \sum_j \mu_i P_{i,j} = \sum_i \mu_j P_{j,i} = \mu_j \sum_i P_{j,i} = \mu_j$.)

Result 2: (Tierney (1994)) (uniqueness) If a Markov chain is Harris recurrent and has a proper invariant distribution. $\mu(\alpha)$, $\mu(\alpha)$ is unique.

Result 3: (Tierney(1994)) (convergence) If a Markov chain with invariant $\mu(\alpha)$ is Harris recurrent and aperiodic, for all $\alpha_0 \in A$ and all A, as $L \to \infty$.

- $||P^L(\alpha_0, A) - \mu(\alpha)|| \to 0$, $||.||$ is the total variation distance.

- For all $h(\alpha)$ absolutely integrable with respect to $\mu(\alpha)$.

- $\lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} h(\alpha^l) \stackrel{a.s.}{\to} \int h(\alpha)\mu(\alpha)d\alpha$.

If chain has a finite number of states, it is sufficient for the chain to be irreducible, Harris recurrent and aperiodic that $P(\alpha^l \in A_1 | \alpha^{l-1} = \alpha_0, y) > 0$, all $\alpha_0, A_1 \in A$.

- Can dispense with the finite number of state assumption.

- Can dispense with the first order Markov assumption.

General simulation strategy:

- Choose starting values α_0, choose a P with the right properties.
- Run MCMC simulations.
- Check convergence.
- Summarize results i.e compute $h(\alpha)$.

1) MCMC methods generate draws which are *correlated* (with normal/basic simulators, posterior draws are iid).

2) MCMC methods generate draws from posterior only after a burn-in period (with normal/basic simulators, the first draw is from the posterior).

3) MCMC methods only need the kernel $\breve{g}(\alpha|y)$ to be operative (no knowledge of the normalizing constants is needed).

4) MCMC can be used in non-Bayesian contexts to explore intractable likelihoods using "data augmentation" technique.

3.3.1 Metropolis-Hastings algorithm

MH is a general purpose MCMC algorithm that can be used when faster methods (such as the Gibbs sampler) are either not usable or difficult to implement.

Starts from an arbitrary transition function $q(\alpha^{\dagger}, \alpha^{l-1})$, where $\alpha^{l-1},\ \alpha^{\dagger} \in A$ and an arbitrary $\alpha^{0} \in A$. For each $l = 1, 2, \ldots L$.

- Draw $\alpha^{\dagger}$ from $q(\alpha^{\dagger}, \alpha^{l-1})$ and draw $\varpi \sim U(0, 1)$.

- If $\varpi < \mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger}) = [\frac{\breve{g}(\alpha^{\dagger}|Y)q(\alpha^{\dagger}, \alpha^{l-1})}{\breve{g}(\alpha^{l-1}|Y)q(\alpha^{l-1}, \alpha^{\dagger})}]$, set $\alpha^{\ell} = \alpha^{\dagger}$.

- Else set $\alpha^{\ell} = \alpha^{\ell-1}$.

These iterations define a mixture of continuous and discrete transitions:

$$\begin{aligned} P(\alpha^{l-1},\ \alpha^{l}) &= q(\alpha^{l-1},\alpha^{l})\mathfrak{E}(\alpha^{l-1},\alpha^{l}) \qquad \text{if } \alpha^{l} \neq \alpha^{l-1} \\ &= 1 - \int_{A} q(\alpha^{l-1},\alpha)\mathfrak{E}(\alpha^{l-1},\alpha)d\alpha \quad \text{if } \alpha^{l} = \alpha^{l-1} \end{aligned} \tag{6}$$

$P(\alpha^{l-1},\ \alpha^{l})$ satisfies the conditions needed for existence, uniqueness and convergence.

• Idea: Want to sample from highest probability region but want to visit as much as possible the parameter space. How to do it? Choose an initial vector and a candidate, compute kernel of posterior at the two vectors. If you go uphill, keep the draw, otherwise keep the draw with some probability.

If $q(\alpha^{l-1}, \alpha^{\dagger}) = q(\alpha^{\dagger}, \alpha^{l-1})$, (Metropolis version of the algorithm) $\mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger}) = \frac{\breve{g}(\alpha^{l-1}|Y)}{\breve{g}(\alpha^{\dagger}|Y)}$. If $\mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger}) > 1$, the chain moves to $\alpha^{\dagger}$. Hence, keep the draw if you move uphill. If the draw moves you downhill stay at α^{l-1} with probability $1 - \mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger})$, and explore new areas with probability equal to $\mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger})$.

Important: $q(\alpha^{l-1}, \alpha^{\dagger})$ is not necessarily equal (proportional) to posterior - histograms of draws not equal to the posterior. This is why we use a scheme which accepts more in the regions of high probability.

How do you choose $q(\alpha^{l-1}, \alpha^{\dagger})$ (the transition probability)?

- Typical choice: random walk chain. $q(\alpha^{\dagger}, \alpha^{l-1}) = q(\alpha^{\dagger} - \alpha^{l-1})$, and $\alpha^{\dagger} = \alpha^{l-1} + v$ where $v \sim \mathbb{N}(0, \sigma_v^2)$. To get "reasonable" acceptance rates adjust σ_v^2. Often $\sigma_v^2 = c * \Omega_\alpha, \Omega_\alpha = [-g''(\alpha^*|y)]^{-1}$. Choose c.

Alternatives:

- Reflecting random walk: $\alpha^{\dagger} = \mu + (\alpha^{l-1} - \mu) + v$

- Independent chain $q(\alpha^{\dagger}, \alpha^{l-1}) = \bar{q}(\alpha^{\dagger})$,$\mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger}) = min[\frac{w(\alpha^{\dagger})}{w(\alpha^{l-1})}, 1]$, where $w(\alpha) = \frac{g(\alpha|Y)}{\bar{q}(\alpha)}$. Monitor both the location and the shape of $\bar{q}$ to insure reasonable acceptance rates. Standard choices for $\bar{q}$ are normal and t.

• General rule for selecting q. A good q must:

a) be easy to sample from

b) be such that it is easy to compute $\mathfrak{E}$.

c) each move goes a reasonable distance in parameter space but does not reject too frequently (ideal acceptance rate 25-40%).

Implementation issues

A) How to draw posterior samples?

