Comparison of MCMC algorithms for a logistic regression model

Abstract

This paper reviews recent uses of Markov chain Monte Carlo methods for exploring and summarizing posterior distribution in generalized Bayesian statistic. We consider a non-linear regression model – the Logit model – according to two methods: a generic Metropolis-Hastings algorithm versus slice sampler method. The results presented here are illustrated by different simulations, and show the superior performances of the Hastings-Metropolis algorithm with a bivariate normal proposal in terms of convergence to the stationary distribution and exploration of the posterior distribution surface against slice sampler algorithm.

Key words: Bayesian statistic; Markov chain Monte Carlo; Metropolis-Hastings algorithm; slice sampler; stationarity; convergence assessment; Gibbs sampler; regression.

1 Introduction

This paper focuses on the comparison between some Markov chain Monte Carlo methods (MCMC) which often appear in the treatment of complex statistical models. The majority of MCMC applications, up to now, have as objective the Bayesian inference. The emphasis and terminology will be Bayesian, from a Bayesian point of view, there is no fundamental difference between the observation and the parameter of a statistical model, both considered as a random variable. Thus, if we note \( D \) the data and \( \theta \) the parameter of the model considered plus the latent data, then a formal inference requires the update of the joint distribution \( f(D, \theta) \) on all the variables. Therefore, the determination of \( \pi(\theta) \) and \( f(D|\theta) \) gives \( f(D, \theta) \) by:

\[
f(D, \theta) = f(D|\theta) \pi(\theta)
\]
Bayesian analysis of Logit model: Metropolis-Hastings versus slice sampler

Having observed $D$, we can use the Bayes’s Theorem to determine the distribution of $\theta$ conditionally to the data (i.e. the posterior law)

$$
\pi(\theta|D) = \frac{f(D|\theta)\pi(\theta)}{\int f(D|\theta)\pi(\theta)d\theta}.
$$

(1)

Indeed, for this approach, the inference on the parameters of interest will not directly be based on the likelihood function $f(D|\theta)$, but on the posterior law of the parameters $\pi(\theta|D)$. For Bayesians, all the characteristics of the posterior law are significant for the inference: moment, quantiles, high posterior density, etc. Some of these quantities can be expressed in terms of conditional expectation of a function of $\theta$ with respect to the posterior law:

$$
E[h(\theta)|D] = \frac{\int h(\theta)f(D|\theta)\pi(\theta)d\theta}{\int f(D|\theta)\pi(\theta)d\theta}.
$$

(2)

Moreover, directly generating samples from a posterior distribution $\pi(\theta|D)$, often high dimensional, joint distribution is in general not possible. However, it is necessary to be able to simulate a sample $(\theta_1, \cdots, \theta_n)$ which is approximately i.i.d. from $\pi(\theta|D)$, in order to determine the confidence regions or the general structure of the law (detection of modes, asymmetries, etc). The methods described in this paper overcome this problem by an indirect approach to the required sampling based on Markov chains.

A Markov Chain Monte Carlo algorithms generate an ergodic Markov chain $\theta^{(t)}$ with the target distribution $\pi(\theta|D)$ as the stationary distribution. A MCMC methods then use the fact that, for $t$ large enough, $\theta^{(t)}$ is approximately $\pi(\cdot|D)$ distributed and (2) can be approximated. Asymptotically the ergodic averages of a function of interest over realizations from a single run of the chain provides a consistent estimator of its expectation. The most commonly used MCMC methods are the Metropolis-Hastings algorithm (Hastings, 1970), and the Gibbs sampler (first introduced by Geman and Geman, 1984).

In the next section, we present a generalized non-linear regression model: the Logit model. In the following sections, we detail the necessary conditions for the use of MCMC algorithms and we expose some Markov chain Monte Carlo algorithms. In particular, we present the random walk Metropolis-Hastings algorithm, and a very general version of the Gibbs sampler, called the slice sampler, from a recent point of view introduced by
Bayesian analysis of Logit model: Metropolis-Hastings versus slice sampler

Damien and Walker (1996). Then, we respectively approximate the posterior distribution of the Logit model by using the random walk Metropolis-Hastings algorithm respectively with a bivariate normal proposal resulting from an approximation and a bivariate Cauchy proposal against slice sampler method. In the last section, we apply some diagnostic methods to control convergence of these MCMC algorithms. The example in this paper illustrates the behaviors and the performances of MCMC algorithms in order to simulate the posterior distribution. From a practical point of view, we conclude that the Metropolis-Hastings algorithm with a bivariate normal or bivariate Cauchy approximation is much more effective, in terms of convergence speed towards the stationary distribution and in terms of surface exploration speed of the posterior distribution, than slice sampler method. The whole results exposed in this paper are illustrated by various simulations with “convergence diagnosis and output analysis software for Gibbs sampling output” (CODA) analysis support for convergence control (see Best et al. 1995).

2 The logistic regression Model

A standard qualitative regression model is the Logit model, where the distribution of $y$ conditionally to the explanatory variables $X \in \mathbb{R}^p$ is, for $\gamma \in \mathbb{R}^p$:

$$P(y = 1) = 1 - P(y = 0) = \frac{\exp(X^t \gamma)}{1 + \exp(X^t \gamma)}.$$  

We consider the logistic dependence between the explanatory variables and the observation. Consider the particular case where $X = (1, x)$ and $\gamma = (\alpha, \beta)$. The binary variables $y_i$'s in $\{0, 1\}$ are associated with the explanatory variables $x_i$'s and modelled following a Bernoulli law of conditional probability:

$$y_i | x_i \sim B \left( \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \right). \quad (3)$$

The Bayesian approach when there is no prior information on the parameters $\alpha$ and $\beta$ is a “generalized” Bayesian approach where the prior distribution is improper. Suppose that our parameters follow an improper prior $\pi(\alpha, \beta) = 1$, in the sense that it is a $\sigma$-finite measure instead of being a regular probability measure. The likelihood of our model, for
Bayesian analysis of Logit model: Metropolis-Hastings versus slice sampler

a sample \((y_1, x_1), \cdots, (y_n, x_n)\) is equal to:

\[
f(y_1, \cdots, y_n | x_1, \cdots, x_n, \alpha, \beta) = \prod_{i=1}^{n} \frac{\exp[(\alpha + \beta x_i)y_i]}{1 + \exp(\alpha + \beta x_i)}.
\]

