MARKOV CHAIN MONTE CARLO METHODS IN BAYESIAN INFERENCE

P. Venkatesan

Scientist-E & Head, Department of Statistics,
Tuberculosis Research Center (ICMR), Chennai-600 031
Email:venkatesanp@icmr.org.in

ABSTRACT

The Markov Chain Monte-Carlo (MCMC) born in early 1950s has recently aroused great interest among statisticians, particularly researchers working in image analysis, discrete optimization, neural networks, genetic sequencing and other related fields. Recent theoretical achievements in resampling procedures provide a new perspective for handling errors in Bayesian inference, which treats all unknowns as random variables. The unknowns include uncertainties in the model such as fixed effects, random effects, unobserved indicator variables and missing data. Only in few cases, the posterior distribution is in standard analytic form. In most other models like generalized linear models, mixture models, epidemiological models and survival analysis, the exact analytic Bayesian inference is impossible. This paper surveys some of the recent advances in this area that allows exact Bayesian computation using simulations and discusses some applications to biomedical data.

Keywords : *Bayesian inference, Markov Chain Monte Carlo, Gibbs, Metropolis, mixture model, hierarchical model, ECM algorithm, panic attack.*

Introduction

Markov Chain Monte Carlo (MCMC) is a powerful technique for performing integration by simulation. In recent years MCMC has revolutionized the application of Bayesian statistics. Many high dimensional complex models, which were formally intractable, can now be handled routinely. MCMC has also been used in specialized non- Bayesian problems. A good introduction on MCMC methods in biostatistical applications can be found in Gilks et al (1996) and Gelman and Rubin (1996). The techniques have been applied in most areas of statistics and Biostatistics namely vaccine efficacy, genomics, proteomics, clinical monitoring, pharmacokinetics, disease mapping, image analysis, genetics and epidemiological research. Gilks et al (1996), Berry and Stangle (1996) describe applications in decision analysis, clinical trial design, and cross over trials, meta analysis, change point analysis, hemodynamics and prenatal mortality. The applications of MCMC in modeling situations involve hierarchical models, missing data, censored data, and spatially correlated data. MCMC methods

have also been used extensively in statistical physics over the last 40 years, in spatial statistics for the l 20 years and in Bayesian image analysis over the last 10 years (Gilks et al 1996). In the last 5 y MCMC has been introduced into significance testing, general Bayesian inference and maxin likelihood estimation.

The use of MCMC was first introduced in statistical mechanics by Metropolis et al (1953 study the equation of the state of a two-dimensional rigid sphere system. The choice made by Metrop et al (1953) was one of many other possibilities. They introduced MCMC as a general method suitable fast computing machines of calculating the properties of any substance considered as composed interacting individual molecules. Now this method has become a miraculous tool of Bayesian anal (Geyer, 1992) and the flag of what has been called as the model liberation movement (Sm 1992). Bayesian calculations not analytically tractable can be performed once a likelihood and prior given (Besag et al, 1995). For non-Bayesian applications MCMC is considered as a very powe numerical device in likelihood analysis or decision theory (Geyer, 1992).

Early 1990 have witnessed a burst of activities in applying MCMC in Bayesian methods simulate Bayesian distributions. The simulation algorithm in its basic form is quite simple and becoming standards in many Bayesian application .A good review is given by Gilks et al (1996). MCI methods have been successfully used to over come problems caused by missing data when using sn networks for conventional statistics. All MCMC methods are ways to produce a stochastic process, wh has a desired distribution as its stationary distribution .The theory of stochastic processes tells as that empirical average of a function of the stochastic process will converge to the expectation of that funct under the desired distribution. MCMC is the idea of using simulations $X_1, X_2, \ldots X_n$ of the Markov Ch to approximate expectations $\mu=E_\pi\{g(X_i)\}$ by sample averages $\mu_n=1/n\Sigma g(X_i)$ where π is equilibri distribution also called invariant distribution, stationary distribution or ergodic limit of the Markov Cha

Gibbs Sampler

The two most commonly used algorithms in MCMC applications are (i) Metropolis Algorithms and Gibbs Sampler. Geman and Geman (1984) present the Gibbs sampler in context of spatial proces involving large number of variables for image reconstruction. They consider situations under which conditional distributions given neighbourhood subsets of the variables uniquely determine the jc distributions. Besag and York (1989) has shown that if the joint distribution P $(\theta_1, \theta_2, \ldots\ldots \theta_d)$ positive over its entire domain, then the joint distribution is uniquely determined by the d conditio distributions P $(\theta_1 / \theta_2, \ldots, \theta_d)$ ..... P $(\theta_d / \theta_1, \ldots, \theta_{d-1})$. Li (1988) applied the Gibbs sampler in

ontext of multiple imputations. Li suggests that the complete data be partitioned into d+1 parts, $_0, X_1, \ldots X_d$, where the observed data X_0 and $X_1, \ldots, X_d$ is a partition of missing data. Li assumes that X_i an be sampled from $P(X_i / X_j, j \neq i)$ and the algorithm is as follows:

tep1: Sample $X_1^{(0)}, \ldots, X_d^{(0)}$ from some distributions.