- Produce one sample (of dimension $n * L + \bar{L}$). Throw away initial $\bar{L}$ observations. Keep only elements $(L, 2L, \ldots, n * L)$ (to eliminate the serial correlation of the draws).

- Produces n samples of $\bar{L} + L$ elements. Use last L observations in each sample for inference.

- Dynare setup to produce n samples. By default it keeps the last 25 percent of the draws of each chain. **Careful: Need to make sure that with 75 percent of the draws each chain has converged**.

B) How long should be $\bar{L}$? How do you check convergence?

- Start from different α^0. Check if sample you keep, for a given $\bar{L}$, has same properties (Dynare approach).

- Choose two points, $\bar{L}_1 < \bar{L}_2$; compute distributions/moments of α after these points. If visually similar, algorithm has converged at $\bar{L}_1$. Could this recursively $\rightarrow$ CUMSUM statistic for mean, variance, etc.(checks if it settles down, no testing required).

For simple problems $\bar{L} \approx 50$ and $L \approx 200$.

For DSGEs $\bar{L} \approx 100,000 - 200,000$ and $L \approx 500,000$. If Multiple modes are present L could be even larger.

C) How do you compute interesting statistics: easy.

- Weak Law of Large Numbers $E(h(\alpha)) \approx \frac{1}{j}\sum_{j=1}^{n} h(\alpha^{jL})$, where α^{jL} is the $j * L$-th observation drawn after $\bar{L}$ iterations are performed.

- $E(h(\alpha)h(\alpha)') = \sum_{-J(L)}^{J(L)} w(\tau)ACF_h(\tau)$; $ACF_h(\tau) =$ autocovariance of $h(\alpha)$ for draws separated by τ periods; $J(L)$ function of L, $w(\tau)$ a set of weights.

- Marginal density $(\alpha_k^1, \ldots \alpha_k^L)$: $g(\alpha_k|y) = \frac{1}{L}\sum_{j=1}^{L} g(\alpha_k|y, \alpha_{k'}^j, k' \neq k)$.

- Predictive inference $f(y_{t+\tau}|y_t) = \int f(y_{t+\tau}|y_t, \alpha)g(\alpha|y_t)d\alpha$.

- Model comparisons: compute marginal likelihood numerically.

3.4 Model Comparison

• $f(y)$ the marginal likelihood (ML) is a measure of fit.

• Can compare the ML of two models. The one with the largest ML is the best.

• Bayes Factor (BF): $\frac{f(y|M_1)}{f(y|M_2)}$

• Posterior odds (PO): $\frac{f(y|M_1)g(M_1)}{f(y|M_2)g(M_2)}$, where $g(M_1), g(M_2)$ are the priors on the two models.

• Rule of thumb: BF (PO) $<$ 3 inconclusive; 3 $< BF <$ 10 favour M_1, $BF >$ 10 strongly favouring M_1

Notes:

- M_1 and M_2 could be two structural models, two time series models or one structural and one time series model

- BIC model selection criteria is an asymptotic expansion of BF.

- Can compare models with different number of parameters (since they are integrated out)

- Can compare models which are non-nested (difficult to do with classical methods).

4 Robustness

• Typically prior chosen to make calculation convenient. How sensitive are results to prior choice?

• Typical approach: repeat estimation for different priors (inefficient).

• Alternative.

i) Select a prior $g_1(\alpha)$ with support included in $g(\alpha)$.

ii) Let $w(\alpha) = \frac{g(\alpha)}{g_1(\alpha)}$. Then any $h_1(\alpha) = \int (h(\alpha)w(\alpha)dg_1(\alpha)$ can be computed using $h_1(\alpha) \approx \frac{\frac{1}{L}\sum_l w(\alpha^l)h(\alpha^l)}{\sum_l w(\alpha^l)}$, where $h(\alpha^l)$ are the statistics computed with $g(\alpha)$.

•Just need the original output obtained and a set of weights!

Example 4.1 $y_t = x_t\alpha + u_t$ $u_t \sim (0, \sigma^2)$. *Suppose* $g(\alpha)$ *is* $\mathbb{N}(0, 10)$. *Then* $g(\alpha|Y)$ *is normal with mean* $\tilde{\alpha} = \tilde{\Sigma}^{-1}(0.1 + \sigma^{-2}x'x\alpha_{ols})$ *and variance* $\tilde{\Sigma} = 0.1 + \sigma^{-2}x'x$, . *If one wishes to examine how forecasts of the model change when the prior variances changes (for example to 5) two alternatives are possible:*

(a) draw from normal $g(\alpha|Y)$ *which has mean* $\tilde{\alpha}_1 = \tilde{\Sigma}_1^{-1}(0.2+\sigma^{-2}x'x\alpha_{ols})$ *and variance* $\tilde{\Sigma} = 0.2 + \sigma^{-2}x'x$, *and compare forecasts.*

(b) Weight draws from the initial posterior distribution with $\frac{g(\alpha)}{g_1(\alpha)}$ *where* $g_1(\alpha)$ *is* $N(0, 5)$.

5 Bayesian estimation of DSGE models

Why using Bayesian methods to estimate DSGE models?

1) Hard to include non-sample information in classical ML (a part from range of possible values).

2) Classical ML is justified only if the model is the GDP of the actual data. Can use Bayesian methods for misspecified models (economic inference may be problematic, no problem for statistical inference).

3) Can incorporate prior uncertainty about parameters and models.

General Principles:

• (log-)linearized DSGE models are state space models whose reduced form parameters α are nonlinear functions of structural θ. Compute the likelihood of θ via the Kalman filter.

• Compute posterior of θ with MH algorithm.

• Use posterior output to compute the marginal likelihood, Bayes factors and any posterior function of the parameters (impulse responses, ACF, turning point predictions, forecasts, etc.).

• Check robustness to the choice of prior.

General algorithm: Given some initial θ_0

[1.] Construct a log-linear solution of the DSGE economy.

[2.] Specify prior distributions $g(\theta)$.

[3.] Transform the data to make sure that is conformable with the model.

[4.] Compute likelihood via Kalman filter.

[5.] Draw sequences for θ using MH algorithm. Check convergence.

[6.] Compute marginal likelihood and compare it to the one of alternative models using Bayes factors.

[7.] Construct statistics of interest. Use loss-based evaluation of discrepancy model/data.

