Monte Carlo Methods and Bayesian Computation: MCMC

Peter Müller

Markov chain Monte Carlo (MCMC) methods use computer simulation of Markov chains in the parameter space. The Markov chains are defined in such a way that the posterior distribution in the given statistical inference problem is the asymptotic distribution. This allows to use ergodic averages to approximate the desired posterior expectations. Several standard approaches to define such Markov chains exist, including Gibbs sampling, Metropolis-Hastings and reversible jump. Using these algorithms it is possible to implement posterior simulation in essentially any problem which allow pointwise evaluation of the prior distribution and likelihood function.

1 Introduction

In Bayesian statistics the posterior distribution $p(\psi|\, y)$ contains all relevant information on the unknown parameters ψ given the observed data y. All statistical inference can be deduced from the posterior distribution by reporting appropriate summaries. This typically takes the form of evaluating integrals

$$J = \int f(\psi)\; p(\psi|\, y)\; d\psi \tag{1}$$

of some function $f(\psi)$ with respect to the posterior distribution. For example, point estimates for unknown parameters are given by the posterior means, i.e., $f(\psi) = \psi$; prediction for future data $\tilde{y}$ is based on the posterior predictive distribution $p(\tilde{y}|\, y) = \int p(\tilde{y}|\psi, y)\; p(\psi|\, y)\; d\psi$, i.e., $f(\psi) = p(\tilde{y}|\psi, y)$, etc. The problem is that these integrals are usually impossible to evaluate analytically. And when the parameter is multidimensional, even numerical methods may fail.

Over the last ten years a barrage of literature has appeared concerned with the evaluation of such integrals by methods collectively known as Markov chain Monte Carlo (MCMC) simulation. The underlying rationale of MCMC is to set up a Markov chain in ψ with ergodic distribution $p(\psi|\, y)$. Starting with some initial state $\psi^{(0)}$ we simulate M transitions under this Markov chain and record the simulated states $\psi^{(j)}$, $j = 1, \ldots, M$. The ergodic sample average

$$\hat{J} = \frac{1}{M} \sum_{j=1}^{M} f(\psi^{(j)}) \tag{2}$$

converges to the desired integral J (subject to some technical conditions), i.e., $\hat{J}$ provides an approximiate evaluation of The art of MCMC is to set up a suitable Markov chain with the desired posterior as stationary distribution and to judge when to stop simulation, i.e, to diagnose when the chain has practically converged.

In many standard problems it turns out to be surprisingly easy to define a Markov chain with the desired stationary distribution. We will review the most important approaches in this entry. The general principle of Monte Carlo simulation, including independent Monte Carlo simulation, is discussed in *Monte Carlo Methods and Bayesian Computation: Overview*.

2 The Gibbs Sampler

Example 1 (Gelfand et al. 1990): Consider a variance components model $y_{ij} = \theta_i + e_{ij}$, $i = 1, \ldots, K$ and $j = 1, \ldots, J$, for data y_{ij} from K groups with J observations in each group. Assume independent normal errors $e_{ij} \sim N(0, \sigma_e^2)$ and a normal random effects model $\theta_i \sim N(\mu, \sigma_\theta^2)$. We assume that $\theta = (\theta_1, \ldots, \theta_k)$, (μ, σ_θ^2), and σ_e^2 are a priori independent with $p(\sigma_\theta^2) = IG(a_1, b_1)$, $p(\mu|\sigma_\theta^2) = N(\mu_0, \sigma_\theta^2)$, and $p(\sigma_e^2) = IG(a_2, b_2)$. Here we use $N(m, s^2)$ to indicate a normal distribution with moments m, s, and $IG(a, b)$ to indicate an inverse gamma distribution with parameters a and b. Let $y = (y_{ij},\ i = 1, \ldots, K,\ j = 1, \ldots, J)$ denote the data vector. It can be shown that the conditional posterior distributions $p(\sigma_\theta^2|y, \mu, \theta, \sigma_e^2)$ and $p(\sigma_e^2|y, \mu, \theta, \sigma_\theta^2)$ are inverse gamma distributions, and $p(\mu|y, \theta, \sigma_\theta^2, \sigma_e^2)$, and $p(\theta|y, \mu, \theta, \sigma_\theta^2, \sigma_e^2)$ are normal distributions.

