

An introduction to Bayesian Econometrics for macroeconomists

V: MCMC; label switching, the Metropolis-Hastings algorithm, Model
comparison

Gianni Amisano

University of Brescia

`amisano@eco.unibs.it`

Milan, 31/01/06

Contents

1	TVP-VAR (continued)	3
1.1	Simulation of the parameters	3
2	Label switching in MS models	4
3	Metropolis-Hastings algorithm	7
3.1	Independence MH	9
3.2	Random walk MH	11
4	Model comparison with MCMC methods	12

1 TVP-VAR (continued)

1.1 Simulation of the parameters

Given a Wishart prior for $\mathbf{H}_{\varepsilon\varepsilon}$, $\mathbf{H}_{\eta\eta}$ and a Gaussian prior on \mathbf{A} , their posterior distribution have same functional form. So they can easily be simulated

Often \mathbf{A} is set to identity matrix.

2 Label switching in MS models

Take simple MS model (no covariates, only an intercept term)

$$p(y_t | s_t, \boldsymbol{\theta}) = N(\mu_{s_t}, h_{s_t}^{-1}) \quad (1)$$

$$\mathbf{P} = \{p(s_t = j | s_{t-1} = i, \boldsymbol{\theta})\} \quad (2)$$

If you permute labelling of the states (ie you call state 2 what you called state 1 before), the likelihood of the model does not change. So we have an identification problem.

Just to fix ideas, let's work with a numeric example

$$p(y_t | s_t = 1, \boldsymbol{\theta}) = N(1, 1) \quad (3)$$

$$p(y_t | s_t = 2, \boldsymbol{\theta}) = N(-1, 10) \quad (4)$$

$$\mathbf{P} = \begin{bmatrix} .9 & .1 \\ .3 & .7 \end{bmatrix} \quad (5)$$

the likelihood of this model is exactly the same as the likelihood of the following

one

$$p(y_t | s_t = 1, \theta) = N(-1, 10) \quad (6)$$

$$p(y_t | s_t = 2, \theta) = N(1, 1) \quad (7)$$

$$\mathbf{P} = \begin{bmatrix} .7 & .3 \\ .1 & .9 \end{bmatrix} \quad (8)$$

This is an identification problem, related to the structural interpretation of the latent states, which can only be solved by imposing constraints.

Examples of constraints (one of them is enough to achieve identification) are

$$\mu_1 > \mu_2 \quad (9)$$

$$h_1 > h_2 \quad (10)$$

$$p_{11} > p_{22} \quad (11)$$

These constraints can be implemented in 2 different ways

1. draw from the relevant conditional posterior distribution that ignores the constraint until you finally satisfy it (works with any priors). For instance if constraint was (9) we draw from $p(\mu_1, \mu_2 | h_1, h_2, \mathbf{P}, \mathbf{y})$ until we get a draw satisfying the constraint.
2. Suppose we use a prior distribution which is symmetric across states, ie

$$p(h_1) = p(h_2) \quad (12)$$

$$p(\mu_1) = p(\mu_2) \quad (13)$$

$$p(p_{11}) = p(p_{22}) \quad (14)$$

we can draw from relevant conditional posteriors and then permute the order of the result to achieve the required ordering. In other words, suppose the draw from the relevant conditional posteriors is

$$\mu_1 = 0.1, \mu_2 = 0.2 \quad (15)$$

$$h_1 = 1.5, h_2 = 2.5 \quad (16)$$

$$p_{11} = .8, p_{22} = .9 \quad (17)$$

and we want to impose (9), then the draw would be permuted as

$$\mu_2 = 0.1, \mu_1 = 0.2 \quad (18)$$

$$h_2 = 1.5, h_1 = 2.5 \quad (19)$$

$$p_{22} = .8, p_{11} = .9 \quad (20)$$

3 Metropolis-Hastings algorithm

The mother of all MCMC algorithms!!!

See Chib, 2001.

A *MH* algorithm to simulate $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})$ works as follows: starting from $\mathbf{x} = \boldsymbol{\theta}^{(i)}$, a draw $\mathbf{y} = \boldsymbol{\theta}^{(i+1)}$ is generated from a 'candidate' density $q(\mathbf{x}, \mathbf{y})$.