ep 2: Sample $X_1^{(j)}$ from $P(X_1/ X_0, X_2^{(j-1)}, \ldots, X_d^{(j-1)})$

Sample $X_2^{(j)}$ from $P(X_2/ X_0, X_1^{(j)}, X_3^{(j-1)}, \ldots, X_d^{(j-1)}), \ldots,$

Sample $X_d^{(j)}$ from $P(X_d/ X_0, X_2^{(j)}, \ldots\ldots X_{d-1}^{(j)})$

ep 2 is repeated until the algorithm converges. Li suggests that multiple paths be considered to check for nvergences. He also illustrated the method in the context of categorical data, latent variables, and nsored life data and provide conditions for the distribution of $(X_1^{(i)} \ldots\ldots X_d^{(i)})$ to converge to $P(0X_1, X_2 \ldots X_d)$. Like Metropolis et.al (1953) and Geman and Geman (1984), Li represents the process as Markov ain with the joint posterior distribution as the stationary distribution.

Tanner and Wong (1987) present the data augmentation algorithm, which is a two-component version the Gibbs sampler. One of the basic contributions of Tanner and Wong (1987) was to develop the mework in which the Bayesian can be performed in the context of iterative Monte Carlo algorithms. oreover, in their rejoinder they sketch a Gibbs sampler approach for handling hierarchical models with t ors. Gelfand and Smith (1990) present a review of data augmentation, the Gibbs sampler and the SIR gorithm due to Rubin (1987). The papers by Gelfand et.al (1990, 1992) and Carlin et al (1992) apply the bbs sampler to a variety of important statistical problems.

etropolis-Hasting Algorithm

The more general updating scheme as a form of the generalized rejection samplings is the etropolis algorithm Hasting (1970) updated in the Metropolis algorithms using arbitrary transition obability function. We first consider the idea in the discrete case. Let $Q = \{ q_{ij} \}$ be a ecified symmetric transition matrix. At a given step, randomly draw state s_j from the i^{th} row of Q. With own probability α_{ij}, move from s_i to s_j, otherwise, remain at step s_i. This construction defines a Markov ain with transition matrix $p_{ij} = \alpha_{ij} q_{ij}$ $(i \neq j)$ and $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$,

Where $\alpha_{ij} = \begin{cases} 1 & if \ \frac{\pi_j}{\pi_i} \geq 1 \\ \frac{\pi_j}{\pi_i} & if \ \frac{\pi_j}{\pi_i} < 1 \end{cases}$

This chain is reversible, since

$$\pi_i p_{ij} = \pi_i \min\left\{1, \frac{\pi_j}{\pi_i}\right\} q_{ij}$$

$$= \text{Min}\{\pi_i, \pi_j\} q_{ij}$$

$$= \text{Min}\{\pi_i, \pi_j\} q_{ji}$$

$$= \pi_j p_{ji}$$

The equilibrium distribution will be unique if Q is irreducible. A sufficient condition for convergence is not constant) is being able to move from any state to any other under Q takes $\alpha_{ij}=\pi_j /(\pi_i+\pi_j)$ (Barker 1965). The resulting chain is reversible, since, $\pi_i p_{ij}=\pi_i \pi_j /(\pi_i+\pi_j)q_{ij}=\pi_j p_{ji}$.

Now we consider the idea in the continuous case: Here π is a density with respect to a measure μ and f (x, y) is a symmetric transition probability function (i.e., f (x, y) = f (y, x)), then the Metropolis algorithm is given by the following two steps.

(a) If the chain is currently at $X_n = x$, then generate a candidate value y* for next location X_{n+1} f(x,y).

(b) With probability $\alpha(x,y^*)=\min \quad 1,\pi(y^*)/ \pi(x)$, accept the candidate value and move the cha $X_{n+1} = y^*$. Otherwise reject and let $X_{n+1} = x$. Thus the Metropolis algorithm yields a series of dependent realizations forming a Markov chain with π as its equilibrium distributi key observation is that the Metropolis algorithm only requires that π be defined up to the normal constant, since the constant drops out in the ratio $\pi(y^*)/ \pi(x)$.

Tierney (1991) presents a number of suggestions for f(x, y). If f(x, y) = f(y-x), then the ch driven by a random – walk process. Possible candidates for f are the multivariate normal, multivariat split t (with a small degrees of freedom). In situations where the multivariate t is used to ger candidate values, one would center the normal or the t at the current state of the chain x, with the vari

covariance matrix possibly equal to some multiple of the inverse information at the posterior mode. Muller (1991) discusses the choice of scale issue in detail. Besides presenting the range of hybrid strategies by cycling/ mixing different chains, Tierney also (1991) presents formal conditions for convergence, rates of convergence, and limiting behavior of averages. Gelfand (1992) presents an analogue to the Gibbs stopper for the Metropolis algorithm.