[8.] Perform robustness exercises.

Step [1.]: can have nonlinear state space models (see later and e.g. Amisano and Tristani (2006), Rubio and Villaverde (2009)) or value function problems (see Bi and Traum (2012)) but computations much more complex.

Recall DSGE models are typically singular! Need to:

i) add measurement errors to use all observables (where to put measurement error? All variables or just enough to complete the probability space?)

ii) find a way to reduce the dimensionality of the system (substituting equations before the solution is computed).

iii) choose the observables optimally (see Canova et al. (2013)).

iv) invent new structural shocks.

In Step [3.] transformations are needed because the model is typically solved in deviation from the steady states. Need to eliminate from the data any long run component. How do you do it? Many ways of doing this (see Canova, 2010) all unsatisfactory.

[4.] is the most computationally intensive step. Considerable gains if this is efficiently done.

In step [5.]: Given a θ^l

i) Draw a candidate $\theta^\dagger$ from the $\mathfrak{P}(\theta^\dagger \theta^l)$. Solve the model.

ii) Compute the likelihood with the kalman filter.

iii) Evaluate the posterior kernel at the draw $\breve{g}(\theta^\dagger|y) = f(y|\theta^\dagger)g(\theta^\dagger)$.

iv) Given the posterior kernel at θ^l i.e $\breve{g}(\theta^0|y) = f(y|\theta^l)g(\theta^l)$, compute $IR = \frac{\breve{g}(\theta^\dagger)}{\breve{g}(\theta^l)}\frac{\mathfrak{P}(\theta^l,\theta^\dagger)}{\mathfrak{P}(\theta^\dagger,\theta^l)}$.

vi) If $IR > 1$ set $\theta^{l+1} = \theta^\dagger$. Else draw $\varpi \sim U(0,1)$. If $\varpi < IR$ set $\theta^{l+1} = \theta^\dagger$ otherwise set $\theta^{l+1} = \theta^l$.

vii) Repeat i)-vi) $\bar{L} + nL$ times. Throw away $\bar{L}$ draws. Keep one every n for inference.

In Step [6.] it is typical to use a modified harmonic mean estimator i.e. approximate $\mathcal{L}(y_t|\mathcal{M}_i)$ using $[\frac{1}{L}\sum_l \frac{f(\alpha_l^i)}{\mathcal{L}(y_t|\alpha_l^i,\mathcal{M}_i)g(\alpha_l^i|\mathcal{M}_i)}]^{-1}$ where α_i^l is the draw l of the parameters α of model i and f is a density with tails thicker than a normal. If $f(\alpha_l^i) = 1$ we have a simple harmonic mean estimator.

Competitors could be a more densely parametrized structural model (nesting the interested one) or more densely parametrized reduced form model (e.g. VAR or a BVAR).

Bayes factors can be computed numerically or via Laplace approximations (to decrease computational burden in large scale systems).

In step [7.] Estimate marginal/ joint posteriors using kernel methods. Compute point estimate and credible sets. Compute continuous functions $h(\theta)$ of interest. Set up a loss function. Compare models using the risk function.

In step [8.] Reweight the draws appropriately.

Example 5.1 *(One sector growth model)*

- *Analytic solution if* $U(c,l) = \ln c$ *and* $\delta = 1$. *Equations are:*

$$\begin{aligned} K_{t+1} &= (1-\eta)\beta A K_t^{1-\eta}\zeta_t + u_{1t} && (7)\\ GDP_t &= A K_t^{1-\eta}\zeta_t + u_{2t} && (8)\\ c_t &= \eta\beta GDP_t + u_{3t} && (9)\\ r_t &= (1-\eta)\frac{GDP_t}{K_t} + u_{4t} && (10) \end{aligned}$$

- ζ_t *technology shock,* u_{jt} *measurement errors added to avoid singularity.*

Parameters: β*: is the discount factor,* $1-\eta$*: the share of capital in production,* σ^2*: variance of technology shock, A: constant in the production function.*

Simulate 1000 points from using $k_0 = 100.0$ *using* $A = 2.86; 1 - \eta = 0.36; \beta = 0.99, \sigma^2 = (0.07)^2$.

Assume $u_{1t} \sim \mathbb{N}(0, 0.1^2)$; $u_{2t}^m \sim \mathbb{N}(0, 0.06^2)$; $u_{3t}^m \sim \mathbb{N}(0, 0.02^2)$; $u_{4t}^m \sim \mathbb{N}(0, 0.08^2)$; *(Note: lots of measurement error!)*

- Keep last 160 as data (to mimic about 40 years of quarterly data).

Interested in $(1 - \eta), \beta$ *i.e (treat* σ^2, A *as fixed).*

Use (9)-(10) to identify the parameters from the data.

Priors: $(1-\eta) \sim$ *Beta(3,5);* $\beta \sim$ *Beta(98,2) (NOTATION DIFFERENT FROM DYNARE)*

Mean of a Beta(a,b) is $(a/a+b)$ *and the variance of a Beta(a,b)is* $ab/[(a+b)^2 * (a+b+1)]$. *Thus prior mean of* $1-\eta = 0.37$, *prior variance* 0.025; *prior mean of* $\beta = 0.98$, *prior variance* 0.0001.

Let $\theta = (1-\eta, \beta)$ *Use random walk to draw* $\theta^{\dagger}$, *i.e.* $\theta^{\dagger} = +\theta^{l-1} + e^{\dagger}$, μ *is the mean and* e_i *is* $U(-0.08, 0.08)$ *for* β *and* $U(-0.06, 0.06)$ *for* η *(roughly about 28% acceptance rate).*

Draw 10000 replications from the posterior kernel. Convergence is fast.

Keep last 5000; use one every 5 for inference.