This draw is either retained, or discarded (setting $\mathbf{y} = \mathbf{x}$ in the following step of the algorithm), subject to a further dichotomic randomisation with probability:

$$\alpha(\mathbf{x}, \mathbf{y}) = \min \left\{ \frac{\pi(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x}, \mathbf{y})}, 1 \right\} \quad (21)$$

The sufficient conditions for the *MH* chain to converge to $\pi(\mathbf{y})$ are very mild and easy to verify: $\pi(\mathbf{x})$ and $q(\mathbf{x}, \mathbf{y})$ be positive and continuous $\forall \mathbf{x}, \mathbf{y} \in \Theta$.

The main issue in the implementation of the *MH* algorithm is connected to the choice of $q(\mathbf{x}, \mathbf{y})$.

It is desirable that $\alpha(\mathbf{x}, \mathbf{y})$ be as stable as possible across draws.

It is also desirable that the acceptance rate (number of times the chain moves in the posterior simulation) be not too high nor too low: around 25-30%.

Like in importance sampling, candidate distribution $q(\mathbf{x}, \mathbf{y})$ should have fatter tails than

target distribution $\pi(\boldsymbol{\theta})$.

Different versions of the algorithm

3.1 Independence MH

In the IMH the candidate pdf does not depend on previous draw

$$q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}) \quad (22)$$

Usually $q(\mathbf{y})$ is multivariate Gaussian (or Student-t) centered on modal value of posterior and covariance matrix given by minus inverse Hessian of log posterior evaluated at modal value

$$q(\mathbf{y}) = N(\tilde{\boldsymbol{\theta}}, \mathbf{V}) \quad (23)$$

$$\tilde{\boldsymbol{\theta}} = \arg \max \ln p(\boldsymbol{\theta}|y)$$

$$\mathbf{V} = -c \times \left\{ \left[\frac{\partial^2 \ln p(\boldsymbol{\theta}|y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \right\}^{-1} \quad (24)$$

c is a tuning constant.

3.2 Random walk MH

In RWMH $q(\mathbf{x}, \mathbf{y})$ is symmetric distribution (eg Gaussian, or Student-t) centered on previous draw and with arbitrary covariance matrix

$$q(\mathbf{x}, \mathbf{y}) = N(\mathbf{x}, \mathbf{V}) \quad (25)$$

In this way

$$q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}, \mathbf{x}) \quad (26)$$

Hence acceptance probability simplifies

$$\alpha(\mathbf{x}, \mathbf{y}) = \min \left\{ \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}, 1 \right\} \quad (27)$$

ie always accept when we go "uphill".

4 Model comparison with MCMC methods

We have seen before that in Bayesian econometrics, model comparison is based on PORs:

$$K_{ij} = \frac{p(M_i|\mathbf{y})}{p(M_j|\mathbf{y})} = K_0 \frac{\int p(\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i|\mathbf{y})d\boldsymbol{\theta}_i}{\int p(\boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j|\mathbf{y})d\boldsymbol{\theta}_j} = K_0 \frac{p(\mathbf{y}|M_i)}{p(\mathbf{y}|M_j)} = K_0 BF_{ij}$$

$$K_0 = \frac{p(M_i|\mathbf{y})}{p(M_j|\mathbf{y})} = \text{prior odds ratio}$$

$$BF_{ij} = \frac{p(\mathbf{y}|M_i)}{p(\mathbf{y}|M_j)} = \text{Bayes factor of model i vs model j}$$

Note that the key element of *POR* computation is the evaluation of marginal likelihoods, i.e. the predictive density of data under the two separate models.

To do model comparison need to evaluate marginal likelihoods.

In some particular contexts, the *POR* can be written in a way that is very convenient for computations. Suppose we are interested in comparing two competing hypotheses concerning $\boldsymbol{\theta}$, the parameter vector of a certain model:

$$\begin{aligned} H_A & : p_A(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_A, \\ H_B & : p_B(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_B, \end{aligned}$$

where Θ_A , and Θ_B are the supports of $\boldsymbol{\theta}$ under H_A and H_B respectively.

POR:

$$POR = \frac{p(H_A|\mathbf{y})}{p(H_B|\mathbf{y})} = \frac{\int_{\Theta_A} p_A(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta}, H_A)d\boldsymbol{\theta}}{\int_{\Theta_B} p_B(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta}, H_B)d\boldsymbol{\theta}}. \quad (28)$$

Let us consider the case of a point hypothesis against a composite competing one.