The posterior law of \((\alpha, \beta)\) results then by a formal application of the Bayes’s theorem:

\[
\pi(\alpha, \beta \mid D) \propto f(y_1, \cdots, y_n | x_1, \cdots, x_n, \alpha, \beta)\pi(\alpha, \beta) \\
\propto \prod_{i=1}^{n} \frac{\exp[(\alpha + \beta x_i)y_i]}{1 + \exp(\alpha + \beta x_i)} = \frac{\exp[\sum_{i=1}^{n} (\alpha + \beta x_i)y_i]}{\prod_{i=1}^{n} [1 + \exp(\alpha + \beta x_i)]}.
\]

3 Condition to use MCMC algorithms

The use of non-informative prior laws, i.e. a \(\sigma\)-finite measure of infinite mass on the parameters’ space, implies that the posterior law derivation using proportionality relation, as \(\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)\), is not necessarily acceptable to implement a Metropolis-Hastings algorithm on \(f(x \mid \theta)\pi(\theta)\), because the corresponding law may not exist, i.e. \(f(x \mid \theta)\pi(\theta)\) is not necessarily integrable. We meet the same difficulty for the Gibbs sampling which, contrary to the Metropolis-Hastings algorithm, functions with conditional laws extracted from \(\pi(\theta_1, \cdots, \theta_q)\) which itself is represented by the above proportionality relation. It may happen that these laws are clearly definite and simulable, but do not correspond to a joint law \(\pi\), i.e. \(\pi\) is not integrable (see Robert 1996 for examples). This fact, rather frequent in a generalized Bayesian approach does not represent a defect for the MCMC algorithms, nor even a simulation problem. We should not however omit the \(\pi\)-existence verification, that we prove in the lemma below. The following hypothesis is introduced:

Hypothesis [H]

Given a sample \((x_1, y_1), \cdots, (x_n, y_n)\) with \(n \geq 4\). We suppose there exists positive \(x_i\) and negative \(x_i\) associated to both of \(y_i = 1\) and \(y_i = 0\).

Lemma 1 The posterior distribution of the Logit model \(\pi(\alpha, \beta \mid D)\), where \(D\) represents
the observed data, is a true law under [H], i.e.
\[
\int \int \pi(\alpha, \beta \mid D) d\alpha d\beta < +\infty.
\]

**Proof:** Define
\[
I = \int \int \prod_{i=1}^{n} \frac{\exp[(\alpha + \beta x_i) y_i]}{[1 + \exp(\alpha + \beta x_i)]} d\alpha d\beta.
\]

Given the index set \{i_1, \ldots, i_p\} for which \(y_i = 1\), then the integral is written:
\[
\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{\exp(\alpha p) \exp(\beta x_0)}{\prod_{i=1}^{n} (1 + \exp(\alpha + \beta x_i))} d\alpha d\beta = I_1 + I_2 + I_3 + I_4,
\]
with \(x_0 = \sum_{i=1}^{n} x_i y_i\) and
\[
I_1 = \int_{-\infty}^{0} \int_{-\infty}^{0} \frac{\exp(\alpha p) \exp(\beta x_0)}{\prod_{i=1}^{n} (1 + \exp(\alpha + \beta x_i))} d\alpha d\beta, \quad I_2 = \int_{-\infty}^{0} \int_{0}^{+\infty} \frac{\exp(\alpha p) \exp(\beta x_0)}{\prod_{i=1}^{n} (1 + \exp(\alpha + \beta x_i))} d\alpha d\beta,
\]
\[
I_3 = \int_{0}^{+\infty} \int_{-\infty}^{0} \frac{\exp(\alpha p) \exp(\beta x_0)}{\prod_{i=1}^{n} (1 + \exp(\alpha + \beta x_i))} d\alpha d\beta, \quad I_4 = \int_{0}^{+\infty} \int_{0}^{+\infty} \frac{\exp(\alpha p) \exp(\beta x_0)}{\prod_{i=1}^{n} (1 + \exp(\alpha + \beta x_i))} d\alpha d\beta.
\]

The denominator is equal to
\[
\prod_{i=1}^{n} (1 + \exp(\alpha + \beta x_i)) = \exp(n\alpha) \exp(\beta \sum_{i=1}^{n} x_i) + \cdots + \exp((p + 1)\alpha) \left[ \sum_{\sigma \in W_{p+1}} \exp(\beta \sum_{j=1}^{p+1} x_{\sigma(j)}) \right] + \exp(p\alpha) \left[ \sum_{\sigma \in W_p} \exp(\beta \sum_{j=1}^{p} x_{\sigma(j)}) \right] + \exp((p - 1)\alpha) \left[ \sum_{\sigma \in W_{p-1}} \exp(\beta \sum_{j=1}^{p-1} x_{\sigma(j)}) \right] + \cdots + 1,
\]
where \(W_q = \{ \text{all the injections } \sigma : \{1, \ldots, q\} \to \{1, \ldots, n\} \}\).

For the first and third integrals, we keep only in the denominator the terms which have \(\exp(\alpha(p - 1))\) as common factor, hence for \(I_1\) we have
\[ I_1 \leq \int_{-\infty}^{0} \int_{-\infty}^{0} \frac{\exp(\alpha p) \exp(\beta x_0)}{\exp(\alpha(p-1)) \left[ \sum_{\sigma \in W_{p-1}} \exp(\beta \sum_{j=1}^{p-1} x_{\sigma(j)}) \right]} \, d\alpha d\beta \\
= \int_{-\infty}^{0} \frac{\exp(\beta x_0)}{\left[ \sum_{\sigma \in W_{p-1}} \exp(\beta \sum_{j=1}^{p-1} x_{\sigma(j)}) \right]} \, d\beta. \]

However:

\[ \sum_{i=1}^{n} x_i y_i = \sum_{k=1}^{p} x_{ik}, \]

where \((i_1, i_2, \ldots, i_p)\) are the indices of \(y_{ik}\) for which \(y_{ik} = 1\). By \([\text{H}]\), there exists \(k \in (i_1, i_2, \ldots, i_p)\) such that \(y_k = 1\) and \(x_k > 0\). Given, without loss of generality, \(i_p\) such that \(y_{ip} = 1\) and \(x_{ip} > 0\), and \(\tilde{\sigma}\) such as: \(\tilde{\sigma}(k) = i_k\) with \(k = 1, 2, \ldots, p - 1\).