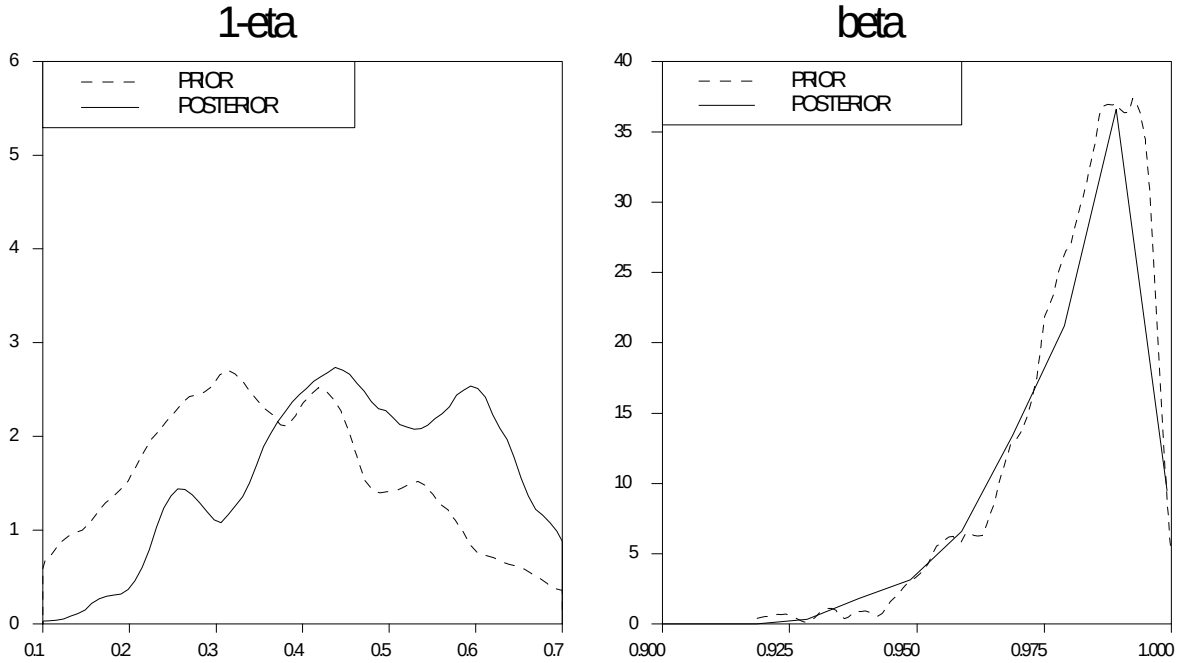


Figure 4: Priors and Posteriors, RBC model

- Prior for β sufficiently loose, posterior similar, data is not very formative.

-Posteriors centered around the true parameters,large dispersion.

Variances/covariances

	true	*posterior 68% range*
var(c)	*0.24*	*[0.11, 0.27]*
var(y)	*0.05*	*[0.03, 0.11]*
cov(c,y)	*0.0002*	*[0.0003, 0.0006]*

Wrong model

- Simulate data from model with habit $\gamma = 0.8$

- Estimate model conditioning on $\gamma = 0$.

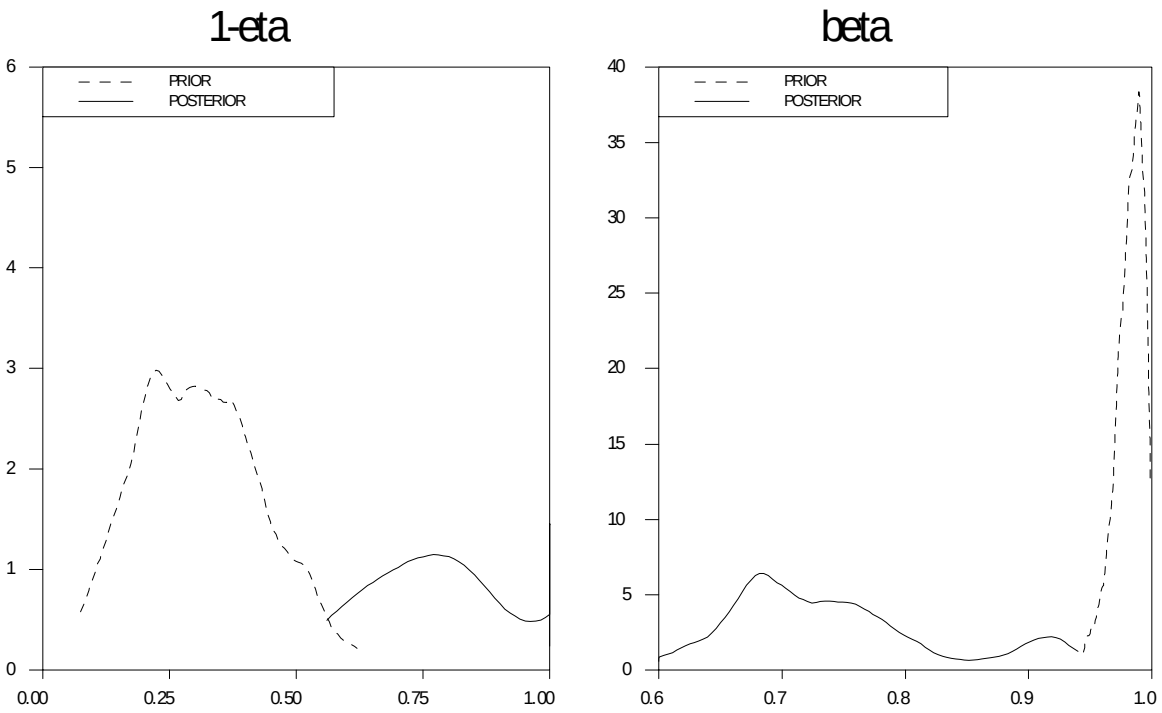


Figure 5: Priors and Posteriors, wrong model

Example 5.2 *(New Keynesian model)*

$$gap_t = E_t gap_{t+1} - \frac{1}{\varphi}(r_t - E_t\pi_{t+1}) + g_t \quad (11)$$

$$\pi_t = \beta E_t \pi_{t+1} + \kappa gap_t + v_t \quad (12)$$

$$r_t = \phi_r r_{t-1} + (1-\phi_r)(\phi_\pi \pi_{t-1} + \phi_{gap} gap_{t-1}) + e_t \quad (13)$$

$\kappa = \frac{(1-\zeta_p)(1-\beta\zeta_p)(\varphi+\vartheta_N)}{\zeta_p}$; ζ_p = *degree of (Calvo) stickiness,* β = *discount factor,* φ = *risk aversion,* ϑ_N = *elasticity of labor supply.* g_t *and* v_t *are AR(1) with persistence* ρ_g, ρ_v *and variances* σ_g^2, σ_v^2; $e_t \sim iid(0, \sigma_r^2)$.