As a simple example of this, let us consider the partition $\boldsymbol{\theta} = [\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2]'$, and the hypotheses

$$\begin{aligned} H_A & : \boldsymbol{\theta}_1 = \mathbf{h}_1, \boldsymbol{\theta}_2 \in \Theta_2, \\ H_B & : \boldsymbol{\theta}_1 \in \Theta_1, \boldsymbol{\theta}_2 \in \Theta_2 \end{aligned}$$

with prior pdf

$$\begin{aligned} p(\boldsymbol{\theta}_1 = \mathbf{h}_1 | H_A) &= 1, \\ & p(\boldsymbol{\theta}_1 | H_B), \\ p(\boldsymbol{\theta}_2 | H_A) &= p(\boldsymbol{\theta}_2 | H_B) \end{aligned}$$

In this case POR is the so-called "Savage density ratio" (Kass and Raftery, 1995):

$$POR = \frac{p(\boldsymbol{\theta}_1 = \mathbf{h}_1 | \mathbf{y}, H_B)}{p(\boldsymbol{\theta}_1 = \mathbf{h}_1 | H_B)}, \quad (29)$$

When the joint posterior distribution under H_B is analytically intractable, a convenient way to calculate the numerator in the expression above is to evaluate analytically the conditional posterior pdf $p(\boldsymbol{\theta}_1 = \mathbf{h}_1 | \boldsymbol{\theta}_2, \mathbf{y}, H_B)$, and average it over draws taken from the posterior distribution $p(\boldsymbol{\theta}_2 | \mathbf{y}, H_B)$. The POR is hence consistently estimated via Monte Carlo simulation as:

$$\overline{POR} = \frac{1}{p(\boldsymbol{\theta}_1 = \mathbf{h}_1 | H_B)} \times \frac{1}{N} \sum_{i=1}^N p(\boldsymbol{\theta}_1 = \mathbf{h}_1 | \boldsymbol{\theta}_2^{(i)}, \mathbf{y}, H_B). \quad (30)$$

In general circumstances, computing PORs is a very tricky issue. Chib (1995) has proposed an approach that can be used when the posterior distribution is sampled by means of Gibbs sampling.

Suppose we use a GS approach with s blocks

Chib's method is based on the identity

$$\ln p(\mathbf{y}) = \ln p(\boldsymbol{\theta}) + \ln p(\mathbf{y}|\boldsymbol{\theta}) - \ln p(\boldsymbol{\theta}|\mathbf{y}) \quad (31)$$

which holds for all models and for all values of the parameters.

Therefore choose a value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^*$. Any $\boldsymbol{\theta}^*$ theoretically would do, but in practice it is much better to choose a high posterior density value such as mode or median or mean) and compute the identity. Most difficult bit is $\ln p(\boldsymbol{\theta}^*|\mathbf{y})$, which can be evaluated as follows

$$\ln p(\boldsymbol{\theta}^*|\mathbf{y}) = \ln p(\boldsymbol{\theta}_1^*|\mathbf{y}) + \ln p(\boldsymbol{\theta}_2^*|\boldsymbol{\theta}_1^*, \mathbf{y}) + \dots + \ln p(\boldsymbol{\theta}_s^*|\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_{s-1}^*, \mathbf{y}) \quad (32)$$

Each of the factors can be estimated by simulation. This method is very computationally burdensome.

In Chib (2001) there is another proposal for the case in which the posterior is sampled via Metropolis-Hastings.

A sensible approach is the one proposed by Gelfand and Dey (1994), which is based on the separate estimation of the marginal likelihoods under both models. Gelfand and Dey's method works as follows: take any function of the parameters that integrates to one over the parameter space, i.e.

$$\int f(\boldsymbol{\theta})d\boldsymbol{\theta} = 1 \quad (33)$$

Then remember Bayes theorem

$$p(\boldsymbol{\theta}|\mathbf{y}, H_A) = \frac{p(\boldsymbol{\theta}|\mathbf{y}, H_A)p(\mathbf{y}|\boldsymbol{\theta}, H_A)}{p(\mathbf{y}|\boldsymbol{\theta}, H_A)} \quad (34)$$