To estimate posterior moments of the type (1) we define a Markov chain in $\psi = (\mu, \theta, \sigma_e^2, \sigma_\theta^2)$. Denote with $\psi^{(t)} = (\mu^{(t)}, \theta^{(t)}, \sigma_e^{2(t)}, \sigma_\theta^{2(t)})$. the state vector of the Markov chain after t transitions. Given the nature of a Markov chain, all we need to define is the transition probability, i.e., given a current value for $\psi^{(t)}$, we need to generate a new value $\psi^{(t+1)}$. We do so by sampling from the complete conditional posterior distributions for $\mu, \sigma_e^2, \sigma_\theta^2$ and θ

1. $\mu^{(t+1)} \sim p(\mu|y, \theta^{(t)}, \sigma_e^{2(t)}, \sigma_\theta^{2(t)})$,
2. $\theta^{(t+1)} \sim p(\theta|y, \mu^{(t+1)}, \sigma_e^{2(t)}, \sigma_\theta^{2(t)})$,
3. $\sigma_e^{2(t+1)} \sim p(\sigma_e^2|y, \mu^{(t+1)}, \theta^{(t+1)}, \sigma_\theta^{2(t)})$,
4. $\sigma_\theta^{2(t+1)} \sim p(\sigma_\theta^2|y, \mu^{(t+1)}, \theta^{(t+1)}, \sigma_e^{2(t+1)})$.

Steps 1 through 4 define a Markov chain $\psi^{(t)}$ which converges to $p(\mu, \theta, \sigma_e^2, \sigma_\theta^2|\, y)$, as desired. Ergodic averages of the type $\hat{J} = 1/M \sum f(\psi^{(t)})$ provide nu-

merical evaluations of any desired posterior integral J.

The described Markov chain Monte Carlo simulation is a special case of a Gibbs sampler. In general, let $\psi = (\psi_1, \ldots, \psi_p)$ denote the parameter vector. The Gibbs sampler proceeds by iteratively, for $j = 1, \ldots, p$, generating from the conditional posterior distributions

$$\psi_j^{(t+1)} \sim p(\psi_j | \psi_1^{(t+1)}, \ldots, \psi_{j-1}^{(t+1)}, \psi_{j+1}^{(t)}, \ldots, \psi_p^{(t)}, y). \quad (3)$$

If practicable it is advisable to generate from higher dimensional conditionals. Compare the discussion in *Monte Carlo Methods and Bayesian Computation: Overview*, Section 2.2.

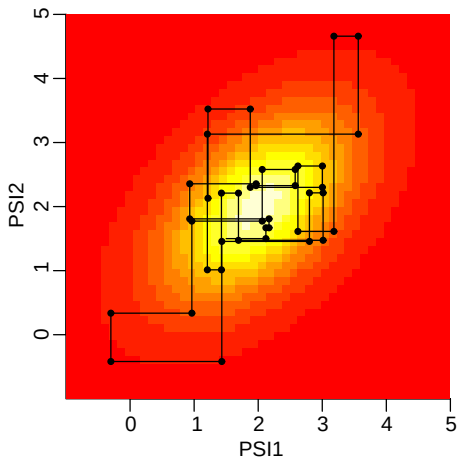

Figure 1: Gibbs sampler. The grey shades show a bivariate posterior distribution $p(\psi_1, \psi_2 | y)$. The connected points show the parameter values $\psi^{(t)}$ generated in $M = 40$ transitions of the MCMC simulation. The transition probabilities in the Gibbs sampler are the full conditional posterior distributions (3), leading to the piecewise horizontal and vertical trajectories seen in the figure. Each horizontal line segment corresponds to generating a new value $\psi_1^{(t+1)} \sim p(\psi_1 | y, \psi_2^{(t)})$. Each vertical line segment corresponds to generating $\psi_2^{(t+1)} \sim p(\psi_2 | y, \psi_1^{(t+1)})$.

Figure 1 illustrates the Gibbs sampling algorithm. The figure shows simulated parameter values for a hypothetical bivariate posterior distribution $p(\psi_1, \psi_2 | y)$.