We have \(\tilde{\sigma} \in W_{p-1}\), therefore we can bound the integral by:

\[ I_1 \leq \int_{-\infty}^{0} \frac{\exp(\beta \sum_{k=1}^{p} x_{ik})}{\exp(\beta \sum_{k=1}^{p-1} x_{\tilde{\sigma}(k)})} \, d\beta = \int_{-\infty}^{0} \frac{\exp(\beta \sum_{k=1}^{p} x_{ik})}{\exp(\beta \sum_{k=1}^{p-1} x_{\tilde{\sigma}(k)})} \, d\beta \]

\[ = \int_{-\infty}^{0} \exp(\beta x_{ip}) \, d\beta < \infty. \]

For the second and fourth integrals, we keep the terms which have common factor \(\exp(\alpha(p+1))\), hence the expression

\[ I_2 \leq \int_{-\infty}^{0} \int_{0}^{+\infty} \frac{\exp(\alpha p) \exp(\beta x_0)}{\exp(\alpha(p+1)) \left[ \sum_{\sigma \in W_{p+1}} \exp(\beta \sum_{j=1}^{p+1} x_{\sigma(j)}) \right]} \, d\alpha d\beta \\
= \int_{-\infty}^{0} \frac{\exp(\beta x_0)}{\left[ \sum_{\sigma \in W_{p+1}} \exp(\beta \sum_{j=1}^{p+1} x_{\sigma(j)}) \right]} \, d\beta \]

By \([\text{H}]\), there exists \(i_{p+1}\) such as \(y_{ip+1} = 0\) and \(x_{ip+1} < 0\). Given \(\tilde{\sigma}(k) = i_k\) with \(k = 1, 2, \ldots, p + 1\). We have \(\tilde{\sigma} \in W_{p+1}\), therefore we have:

\[ I_2 \leq \int_{-\infty}^{0} \frac{\exp(\beta \sum_{k=1}^{p} x_{ik})}{\exp(\beta \sum_{j=1}^{p+1} x_{\tilde{\sigma}(k)})} \, d\beta = \int_{-\infty}^{0} \frac{\exp(\beta \sum_{k=1}^{p} x_{ik})}{\exp(\beta \sum_{j=1}^{p+1} x_{\tilde{\sigma}(k)})} \, d\beta \]
$$= \int_{-\infty}^{0} \exp(-Bx_{p+1})d\beta < +\infty.$$ 

And it goes in the same way for the other integrals.

4 The Metropolis-Hastings algorithm

The generic Metropolis-Hastings algorithm is based on the use of a conditional density $q(y|x)$ with respect to the dominant measure for the model. It can be put in practice only if $q(.|x)$ can be quickly simulated and is, either analytically available up to a constant independent of $x$, or symmetrical, i.e. such as $q(y|x) = q(x|y)$. The Metropolis-Hastings algorithm associated with the objective law $\pi$ and the conditional law $q$ produces a Markov chain $(x^{(t)})$ based on the following transition:

Given $x^{(t)}$

1. Generate 

$$y_t \sim q(y|x^{(t)}),$$

2. Take 

$$x^{(t+1)} = \begin{cases} 
y_t & \text{with probability } \rho(x^{(t)}, y_t) \\
x^{(t)} & \text{with probability } 1 - \rho(x^{(t)}, y_t) 
\end{cases}$$

where

$$\rho(x^{(t)}, y_t) = \min \left\{ \frac{\pi(y_t) q(x^{(t)}|y_t)}{\pi(x^{(t)}) q(y_t|x^{(t)})}, 1 \right\}.$$ 

The law $q$ will be called proposal law. This algorithm systematically accepts simulations $y_t$ such that the ratio $\pi(y_t)/q(y_t|x^{(t)})$ is higher than the preceding value $\pi(x^{(t)})/q(x^{(t)}|y_t)$. It is only in the symmetrical case that the acceptance is controlled by the ratio $\pi(y_t)/\pi(x^{t})$.

4.1 Random walk Metropolis-Hastings algorithm

A particular case of the Metropolis-Hastings algorithm is the random walk algorithm, for which $q(y|x) = g(|y - x|)$. For example, when $x$ is continuous, $q(.|x)$ can be a multivariate normal distribution of mean $x$ and with a constant variance-covariance matrix $\Sigma$. A
Bayesian analysis of Logit model: Metropolis-Hastings versus slice sampler

careful choice of the proposal distribution results in generating a small step $y - x_t$ which generally gives a high rate of acceptance and a slowly mixing chain. A bad choice of the proposal distribution leads to an excessive step which provides movements of the center to the tail of the distribution, and in general produces small values of $\pi(y)/\pi(x_t)$ and a small rate of acceptance. Such a chain also leads to a slow mixing. The ideal solution to avoid these two cases is to use a scale parameter for the proposal law. The algorithm [A1] authorizing this dependence, $q(y|x)$ can thus be of the form $g_\tau(|y - x|)$, i.e. $y_t$ can be written in the form $x_t + \tau \xi_t$, $\xi_t$ being a random perturbation of distribution $g$, independent of $x_t$, and $\tau$ being a scale parameter. The Markov chain associated at $q_\tau$ is then a random walk on $X$. Note that the choice of $g$ as a symmetrical function $g(-t) = g(t)$ allows to express the algorithm [A1] in the following form:

1. Generate $y_t \sim g_\tau(y - x^{(t)})$.
2. Take $x^{(t+1)} = \begin{cases} y_t & \text{with probability } \min\left\{1, \frac{\pi(y_t)}{\pi(x^{(t)})}\right\} \\ x^{(t)} & \text{otherwise} \end{cases}$ [A2]

Compared to other algorithm, the random walk Metropolis-Hastings algorithm requires a specific rate of acceptance analysis, because of the dependence of the proposal law on the value previously accepted. A high rate of acceptance does not state that the algorithm evolves correctly. Conversely, if the average rate of acceptance is weak, the successive values of $\pi(y_t)$ are frequently small compared to $\pi(x_t)$, i.e. the random walk moves quickly on the surface of $\pi$ (but can visit too much the tails of $\pi$). An automatic parameterization method cannot guarantee optimal performances for the random walk Metropolis-Hastings algorithm, and the choices of rates operated here do not inevitably lead to optimality. Gelman, Gilks and Roberts (1996), recommend rates of acceptance close to 50% for models of dimension 1 or 2.
5 Two strategies

We present in this section two schemes based on the Metropolis-Hastings algorithm versus slice sampler that will be used on the simulation study.