Take 33 and multiply numerator and denominator of the integrand by the posterior distribution:

$$\begin{aligned}
 \int f(\boldsymbol{\theta})d\boldsymbol{\theta} &= \int f(\boldsymbol{\theta})\frac{p(\boldsymbol{\theta}|\mathbf{y}, H_A)}{p(\boldsymbol{\theta}|\mathbf{y}, H_A)}d\boldsymbol{\theta} = & (35) \\
 &= \int f(\boldsymbol{\theta})\frac{p(\mathbf{y}|H_A)}{p(\boldsymbol{\theta}|H_A)p(\mathbf{y}|\boldsymbol{\theta}, H_A)}p(\boldsymbol{\theta}|\mathbf{y}, H_A)d\boldsymbol{\theta} = 1 \\
 &\Leftrightarrow \int f(\boldsymbol{\theta})\frac{1}{p(\boldsymbol{\theta}|H_A)p(\mathbf{y}|\boldsymbol{\theta}, H_A)}p(\boldsymbol{\theta}|\mathbf{y}, H_A)d\boldsymbol{\theta} = [p(\mathbf{y}|H_A)]^{-1}
 \end{aligned}$$

The last 2 lines of expression 35 mean that if we have a sample of N draws from the posterior distribution of $\boldsymbol{\theta}$ under H_A , we could consistently estimate the inverse marginal likelihood of the model A as:

$$\frac{1}{N} \sum_{i=1}^N \frac{f(\boldsymbol{\theta}^{(i)})}{p(\boldsymbol{\theta}^{(i)}|H_A)p(\mathbf{y}|\boldsymbol{\theta}^{(i)}, H_A)}$$

These estimated marginal likelihoods can be used to construct PORs.

How to choose $f(\cdot)$? Use function that integrates to one and has thinner tails than the posterior (to have quantity inside the summation bounded). Geweke

(1999) suggests a truncated multivariate distribution

$$\begin{aligned}
 f(\boldsymbol{\theta}^{(i)}) &= \frac{1}{q} \left(\frac{1}{\sqrt{2\pi}} \right)^{k/2} |\bar{\boldsymbol{\Omega}}_{\boldsymbol{\theta}}|^{1/2} \times \\
 &\quad \times \exp \left[-\frac{1}{2} (\boldsymbol{\theta}^{(i)} - \bar{\boldsymbol{\theta}})' \bar{\boldsymbol{\Omega}}_{\boldsymbol{\theta}} (\boldsymbol{\theta}^{(i)} - \bar{\boldsymbol{\theta}}) \right] \times \\
 &\quad \times I(\boldsymbol{\theta}^{(i)} \in \Theta_q)
 \end{aligned} \tag{36}$$

where

$$q \in (0, 1) \tag{37}$$

$$\bar{\boldsymbol{\theta}} = \frac{1}{M} \sum_{i=1}^M \boldsymbol{\theta}^{(i)} \tag{38}$$

$$\bar{\boldsymbol{\Omega}}_{\boldsymbol{\theta}} = \frac{1}{M} \sum_{i=1}^M (\boldsymbol{\theta}^{(i)} - \bar{\boldsymbol{\theta}}) (\boldsymbol{\theta}^{(i)} - \bar{\boldsymbol{\theta}})' \tag{39}$$

$$\Theta_q = \left\{ \boldsymbol{\theta} : (\boldsymbol{\theta}^{(i)} - \bar{\boldsymbol{\theta}})' \bar{\boldsymbol{\Omega}}_{\boldsymbol{\theta}} (\boldsymbol{\theta}^{(i)} - \bar{\boldsymbol{\theta}}) < z_{(q)}^{(k)} \right\} \tag{40}$$

and $z_{(q)}^{(k)}$ is q -quantile of χ^2 with k degrees of freedom.

Intuition: truncate domain to area with high probability mass.

Computational problem: using Geweke's modified Gelfand and Dey method we basically fit a Gaussian distribution on the posterior.

Domain of posterior should be unrestricted.

Hence: needed to transform parameters to achieve unrestricted domain.

For instance take following transformations:

$$\omega_i = \ln(h_i) \quad (41)$$

$$\zeta_{ii} = \ln\left(\frac{p_{ii}}{1 - p_{ii}}\right) \quad (42)$$

In other words, if the original parameterisation of the model was in term of γ , a vector comprising constrained parameters, we could reparameterise it in terms of θ where θ contains unconstrained parameters.

Likelihood would be the same

$$p(\mathbf{y}|\theta^{(i)}) = p(\mathbf{y}|\gamma^{(i)}) \quad (43)$$

but prior should be modified taking into account Jacobian of transformation

$$p(\theta^{(i)}) = p(\gamma^{(i)}) \left| \frac{\partial \gamma}{\partial \theta} \right|_{\theta=\theta^{(i)}} \quad (44)$$