The seminal paper by Gelfand and Smith (1990) and the companion paper by Gelfand et al. (1990) popularized the Gibbs sampler for posterior simulation in a wide class of important problems. Many earlier papers used essentially the same method in specific problems. For example, a special case of the Gibbs sampler occurs in problems with missing data. In many problems, the actually observed data y can be augmented by missing data z in such a way that simulation from $p(\psi | y, z)$ and $p(z | \psi, y)$ can be implemented in computationally efficient ways, even when simulation from the original posterior distribution $p(\psi | y)$ is difficult. Tanner and Wong (1987) propose what is essentially a Gibbs sampler for the augmented posterior distribution $p(\psi, z | y)$. Geman and Geman (1984) proposed the Gibbs sampler for posterior simulation in a spatial model with a Markov random field prior.

3 The Metropolis-Hastings Algorithm

The Gibbs sampler owes some of its success and popularity to the fact that in many statistical models the complete conditional posterior distributions $p(\psi_j | \psi_i, i \neq j, y)$ take the form of some well-known distributions, allowing efficient random variate generation. But there remain many important applications where this is not the case, requiring alternative MCMC schemes. Possibly the most generic such scheme is the Metropolis scheme (Metropolis et al., 1953). The general form of the algorithm is defined in *Monte Carlo Methods and Bayesian Computation: Overview*, Section 2.1. Consider generating from a posterior distribution $p(\psi | y)$. Denote with ψ the current state of the Markov chain. One transition is defined by the following steps:

1. Generate a proposal $\tilde{\psi}$ from some proposal generating distribution $q(\tilde{\psi} | \psi)$. The choice of the proposal distribution $q(\cdot)$ is discussed below.
2. Compute
$$a(\psi, \tilde{\psi}) = \min \left\{ 1, \frac{p(\tilde{\psi} | y)}{p(\psi | y)} \cdot \frac{q(\psi | \tilde{\psi})}{q(\tilde{\psi} | \psi)} \right\} \quad (4)$$
3. With probability a replace ψ with the proposal $\tilde{\psi}$. Otherwise, leave ψ unchanged.

Figure 2 illustrates the algorithm. The figure shows the proposals $\tilde{\psi}$ and the (accepted) states $\psi^{(t)}$ for the first 40 iterations of a Metropolis chain simulation for a hyptothetical bivariate posterior distribution. The choice of the proposal distribution $q(\tilde{\psi} | \psi)$ is essentially arbitrary, subject only to some technical constraints. Using a symmetric proposal distribution with $q(\tilde{\psi} | \psi) = q(\psi | \tilde{\psi})$, for example a normal centered at ψ, has the practical advantage that the ratio

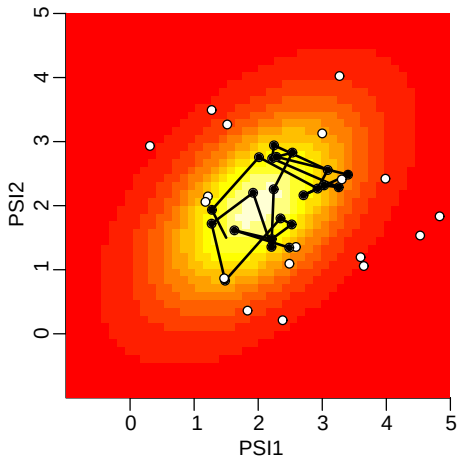

Figure 2: Metropolis sampler. The grey shades show a posterior distribution $p(\psi|\, y)$. The connected solid points show the parameter values $\psi^{(t)}$ generated in $M = 40$ transitions of a Metropolis chain with bivariate normal proposals $\tilde{\psi} \sim N(\psi, 0.75\, I)$, where I denotes the 2×2 unit matrix. The empty circles show generated proposals $\tilde{\psi}$ which were rejected using the acceptance probabilities (4). Compare with Figure 1 which shows a Gibbs sampler with an equal number of transitions.