5.1 A Normal-Based approximation

Let $\hat{\theta}$ be the maximum likelihood estimator (MLE) of $\theta$ computed from the data $y$. We have the following asymptotic distribution:

$$\sqrt{n}(\theta - \hat{\theta}) \approx \mathcal{N}(0, C),$$

where $C = I^{-1}(\theta)$ is the Fisher information for an observation,

$$I(\theta|y) = E \left[ -\frac{\partial^2 I(\theta|y)}{\partial \theta \partial \theta^t} \right]$$

(in the i.i.d. case). For Bayesians, $\hat{\theta}$ is fixed conditionally on the data $y$ and $\theta$ is the variable. Knowing the model and the data, (5) implies that the posterior density of $\theta$ is asymptotically normal of average $\hat{\theta}$, and of variance-covariance matrix $C$. The Bayesian justification of (5) comes from the Taylor expansion of the loglikelihood of the posterior density about the fixed value $\hat{\theta}$:

$$l(\theta|y) = l(\hat{\theta}|y) + (\theta - \hat{\theta})^t S(\hat{\theta}|y) - \frac{1}{2} (\theta - \hat{\theta}) I(\hat{\theta}|y)(\theta - \hat{\theta})^t + r(\theta|y),$$

with

$$I(\hat{\theta}|y) = -\frac{\partial^2 I(\theta|y)}{\partial \theta \partial \theta^t}|_{\hat{\theta}}.$$  

Note that $E[I(\theta|y)]$ is the Fisher information matrix. Since $S(\hat{\theta}|y) = 0$, by assuming that $r(\theta|y)$ can be neglected, the posterior density is proportional to the multivariate normal density of average $\hat{\theta}$ and of variance-covariance matrix $C = I^{-1}(\theta|y)$ (see Tanner, 1993).

5.1.1 Application to the Logit model: bivariate normal approximation

We reconsider the Logit model in order to approach the posterior law of $(\alpha, \beta)$ by implementing the random walk Metropolis-Hastings algorithm with a bivariate normal proposal:

$$\left( \begin{array}{c} \alpha^{(t+1)} \\ \beta^{(t+1)} \end{array} \right) \sim \mathcal{N}_2 \left( \left( \begin{array}{c} \alpha^{(t)} \\ \beta^{(t)} \end{array} \right), \Sigma^{(t)} \right).$$
To calculate $\Sigma^{(t)}$, we take the Taylor expansion of the logarithm of the target law about the value $(\hat{\alpha}, \hat{\beta})$, which is given by:

$$\log \pi(\alpha, \beta) = \log \pi(\hat{\alpha}, \hat{\beta}) + \frac{1}{2} \left( \alpha - \hat{\alpha}, \beta - \hat{\beta} \right) \nabla \nabla^T \log \pi(\hat{\alpha}, \hat{\beta}) \left( \alpha - \hat{\alpha}, \beta - \hat{\beta} \right)^T.$$ 

Then we replace $E \left[ \nabla \nabla^T \log \pi(\alpha, \beta) \right]$ by its observation. That implies the calculation of:

$$\nabla \log \pi(\alpha, \beta) = \left( \sum_{i=1}^{n} \frac{y_i - \sum_{i=1}^{n} \frac{\exp(\alpha + \beta z_i)}{1 + \exp(\alpha + \beta z_i)}}{\sum_{i=1}^{n} \frac{\exp(\alpha + \beta z_i)}{1 + \exp(\alpha + \beta z_i)}} , \sum_{i=1}^{n} \frac{\exp(\alpha + \beta z_i)}{1 + \exp(\alpha + \beta z_i)} \right),$$

$$\nabla \nabla^T \log \pi(\alpha, \beta) = -\left( \sum_{i=1}^{n} \frac{\exp(\alpha + \beta z_i)}{1 + \exp(\alpha + \beta z_i)} , \sum_{i=1}^{n} \frac{\exp(\alpha + \beta z_i)}{1 + \exp(\alpha + \beta z_i)} \right)^2 \sum_{i=1}^{n} \frac{\exp(\alpha + \beta z_i)}{1 + \exp(\alpha + \beta z_i)} .$$

Moreover, we consider the matrix $\Sigma^{(t)}$ adjusted using a scale factor $\tau$:

$$\Sigma^{(t)} = \tau^2 \left( \sum_{i=1}^{n} \frac{\exp(\alpha + \beta z_i)}{1 + \exp(\alpha + \beta z_i)} , \sum_{i=1}^{n} \frac{\exp(\alpha + \beta z_i)}{1 + \exp(\alpha + \beta z_i)} \right)^{-1} \sum_{i=1}^{n} \frac{\exp(\alpha + \beta z_i)}{1 + \exp(\alpha + \beta z_i)} .$$

The subjacent idea is to calibrate the factor $\tau$ according to the rate of acceptance of the algorithm (see Robert, 1996b). Lastly by substituting $(\alpha, \beta)$ with $(\alpha^t, \beta^t)$, maximum likelihood, we obtain $\Sigma^{(t)}$. Since the proposal law in this case is the bivariate normal $\mathcal{N}(0, \Sigma)$, then the algorithm application can respectively pass by the generation of the marginal and the conditional laws:

$$X_1 \sim \mathcal{N}(0, \sigma_1^2), \quad X_2 X_1 \sim \mathcal{N}(\rho \sigma_2 \sigma_1, \sigma_2^2 (1 - \rho^2)).$$