$\theta = (\beta, \varphi, \vartheta_N, \zeta_p, \phi_\pi, \phi_{gap}, \phi_r, \rho_g, \rho_v, \sigma_v^2, \sigma_g^2, \sigma_r^2)$.

Assume $g(\theta) = \prod g(\theta_i)$

Assume $\beta \sim Beta(98, 3), \varphi \sim \mathbb{N}(1, 0.375^2), \vartheta_N \sim \mathbb{N}(2, 0.75^2), \quad \zeta_p \sim Beta(9, 3), \phi_r \sim Beta(6, 2), \phi_\pi \sim Normal(1.5, 0.1^2),$
$\phi_{gap} \sim \mathbb{N}(0.5, 0.05^2), \ \rho_g \sim Beta(17, 3), \rho_v \sim Beta(17, 3) \ \sigma_i^2 \sim IG(2, 0.01), i = g, v, r.$

Use US linearly detrended data from 1948:1 to 2002:1 to estimate the model.

Use random walk MH algorithm to draw candidates.

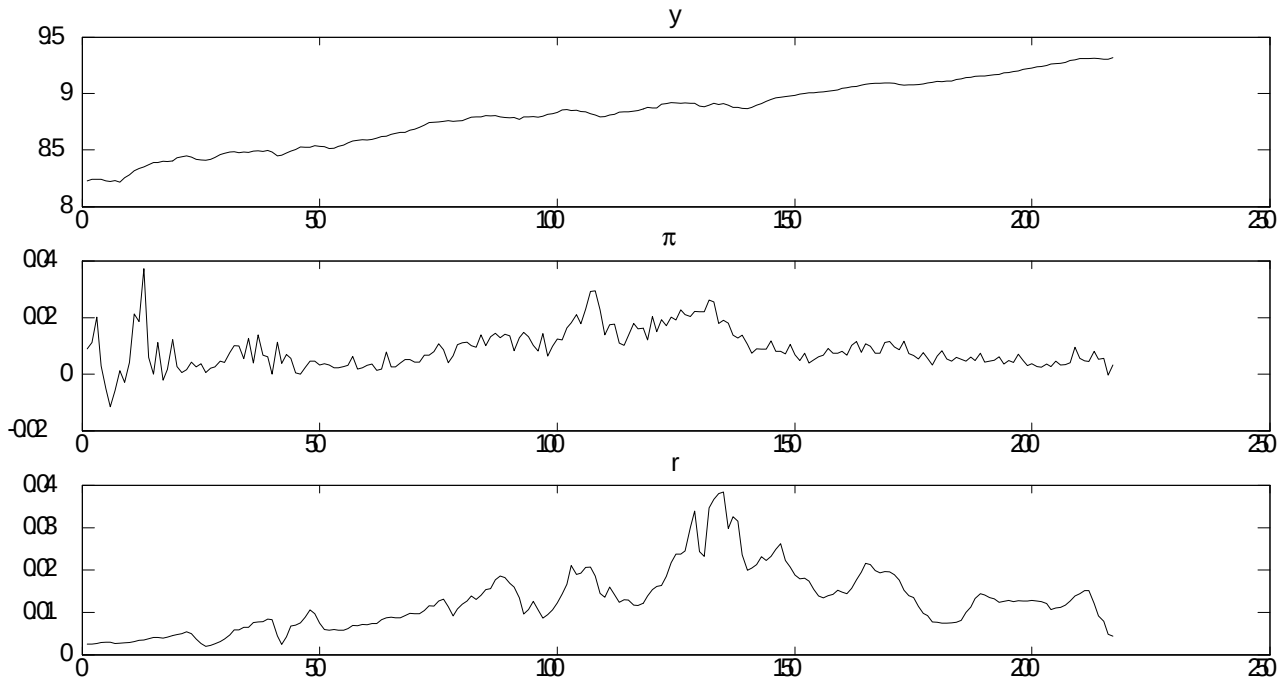


Figure 6: Raw Time series

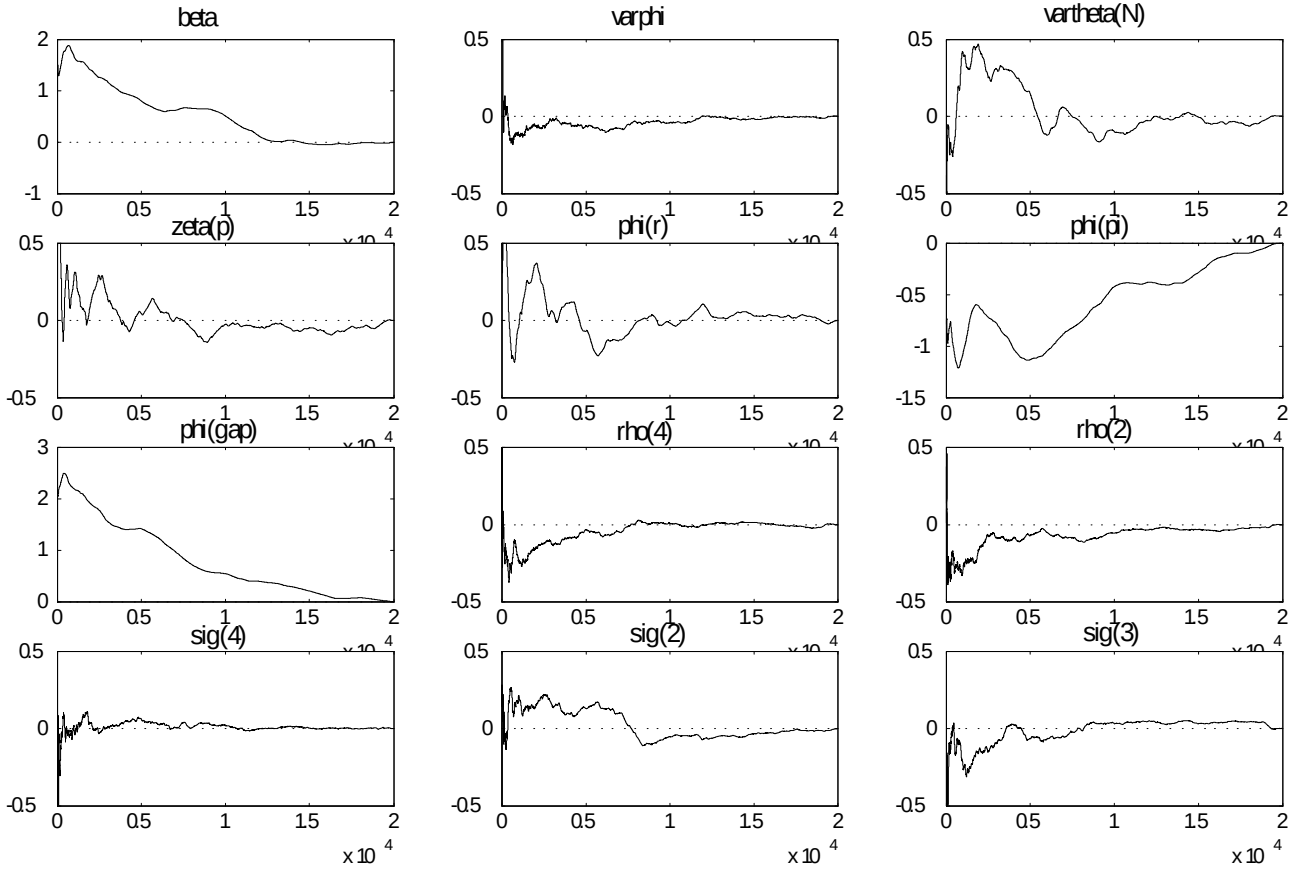


Figure 7: CUMSUM statistics

Prior and Posterior statistics

	Prior		Posterior				
	mean	*std*	*median*	*mean*	*std*	*max*	*min*
β	*0.98*	*0.01*	*0.992*	*0.991*	*0.003*	*0.999*	*0.998*
φ	*1.00*	*0.37*	*0.826*	*0.843*	*0.123*	*1.262*	*0.425*
ϑ_N	*2.00*	*0.75*	*1.825*	*1.884*	*0.768*	*3.992*	*0.145*
ζ_p	*0.75*	*0.12*	*0.743*	*0.696*	*0.195*	*0.997*	*0.141*
ϕ_r	*0.75*	*0.14*	*0.596*	*0.587*	*0.154*	*0.959*	*0.102*
ϕ_π	*1.50*	*0.10*	*1.367*	*1.511*	*0.323*	*2.33*	*1.042*
ϕ_{gap}	*0.5*	*0.05*	*0.514*	*0.505*	*0.032*	*0.588*	*0.411*
ρ_g	*0.85*	*0.07*	*0.856*	*0.854*	*0.036*	*0.946*	*0.748*
ρ_u	*0.85*	*0.07*	*0.851*	*0.851*	*0.038*	*0.943*	*0.754*
σ_g	*0.025*	*0.07*	*0.025*	*0.025*	*0.001*	*0.028*	*0.021*
σ_v	*0.025*	*0.07*	*0.07*	*0.07*	*0.006*	*0.083*	*0.051*
σ_r	*0.025*	*0.07*	*0.021*	*0.021*	*0.005*	*0.035*	*0.025*

- Little information in the data for some parameters (prior and posterior overlap).

- For parameters of the policy rule: posteriors move and not more concentrated.

-Posterior distributions roughly symmetric except for ϕ_π and ζ_p (mean and median coincide).

-Posterior distribution of economic parameters reasonable (except φ).

- Posterior for the AR parameters has a high mean, but no pile up at one.