of proposal distributions $q(\psi|\tilde{\psi})/q(\tilde{\psi}|\psi)$ cancels out of the expression for $a(\cdot)$. Often *Metropolis chain* is used to refer to this special case only. Another practically interesting variation is the use of an independent probing distribution $q(\tilde{\psi})$, i.e., the proposal is independent of the current state. Tierney (1994) refers to such algorithms as *independence chains*. Hastings (1970) proposes a larger class of similar algorithms based on a more general expression for the acceptance probability. Chib and Greenberg (1995) give a tutorial introduction to the Metrlopolis-Hastings algorithm. See section 2.1. in *Monte Carlo Methods and Bayesian Computation: Overview* for a more detailed discussion. Section 2.2. in *Monte Carlo Methods and Bayesian Computation: Overview* explains generalizations of the Metropolis-Hastings algorithm to multiple-block updating.

4 Convergence

The use of integral estimates (2) requires the verification of two conditions related to convergence.

First, the chain has to theoretically, i.e., for $M \to \infty$, converge to the desired posterior distribution. Second, even if convergence for $M \to \infty$ is established, we need a convergence diagnostic to decide when we can terminate simulations in a practical implementation.

Tierney (1994, Theorem 1) shows convergence (in total variation norm) under three conditions: irreducibility, aperiodicity and invariance.

The Markov chains which are used in MCMC schemes generally use a continuous state space, i.e., $\psi^{(t)}$ is a real valued vector. For such continuous state spaces the notion of irreducibility is formally defined as π-irreducibility, with respect to some measure π on the state space. For the purpose of the present discussion we only consider $\pi(\psi) = p(\psi|\, y)$, i.e., π denotes the desired stationary distribution. A Markov chain is π-irreducible if for any state ψ and any set B of states with $\pi(B) > 0$ there exists an integer $n \geq 1$ such that in n iterations the chain can with positive probability make a transition from ψ to some state in B.

Invariance refers to the property that if we start with a state vector generated from the desired posterior distribution, i.e., $\psi^{(t)} \sim \pi$, then a further transition in the Markov chain leaves the marginal sampling distribution of ψ unchanged, i.e., $\psi^{(t+1)} \sim \pi$.

The Gibbs sampler and the Metropolis-Hastings scheme define Markov chains which by construction are invariant with respect to the desired posterior distribution. Irreducibility and aperiodicity need to be verified, but are usually not a problem. However, sometimes MCMC implementations suffer from practical violations of irreducibility. There might be some subsets of the parameter space which are such that once the Markov chain simulation enters this set it is very unlikely to leave this subset again within any reasonable number of iterations. Such situations occur, for example, in independence chains if the proposal distribution $q(\tilde{\psi})$ has thinner tails than the desired posterior $\pi(\psi)$. The acceptance probabilties (4) include the ratios $\pi(\psi)/q(\psi)$. Assume the chain has generated a parameter value ψ far out in the tail, with very large ratio $\pi(\psi)/q(\psi)$. The chain will then reject any proposed move until a new proposal $\tilde{\psi}$ equally far out in the tail is generated.

Practically more important than establishing theoretical convergence is to recognize practial convergence, i.e., to judge when sufficiently many transitions M have been simulated to obtain ergodic averages $\hat{J}$ close to the desired posterior expectations J. The simplest procedure is to plot the trajectories $\psi^{(t)}$ against iteration number t and judge convergence if an informal visual inspection of the plot does not reveal obvious trends. Figure 3 shows a typical trajectory. Several more formal convergence diagnostics have been proposed in the recent litera-

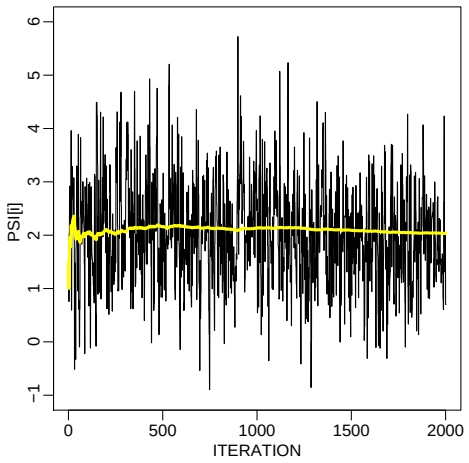

Figure 3: Convergence. The figure plots the first 2000 steps for the Gibbs sampler shown in Figure 1. The thin black curve plots $\psi_1^{(t)}$ against iteration t. The thick grey curve plots $\hat{J}_t = 1/t \sum_{j=1}^{t} \psi_1^{(t)}$. After about 500 iterations the estimated posterior mean has practically converged. See the text for a discussion of more formal convergence diagnostics.