5.1.2 Application to the Logit model: bivariate Cauchy approximation

The Cauchy law has much thicker tails than the normal law. We reconsider our Logit model to approach the posterior law of $(\alpha, \beta)$ by the random walk Metropolis-Hastings algorithm with

$$\left( \begin{array}{c} \alpha^{(t+1)} \\ \beta^{(t+1)} \end{array} \right) = \mathcal{C}_2 \left( \left( \begin{array}{c} \alpha^{(t)} \\ \beta^{(t)} \end{array} \right) , \Sigma^{(t)} \right),$$

where $\Sigma^{(t)} = I^{-1}(\theta)$ is the Fisher information up to a calibration constant as already explained in the preceding section.
If $X \sim C_2(0, \Sigma)$, then we can transform it into $Y = \Sigma^{-\frac{1}{2}}X \sim C_2(0, I)$, and the density of $Y$ is given by:

$$f(y_1, y_2) = \frac{1}{2\pi (1 + y_1^2 + y_2^2)^{3/2}}.$$ 

The application of the random walk Metropolis-Hastings algorithm with a proposal $C_2(0, I)$, can respectively pass by the generation of the marginal and the conditional laws:

$$Y_1 \sim \frac{1}{\pi (1 + y_1^2)}, \quad Y_2 \mid Y_1 \sim \frac{1 + y_1^2}{2 (1 + y_1^2 + y_2^2)^{3/2}}.$$ 

Finally, to simulate $X \sim C_2(0, \Sigma)$ we pass by the linear transformation $X = \Sigma^{\frac{1}{2}}Y$, with $Y \sim C_2(0, I)$:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \Sigma^{1/2} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$ 

5.2 Damien and Walker method

5.2.1 Principle

Let $\pi$ be a continuous density function definite at $\mathbb{R}^d$. We address the problem of generating a random variate $X$ from $\pi$. The basic idea of Damien and Walker (1996) is to introduce a auxiliary variable $Y$ in order to build the joint density $g(x, y)$ of $(X, Y)$ with marginal density for $X$ given by $\pi$,

$$\pi(x) = \int g(x, y) dy.$$ 

A Gibbs sampler can now be implemented (see Smith and Roberts, 1993) to generate random variates from $g$ if obviously the full conditionals $g(y|x), g(x|y)$ are of known types. The Damien and Walker method in general leads to easily simulable conditional distributions (Besag and Green, 1993). The iterative algorithm is based on the:

1. Simulate $Y^{t+1} \sim f(y|X = x^t)$,
2. Simulate $X^{t+1} \sim f(x|Y = y^{t+1})$. [A3]

By applying the Gibbs sampling algorithm [A3], we generate an ergodic Markov chain $(X^t, Y^t)$ from a transition kernel of stationary law $g(x, y)$, which moreover guarantees
convergence in law towards \( g \) for \( T \) rather large. We can then consider \( X^T \sim \pi \).

**Theorem 1 (Damien and Walker, 1996)** If

\[
\pi(x) \propto f_0(x) \prod_{i=1}^{n} f_i(x),
\]

where the \( f_i(x) \) are nonnegative invertible functions (not necessarily densities), that is, if \( f_i(x) > u \), then it is possible to obtain the set \( A_k(u) = \{ x : f_i(x) > u \} \), and \( f_0(x) \) is a density of a known type. Then we can write \( \pi \) as follows:

\[
\pi(x) \propto f_0(x) \prod_{i=1}^{n} \int 1_{\{u \leq f_i(x)\}} du_i
\]

**Proof:** We introduce the auxiliary variables \((u_1, \ldots, u_n)\), defined at \((0, \infty)\) such that the joint density of this form:

\[
g(x, u_1, \ldots, u_n) \propto f_0(x) \prod_{i=1}^{n} 1_{\{u_i \leq f_i(x)\}}.
\]

Clearly the marginal density of \( x \) is \( \pi \), and obviously the full conditionals for each \( u_i \) is uniform density in the interval \((0, f_i(x))\). In the same way, the full conditional for \( x \) is given by \( \pi(x) \) restricted to the set \( A(u) = \{ x : u_i < f_i(x); i = 1, \ldots, n \} \). Consequently, the Gibbs sampling could be easily applied.

### 5.2.2 Application to the Logit model: Damien and Walker method

To simulate the posterior law of the Logit model following the Damien and Walker method, we notice, by a simple application of the preceding theorem, that the full conditional laws are available and allow to approach the posterior distribution by the general version of the Gibbs sampler, called the slice sampler (see Damien et al. (1999), Wakefield et al. (1999), and Tierney and Mira 1999). We introduce the \( n \) independent random variables \( w_1, \ldots, w_n \) where the auxiliary variables \( w_i \) are defined on \((0, 1)\), such that their joint density with \( \alpha \) and \( \beta \) is given by:

\[
\pi_1(\alpha, \beta, w_1, \ldots, w_n \mid D) \propto \prod_{i=1}^{n} 1 \left( \frac{\exp(\alpha + \beta x_i) y_i}{1 + \exp(\alpha + \beta x_i)} \right).
\]
A direct Gibbs sampling implementation in this context leads to simulate \( \pi_1(\alpha, \beta, w \mid D) \).
The generation of this model consists of the following stages:

where

\[
\rho_i(\alpha, \beta) = \frac{\exp(\alpha + \beta x_i) y_i}{1 + \exp(\alpha + \beta x_i)}
\]

And

\[
\alpha, \beta \mid w_1, \cdots, w_n \sim \mathcal{U}_{\{(\alpha, \beta) : w_i \leq \rho_i(\alpha, \beta)\}},
\]

or

\[
\alpha, \beta \mid w_1, \cdots, w_n \sim \mathcal{U}_{\left\{\left\{\cap_{i=1} \{\alpha, \beta : \alpha + \beta x_i \geq \log\left(\frac{w_i}{1 - w_i}\right)\}\right\} \cap \left\{\cap_{i=0} \{\alpha, \beta : \alpha \leq \log\left(\frac{1 - w_i}{w_i}\right) - \beta x_i\}\right\}\right\}}.
\]