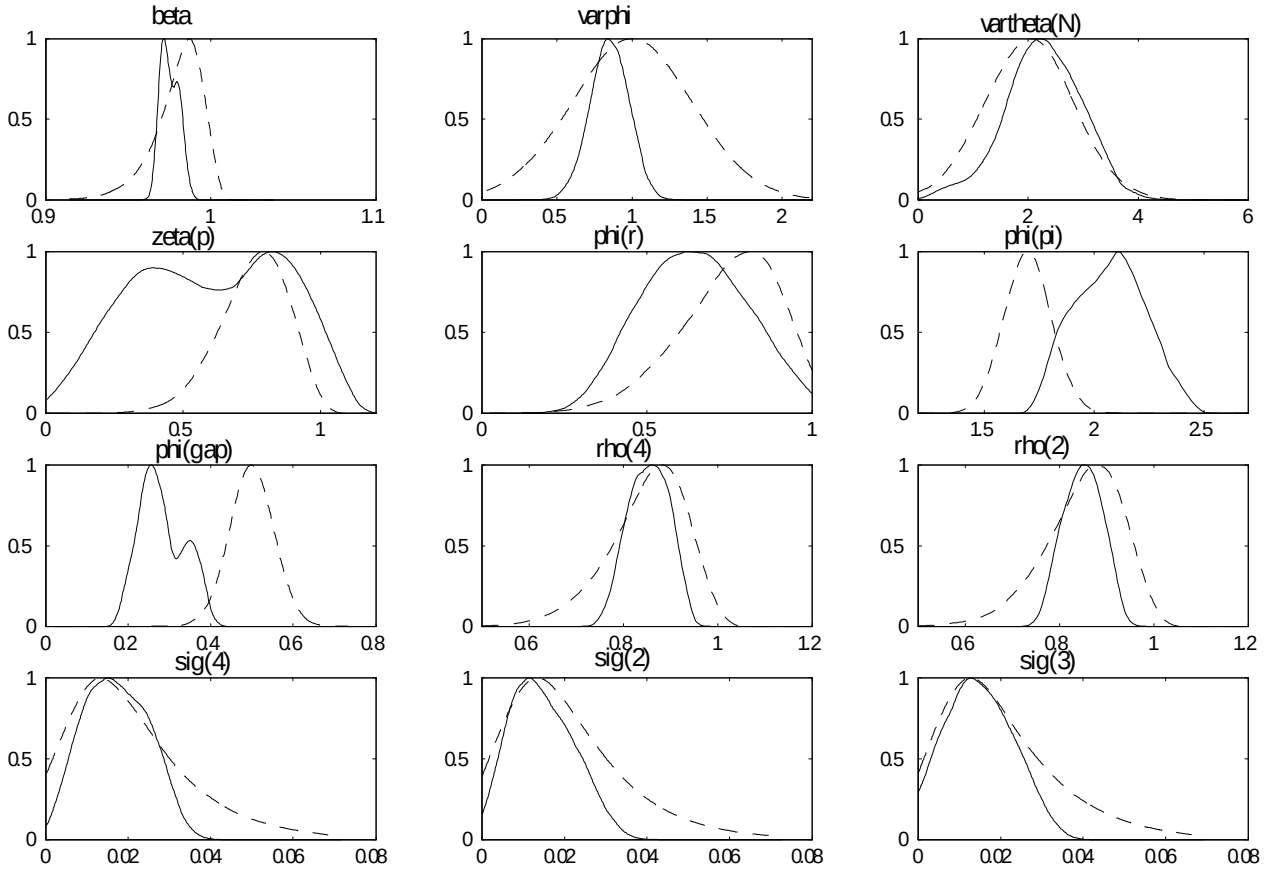


Figure 8: Priors and Posteriors, NK model

Model comparisons

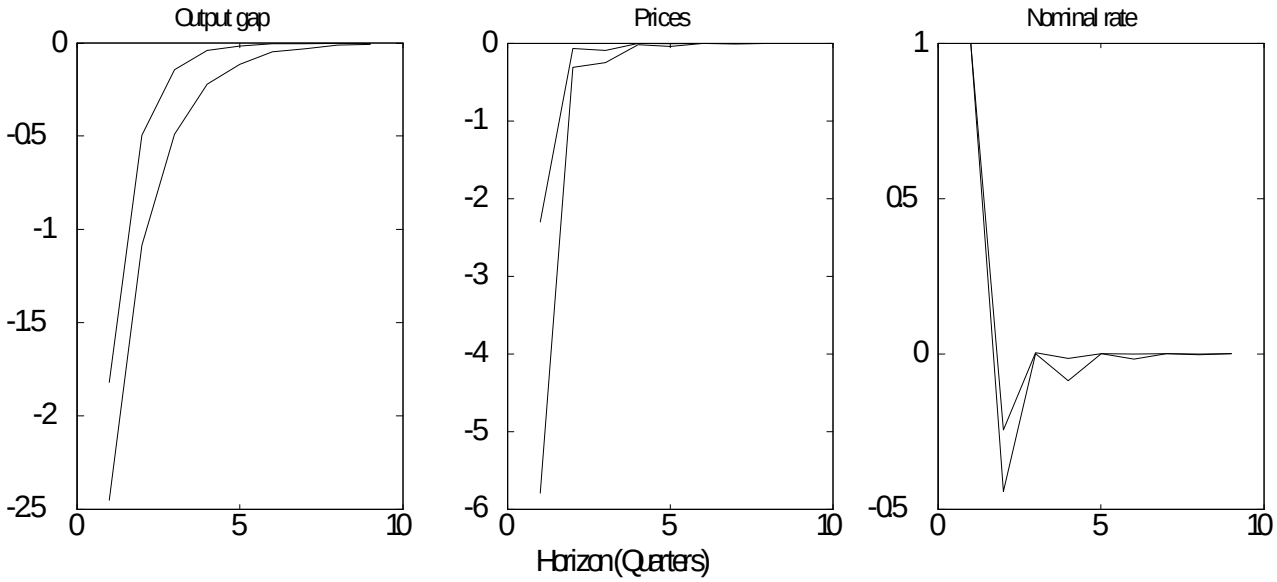
Compare ML against flat prior VAR(3) or a BVAR(3) with Minnesota prior and standard parameters (tightness=0.1, linear lag decay and weight on other variables equal 0.5), both with a constant.

Bayes factor are very small $\approx$ 0.02 in both cases.