ture. Gelman and Rubin (1992) propose to consider several independent parallel runs of the MCMC simulation. Convergence is diagnosed if the differences of $\hat{J}$ across the parallel runs are within a reasonable range. Gelman and Rubin (1992) formalize this with an ANOVA type statistic. Geweke (1992) proposes to compare an ergodic average based on early simulations (say the first 10% of the iterations) with an ergodic average based on later iterations (say the last 50%). Under convergence the two ergodic averages should be approximately equal. Using an approximate sample standard deviation based on spectral density estimates allows a formal test. Section 2.1. in *Monte Carlo Methods and Bayesian Computation: Overview* discusses an approach based on tracking autocorrelation times.

These and other convergence diagnostics are discussed in Best et al. (1995) and implemented in the public domain software BOA described there.

5 Limitations and Further Reading

The Gibbs sampler and the Metropolis-Hastings chain implicitely require a fixed dimension parameter space, i.e., the dimension of ψ must not change across different values. This excludes, for example, a regression model with an unknown number of covariates. In other words, the Gibbs sampler or the Metropolis-Hastings algorithm can not be used for model selection. Extensions of the basic MCMC schemes which allow model comparison and simulation across models with different dimension parameter spaces are discussed in Section 4 in *Monte Carlo Methods and Bayesian Computation: Overview*.

Several recent monographs provide more complete reviews of MCMC methods. Tanner (1996) provides an introduction including related schemes such as importance sampling. Assuming basic familiarlity with the algorithms, Gilks et al. (1996) discuss Markov chain Monte Carlo simulation in the context of important statistical models. Gamerman (1997) and Robert and Casella (1999) review alternative algorithms and related theory. Relevant references for specific models are listed in Section 5 of *Monte Carlo Methods and Bayesian Computation: Overview*.

References

Best, N. G., Cowles, M. K., and Vines, S. K., (1995), *Convergence diagnosis and output analysis software for Gibbs sampling output, Version 0.3*, Cambridge, UK: MRC Biostatistics Unit.

Carlin, B. and Chib, S., (1995), "Bayesian model choice via Markov chain Monte Carlo," *JRSS B*, 57, 473–484.

Chib, S. and Greenberg, E., (1995), "Understanding the Metropolis-Hastings algorithm," *The American Statistician*, 49, 327–335.

Gamerman, D., (1997), *Markov Chain Monte Carlo*, London: Chapman and Hall.

Gelfand, A. and Smith, A., (1990), "Sampling based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelfand, A.E., Hills, S.E., Racine–Poon, A., and Smith, A.F.M., (1990), "Illustration of Bayesian inference in normal data models using Gibbs sampling," *Journal of the American Statistical Association*, 85, 972–985.

Gelman, A. and Rubin, D., (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457 – 473.

Geman, S. and Geman, A., (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6), 721–740.

Geweke, J., (1992), "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 169–194, Oxford: Oxford University Press.

Gilks, W., Richardson, S., and Spiegelhalter, D., (1996), *Markov chain Monte Carlo in practice*, Chapman and Hall.

Green, P., (1995), "Reversible jump Markov chain Monte Carlo compuationa and Bysian model determination," *Biometrika*, 82, 711–732.

Hastings, W., (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953), "Equations of state calculations by fast computing machines," *Journal of Chemical Physics*, 21, 1087–1091.

Robert, C. and Casella, G., (1999), *Monte Carlo Statistical Methods*, New York, NY, USA: Springer-Verlag.

Tanner, M., (1996), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd. ed., New York: Springer-Verlag.

Tanner, M. and Wong, W., (1987), "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 528–550.

Tierney, L., (1994), "Markov chains for exploring posterior distributions," *Annals of Statistics*, 22, 1701–1728.