The Slice Sampler Algorithm is written as:

Simulate

1. \( w_1 \mid \alpha, \beta \sim \mathcal{U}_{\{0, \rho_1(\alpha, \beta)\}}; \)
2. \( w_n \mid \alpha, \beta \sim \mathcal{U}_{\{0, \rho_n(\alpha, \beta)\}}; \)

\[
\begin{align*}
n+1. \alpha \mid \beta, w_1, \cdots, w_n & \sim \mathcal{U}_{\left\{\left\{\cap_{i=1} \{\alpha : \alpha \geq \log\left(\frac{w_i}{1 - w_i}\right) - \beta x_i\}\}\right\} \cap \left\{\cap_{i=0} \{\alpha : \alpha \leq \log\left(\frac{1 - w_i}{w_i}\right) - \beta x_i\}\right\}\right\}; \quad [44] \\
n+2. \beta \mid \alpha, w_1, \cdots, w_n & \sim \mathcal{U}_{\left\{\left\{\cap_{i=1, x_i > 0} \{\beta : \beta > \frac{\log\left(\frac{w_i}{1 - w_i}\right) - \alpha}{x_i}\}\} \cap_{i=0, x_i < 0} \{\beta : \beta > \frac{\log\left(\frac{1 - w_i}{w_i}\right) - \alpha}{x_i}\}\right\} \cup \left\{\cap_{i=1, x_i < 0} \{\beta : \beta < \frac{\log\left(\frac{w_i}{1 - w_i}\right) - \alpha}{x_i}\}\} \cup_{i=0, x_i > 0} \{\beta : \beta < \frac{\log\left(\frac{1 - w_i}{w_i}\right) - \alpha}{x_i}\}\right\}.
\end{align*}
\]

Note, Roberts and Rosenthal (1998) studied the slice sampler and showed that it usually enjoys good theoretical properties. In particular, they proved that the slice sampler allows small sets and the associated chain is geometrically ergodic when the density \( \pi \) is bounded. Tierney and Mira (1999) have also established uniform ergodicity when \( n = 1 \). As \( n \)
increases, the determination of the set $A(u) = \{x : u_i < f_i(x); i = 1, \cdots, n\}$ may get increasingly complex, and this may create convergence slowdowns. Also, the Gibbs sampler is forced to work in a highly restricted space when $n$ is large.

6 Diagnosing Convergence

Even if MCMC methods are applicable to a vast class of models, they suffer from a significant problem in the applicability: how to determine the moment when we can conclude their convergence, in other words, when should one stop the chain and use the observations in order to estimate the distribution characteristics considering that the sample is sufficiently representative of the stationary distribution. We have for the moment few methods which have practical results to handle this convergence control problem.

In this paper, we have applied MCMC algorithms. Under minimal regularity conditions, these algorithms converge towards the target distribution and, under mild constraining conditions, this convergence occurs at geometrical speed. However, the practical implementation of these algorithms still poses many problems, at the same time from a purely algorithmic point of view (how many simulations is it necessary to perform? Does one have to “warm-up” the algorithm before taking simulations into account? Is it preferable to start from several initial values? Which algorithm converges most quickly? etc), as from a statistical point of view (up to which point can we use the sequence $\theta^{(t)}$ as a sample of $\pi$? How to compare the samplings in terms of quadratic risk? Does we have to condition on the nuisance parameters? etc). In practice, note that the available diagnosis try to check necessary conditions which are not sufficient to ensure the convergence (see Cowles and Carlin 1996, Brooks and Roberts 1999. As noted in Cowles and Carlin (1996), it is not unfortunately possible, even in recent works, to think of proposing automatic stopping criteria.
6.1 Convergence control criteria

According to Cowles and Carlin (1996) and Brooks and Roberts (1998), we can distinguish three convergence degrees for which a control is necessary. The first one decides if the variables $\theta^{(t)}$ are distributed according to the stationary distribution $\pi$. For the second, it should be noted that even if $\theta^{(t)} \sim \pi$, the exploration of the complexity of the support of $\pi$ from the chain $(\theta^{(t)})$ may requires many iterations. We must thus make sure that the chain has updated well all specificity of $\pi$, as the set of modes. Brooks and Roberts (1995) associate this convergence to the chain mixing speed, in the informal sense of a more or less fast exploration of the support of $\pi$. It is, in particular, a question of determining the minimal value of $T$ authorizing the approximation of $E_{\pi}[h(\theta)]$ by the classical Monte Carlo estimator:

$$
\delta_T = \frac{1}{T} \sum_{t=1}^{T} h(\theta^{(t)}). \tag{6}
$$

Although (6) converges almost surely towards the posterior expectation $E_{\pi}[h(\theta)]$ when $T$ goes to $+\infty$, it is obviously preferable to be able to control the convergence speed or in an equivalent manner the precision of the approximation $E_{\pi}[h(\theta)]$ from (6) by means of a central limit theorem

$$
\sqrt{T}(\delta_T - E_{\pi}[h(\theta)]) \approx \mathcal{N}(0, \sigma^2_\pi).
$$

The methods of parallel chains and “batch sampling” are of the third convergence type, to guarantee the independence or quasi-independence of the simulated variables (see Mengersen, Robert, and Guillaume-Jouaux 1999).