- *The restrictions the model imposes are false. Need to add features to the model that make dynamics of the model more similar to those of a VAR(3).*

Posterior analysis

How do responses to monetary shocks look like? No persistence!



How much of the output gap and inflation variance explained by monetary shocks? Almost all!!

Detecting a multimodal posterior

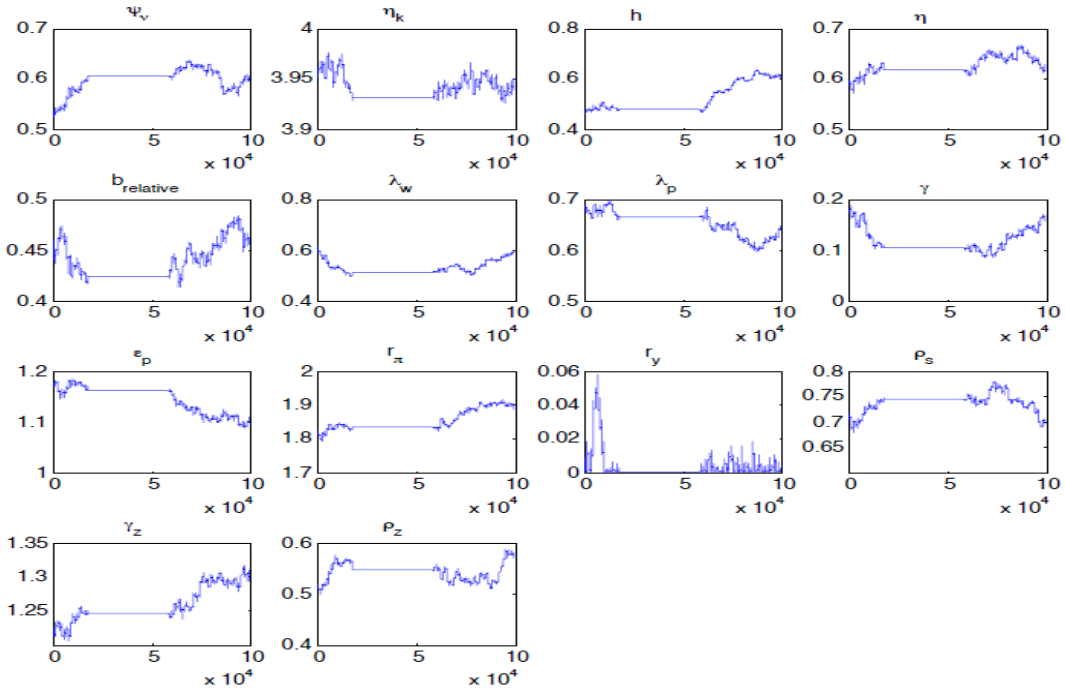
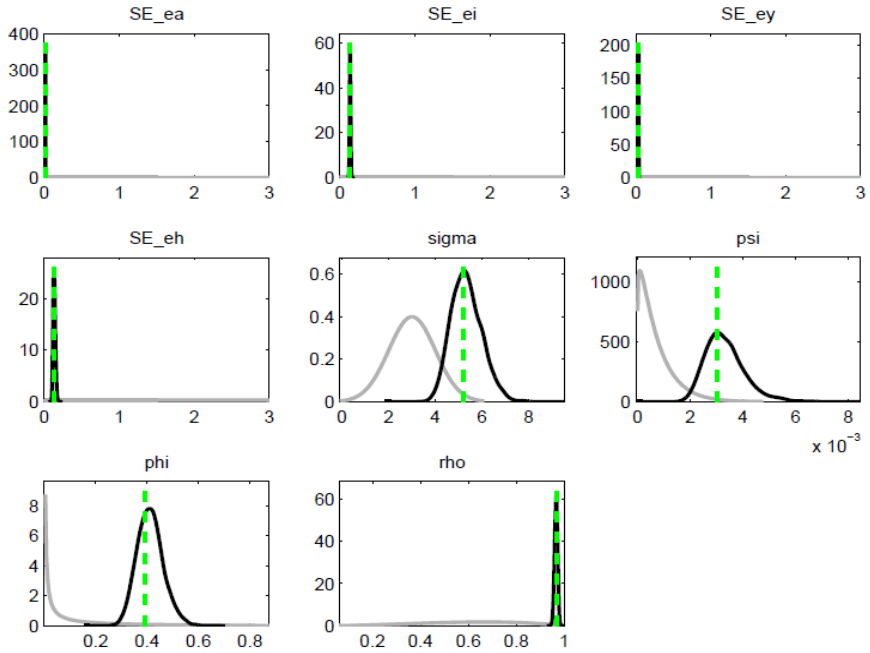