6.2 Graphical and empirical methods

A natural convergence control approach is to graphically examine the behavior of a simulated chain. We can thus trace the sequence of values $\theta^{(t)}$ depending on $t$, as done by many authors (see figure 1), but this approach allows in maximum to detect non-strong stationarities (see Gelfand and Smith, 1990). In general, the graph of the series $(\theta^{(t)})$ is very ineffective to detect the stationarity or the convergence. Depending on the quantities of interest $E_{\pi}[h(\theta)]$, we generally prefers the evolution of the averages cumulated
(6) according to $T$ as we did below (see figure 5). A necessary convergence condition is the stationarity of the sequence $\delta_T$, even if an apparent stabilization of this sequence correspond to exploration a single-mode of the density $\pi$. Robert (1996a) proposes a more robust approach of graphical convergence evolution: it consists in employing the chain $(\theta^{(t)})$ through several estimation procedures, all converge, until these various estimations coincide. The selected estimation techniques are, in addition to the empirical average $\delta_T$, the conditional version of this average (Rao-Blackwellisation), either on its nonparametric form for the Metropolis-Hastings algorithm, or in parametric form for the Gibbs sampler

$$
\delta_T^C = \frac{1}{T} \sum_{t=1}^{T} E \left[ h(\theta) | \eta^{(t)} \right].
$$

(7)

### 6.3 Application to the Logit model

Within the framework of the Logit model, we consider the explanatory variables $x_i$ distributed following the exponential law of parameter $\lambda = 1$ and we observe this model with $\alpha = -5$, $\beta = 8$. For the data creation of this model, we consider the data size $n = 1000$.

#### 6.3.1 How to compare the two methods?

We need to precise how we run both algorithms and on which basis we compare those, from the fact that both algorithms in essence is a single chain method. This is actually the reason why we have chosen to compare two single chains issued from each methods, keeping in mind that the very nature of the Metropolis-Hastings requires an exploratory stage (to be understand as burn-in). The random walk Metropolis-Hastings essentially uses these burn-in iterations to escape from the starting position (i.e. to obtain values that do not depend on the starting point), and (presumably) to obtain a chain approximately $\pi$ distributed (this is what burn-in is usually for). In this example, we decided to compare both methods on the basis of one single chain of length 50,000 issued from each algorithm including the burn-in iterations for the Metropolis-Hastings and the rates of acceptance operated here close to 50%. This is “fair” from the duration of the resulting chains point of view, but may be unfair to random walk Metropolis-Hastings algorithm, since it does
not take the benefit of its burn-in stage. What is interesting, we show that without this exploration stage which reduces its implementation cost, the Metropolis-Hastings version produces a really better chain (this maybe thanks to a good choice of the starting points) as shown by the simulations and convergence diagnosis.

The results of this experience show clearly the defects of slice sampler method. Indeed, from a practical point of view, slice sampler method needs much more iterations in order to explore the support and to arrive at convergence properties close to those of the Metropolis-Hastings algorithm. That comes from the number of the auxiliary variables to simulate, i.e. more the size of data increases, more there will be auxiliary variables to simulate. In fact, in slice sampler if highly correlated scalar components are treated individually, there could be painfully slow convergence of the chain to the stationarity distribution as a result of very little movement at each conditional random variate generation steps. Which increase the calculating time since decreases the convergence speed and the surface exploration speed of the posterior distribution.

Conversely, for the Metropolis-Hastings algorithm, the increase in the number of data allows a fast exploration of the set of modes of the objective law and a good convergence speed towards the stationary distribution, with less iterations. Of course both of the methods lead to the same posterior distribution as shown in the figures 1 and 2.
Figure 1: Trace and density comparison of the parameters $\alpha, \beta$ following the Metropolis-Hastings algorithm with a bivariate normal (chain 1), a bivariate Cauchy (chain 2) and slice sampler (chain 3). The illustration is made for the 5000 last iterations of the chains.

Figure 2: (left) Displacement comparison of the chain $(\alpha^{(t)}, \beta^{(t)})$ on the contour of the posterior law for the 100 last iterations of the chain. (right) Histogram comparison.
Figure 3 gives the autocorrelations for the parameters $\alpha, \beta$ of the Logit model, according to slice sampler method, and the Metropolis-Hastings algorithm with respectively a bivariate normal and a bivariate Cauchy approximation. We see that the first method provides a high autocorrelation. This indicates a very weak mixing and induces a slow convergence for this algorithm, characterized in figure 3 by a slow topdown oscillation. On the contrary, the Metropolis-Hastings algorithm with a bivariate normal or Cauchy proposal induces weak autocorrelations but approximately of the same order, i.e. a quick convergence speed, indicated in figure 3 by a rapid topdown oscillation. It seems obvious that the very strong autocorrelation for the Gibbs sampler method considerably limits the displacement of the chain on the contour of the posterior distribution (see figure 2 left), and this undoubtedly explains the need for a large iteration number to achieve convergence.

However, we observe that the Metropolis-Hastings with a Cauchy proposal, has the same advantage as that associated with a normal proposal with regard to the exploration of the support of the posterior distribution and we can roughly arrive at the same convergence property. Moreover, we note that the displacement of the chain $(\alpha^{(t)}, \beta^{(t)})$ on the support is satisfactory and is approximately of the same order as the displacement of the Metropolis-Hastings algorithm for a normal law. Note also, that the displacement entirely exceeds the support, which is not the case for the same algorithm with a bivariate normal proposal. That undoubtedly comes owing to the fact that the Cauchy law has thicker tails than the normal law, which enables it to explore less in detail the support of the posterior distribution (see figure 2 left).
From the Geweke (1992) point of view to convergence control of MCMC algorithms, we observe in figure 4 that most of the values of its statistics \( Z_n \) for the Gibbs sampler algorithm are out of the interval \( \pm 1.96 \) of the reduced centered normal distribution \( N(0, 1) \), meaning the failure of the convergence test (they must be inside this interval in the case of convergence). In contrast, for Metropolis-Hastings algorithm, the majority of the values are in the interval \( \pm 1.96 \), which supposes a strong possibility of convergence.
Figure 4: Geweke convergence diagnosis comparison for the Logit model according to Metropolis-Hastings with a bivariate normal proposal (chain 1), Metropolis-Hastings with a bivariate Cauchy proposal (chain 2) against slice sampler method (chain 3).

In addition to the two preceding methods of diagnosis (autocorrelation and Geweke), which encourage the rejection of slice sampler method, the effectiveness of the Metropolis-Hastings algorithm, associated with the normal or Cauchy approximation, is confirmed by the diagnosis of Raftery and Lewis (1992a,b). Table 1 gives the evolution of $k$, minimum sampling step, $t_0$, number of minimum iterations necessary to obtain the stationarity and $T$, a total number of iterations ensuring convergence. Indeed, Table 1 indicates that the slice sampler method, with a test chain of 50,000 iterations, we need 607,350 (307,360) as the total number of iterations, 3746 as the minimum numbers of iterations, 570 (260) initial iterations to reject and a step batch size of 30 (20) for the parameters $\alpha$ ($\beta$) respectively. While that of Metropolis-Hastings with a normal proposal, requires a step of 4 (2) with a rejection proportion of 4% (6%) and requires a total number of iterations 4.5% (6.1%) of that of slice sampler with an identical minimum numbers of iterations respectively for the parameter $\alpha$ and ($\beta$). As said before, the illustrations are based on 50,000 iterations obtained by the CODA software (see Robert, 1996, Chap.8).
Table 1: Table of Raftery and Lewis diagnosis corresponding to this experiment.

<table>
<thead>
<tr>
<th>Chain</th>
<th>Variable</th>
<th>Thin ((k))</th>
<th>Burn-in ((t_0))</th>
<th>Total ((T))</th>
<th>Lower bound ((T_{\text{min}}))</th>
<th>Dependence factor ((I))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain 1</td>
<td>(\alpha)</td>
<td>4</td>
<td>24</td>
<td>27700</td>
<td>3746</td>
<td>7.39</td>
</tr>
<tr>
<td></td>
<td>(\beta)</td>
<td>2</td>
<td>18</td>
<td>18614</td>
<td>3746</td>
<td>4.97</td>
</tr>
<tr>
<td>Chain 2</td>
<td>(\alpha)</td>
<td>14</td>
<td>98</td>
<td>113078</td>
<td>3746</td>
<td>30.2</td>
</tr>
<tr>
<td></td>
<td>(\beta)</td>
<td>10</td>
<td>90</td>
<td>97950</td>
<td>3746</td>
<td>26.1</td>
</tr>
<tr>
<td>Chain 3</td>
<td>(\alpha)</td>
<td>30</td>
<td>570</td>
<td>607350</td>
<td>3746</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td>(\beta)</td>
<td>20</td>
<td>260</td>
<td>307360</td>
<td>3746</td>
<td>82.1</td>
</tr>
</tbody>
</table>

Figure 5: Comparison of the approximations of

\[
E \left[ \sum_{i=1}^{n} x_i \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \right] = \sum_{i=1}^{n} x_i y_i
\]

following the Metropolis-Hastings algorithm with a bivariate normal (solid lines), and a bivariate Cauchy (dots) versus slice sampler method (dashes).

The diagnosis of Heidelberger and Welch (1983) indicates that slice sampler method does not satisfy the stationarity test, even if we repeat the rejection procedures until a proportion of 50% of iterations. That means that we need a longer chain to arrive at convergence. Contrary to the Metropolis-Hastings algorithm with a bivariate normal proposal, the table 2 shows that the stationarity test is immediately completed without any value to reject for each parameter of interest. With regard to the Metropolis-Hastings
algorithm with a bivariate Cauchy, the convergence is accomplished after having eliminated approximately a thousand iterations.

\[ \Delta \]

<table>
<thead>
<tr>
<th>Chain</th>
<th>Variable</th>
<th>Stationarity test</th>
<th>Total iters to keep</th>
<th>iters to discard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain 1</td>
<td>$\alpha$</td>
<td>passed</td>
<td>30000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>passed</td>
<td>30000</td>
<td>0</td>
</tr>
<tr>
<td>Chain 2</td>
<td>$\alpha$</td>
<td>passed</td>
<td>28900</td>
<td>1100</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>passed</td>
<td>28900</td>
<td>1100</td>
</tr>
<tr>
<td>Chain 3</td>
<td>$\alpha$</td>
<td>failed</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>failed</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 2: Table of Heidelberger and Welch diagnosis of the stationarity tests corresponding to this experiment. The stationarity tests are made for the 30,000 last iterations of the chains.

7 Discussion and Conclusions

With the increasing use of the MCMC algorithms in the Bayesian analysis, faster, efficient, and simpler methods for the generating random variates are required. In this paper, we have implemented and illustrated the use of popular algorithms from a Bayesian perspective. The comparison between the Gibbs sampler algorithm and an arbitrary Metropolis-Hastings algorithm would seem a priori to favor Gibbs sampling, since it draws its conditional distributions of the true law $\pi$, while the Metropolis-Hastings algorithm is, at best, based on a proposal law $g$ which is an approximation of $\pi$. Indeed, the slice sampler method is, by construction, more direct than the Metropolis-Hastings algorithm, because it is not subjected to a "bad" choice of the instrumental distribution and avoids useless simulations rejection. However, we notice in our study that the availability and the objectivity of the Gibbs sampling method are not necessarily arguments in its favour.

The simulation of a single component at each iteration of the Gibbs sampler algorithm [A4] strongly restricts the possible displacements of the chain $(\alpha^{(i)}, \beta^{(i)})$ and this makes the Gibbs sampler algorithms be slow to converge, since it is slow to explore the surface of $\pi$. This slowness proper to Gibbs sampler leads to a strong attraction towards the closest local modes, which induces enormous difficulties in exploring the entire range of
the support of $\pi$ in the case of multimodality.

In contrast, the defects of the Metropolis-Hastings algorithms are of another nature, because they more often come from a bad adjustment between the objective distribution $\pi$ and the proposal distribution $g$ than from a too strong approximation between the two distributions. The freedom given by the Metropolis-Hastings methods allows to cure these defects by increasing certain variation parameters. The essential disadvantage of the Metropolis-Hastings compared to the Gibbs sampling, is rather not always seize the details of the distribution $\pi$ because of a not very exact simulation scale.

In our study, we applied the slice sampler method by introducing auxiliary variables to have an easy access to the full conditional distributions which are essential to the Gibbs algorithm implementation. The Completion of $\pi(\alpha, \beta)$ by $\pi_1(\alpha, \beta, w)$ to have an inference on $\pi$ passes by the simulation of $\pi_1(\alpha, \beta, w)$ and the rejection of the auxiliary variables which are considered, in this case, as nuisance parameters. This considerably multiplies the calculating time and reduces the convergence speed of the Gibbs sampling method.

References