Figure 1. Draws of parameters

Normal vs. MCMC approximations



5.1 Interpreting results

- Most of the shocks of DSGE models are non-structural (alike to measurement errors). Careful with interpretation and policy analyses with these models (see Chari et al. (2009)).

- A model where "measurement errors" explain a large portion of main macro variables is suspicious (e.g. in Smets and Wouters (2003) markup shocks dominate).

- If the standard error of one the shocks is large relative to the others: evidence of misspecification.

- Compare estimates with standard calibrated values. Are they sensible? Often yes, but because of tight priors are centered at calibrated values.

5.2 Bayesian methods and identification

- Likelihood of a DSGE typically flat. Could be due to marginalization (use only a subset of economic relationships), or to lack of information.

- Difficult to say a-priori which parameters is identified and which is not: we do not have an analytic relationship between the reduced form and the structural parameters.

- Could go a long way to understand if problems exists by numerically constructing the likelihood as a function of the structural parameters, see Canova and Sala (2009).

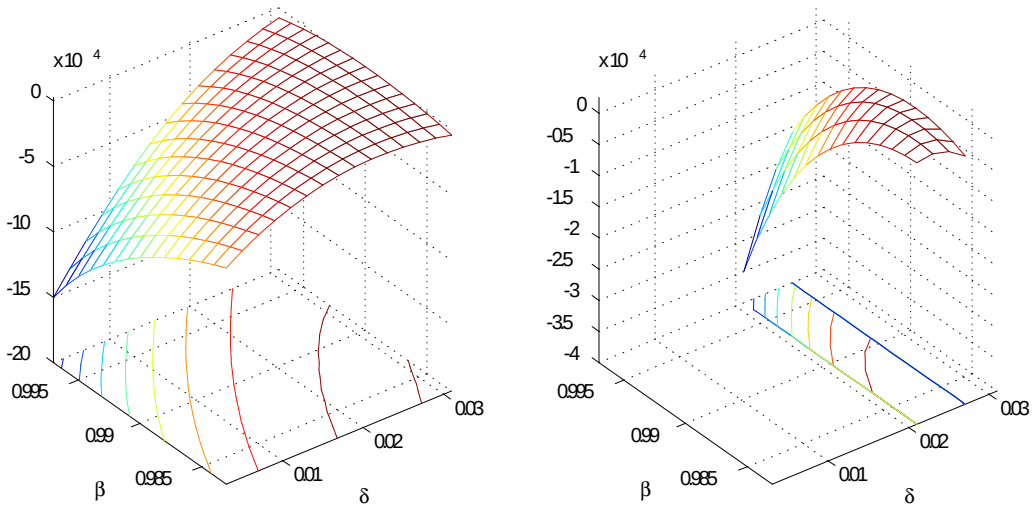
- Standard remedy when some parameters are hard to identify: calibrate them. Problem if parameter not fixed at a consistent estimator $\longrightarrow$ biases could be extensive! (see Canova and Sala, 2009).

- Alternative: add a prior. A prior may increases the curvature of the likelihood $\longrightarrow$ underidentification may be hidden!. Posterior may look nice because the prior does the job!!.

- Important

i)In general if $\mathcal{L}(\theta_1, \theta_2|Y^T) = \bar{\mathcal{L}}(\theta_1|Y^T)$ then $g(\theta_1, \theta_2|Y^T) = g_1(\theta_1|Y^T)$ $g(\theta_2|\theta_1)$, i.e no updating of conditional prior of θ_2.

ii) However, updating possible even if no sample information is present if θ_1, θ_2 are linked by economic or stability conditions!!



Likelihood and Posterior, δ and β in a RBC model

- Posterior nicer than the likelihood because the prior rules out certain part of the parameter space. Many examples of this. see Williams (2009) study of the Smets and Wouter model or Herbst and Schorfheide (2013) study of Schmitt-Grohe and Uribe (2012) model.

- If prior $\approx$ posterior two possibilities: weak data information or prior which is too much data based.

- How do you detect which is the problem?

i) Move the location of the prior. If the posterior follows the prior, then there is weak information in the data.

ii) Change the sample size and see how the posterior changes.

- Formal methods to detect underidentification: Iskrev (2010); Komunjer and Ng (2012); Mueller (2012); Qu and Thachenko (2012).

- Formal methods have hard time to detect weak and partial identification problems. Graphical methods of Canova and Sala (2009) useful.