A earliest generalization of the Metropolis algorithm was due to Hastings (1970) which defines

$$\alpha(x, y) = \begin{cases} \min\{1, \pi(y)q(y,x) / \pi(x)q(x,y)\} & \text{if } \pi(x)q(x,y) > 0 \\ 1 & \text{if } \pi(x)\, q\,(x, y) = 0 \end{cases}$$

where q (x, y) is an arbitrary transition probability function. If q is symmetric i.e., q (x, y) = q (y, x), as would be the case in using a multivariate normal or multivariate t to drive the algorithm, then the Hastings algorithm reduces to the Metropolis algorithm. Hastings (1970) considers the case where q (x, y) = q (y), which is closely related to importance sampling. Tierney (1991) calls this as independence chains. This form of acceptance probability is not unique since there may be many acceptance functions, which provide a chain with the desired properties. However Peskun (1973) showed that this form is optimal in that suitable candidates are rejected less often and so efficiency is maximized.

Bayesian Inference using Simulation

Given a set of posterior simulation draws $\theta^1,\ldots,\theta^N$ of a vector parameter θ, one can estimate the posterior distribution of any quantity of interest. For example with N=1000 simulation draws one can estimate a 95% posterior interval for any function $\phi(\theta, y)$ of parameters and data by the 25^{th} largest and 975^{th} largest simulated values of $\phi(\theta^L, y)$, $L = 1,\ldots,1000$. In some problems such as the normal linear regression model random draws can be obtained from the posterior distribution directly in one step (Gelman et al 1995). In other complicated cases such as normal linear regression model with unknown variance, the parameter vector can be partitioned into two sub vectors $\theta = (\theta_1,\theta_2)$ such that the posterior distribution of θ_1, $p(\theta_1/y)$ and the conditional posterior distribution of θ_2 given θ_1, $p(\theta_2/\theta_1,y)$ are both standard distributions from which simulations can easily be drawn.

Many problems such as generalized linear models and hierarchical models direct simulation is not possible even with two or more steps. Until recently these problems have been attacked by approximating the desired posterior distributions by normal or transformed normal distributions from which direct simulation can be drawn in recent years iterative simulation methods such as MCMC have been developed

to draw from general distributions without any direct need for normal approximation (Gelman Rubin1996). The advantage of these iterative methods is that they can be setup with virtually any m that can be setup in statistics. The main limitation is that they currently require extensive programr and debugging. In Bayesian posterior distribution, the goal of iterative simulation is the inference a the target distribution and not merely some moments of the target distribution. So it is desirable to ch starting points that widely dispersed in the target distribution over dispersed starting points ar important design feature of MCMC for two major reasons.

1. Starting far apart can make lack of convergence apparent.
2. Starting over dispersed can ensure that all major reasons of the target distributions represented in the simulation.

The chain of the class of models where MCMC is easy to use, assessing the convergence, guidelines for starting values. Methods assessing the behavior of the chain and useful softwares extensively discussed by many authors.(e.g. Kass et al 1998,Spiegalhalter et al 1995,Cowles and C 1996, Gelman and Rubin 1992, Boscardin 1996, Gerlman 1996, Besag and Green 1993, Geyer Thompson 1995, Gelfand et al 1995,1996, Besag et al 1995 etc.)

An Application to Medical Data.

We illustrate the application of MCMC with a mixed model to data obtained from patients with p attack (PD). This application is complicated that Markov Chain simulation methods are the most effe tool for exploring the posterior distribution

Panic Attack Data

In the experiment under study of 20 subjects – 10 controls (5 males and 5 females) an patients (6 males and 4 females) had their ECG measured continuously for 60 beats. We briefly re the basic statistical approach here. The R-R intervals at 60 beats on standing were compared for cor and patients. The patients with PD showed a highly significant decrease in R-R variance upon star from supine posture when compared to that of controls. During an induced panic attack with 5% inhalating in PD patients, the R-R variance further decreased when compared to the levels befor attack.

Finite mixture likelihood model

To address the problem the following basic model was fit. R-R variance of the controls is described by random effect model in which the responses Y_{ij} (i = 1, 2,...,20) of person j (j = 1, 2,...,10) are normally distributed with distinct mean α_j and common variance σ_y^2. To reflect the response of PD. patients (j = 11 to 20) are modeled as a two compartment mixture with probability (1 - λ) for controls and R – R variance is normally distributed with mean α_j and variance σ_y^2 and with probability λ for R-R variance of the PD patients with mean $\alpha_j - \tau$ and the same variance σ_y^2 who exhibits symptoms such as palpitation, swatting, dizziness, shortness of breath and other atomic symptoms (Bharathi et.al 1994). These are attributed to the dysfunction of both sympathetic and parasympathetic systems in these patients.

Hierarchical Population Model

The comparison of the components of $\alpha = (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{20})$ for PD patients verses controls addresses the magnitude of decrease in cholinergic activity. We include a hierarchical parameter β measuring the activity. Specifically variation among the individuals is modeled by having the means α_j follow a normal distribution with mean μ for controls and $\mu + \beta$ for PD patients with each distribution having variance σ_α^2. i.e., the mean of α_j in the population distribution is $\mu + \beta S_j$ where S_j is an indicator variable with 1 if the person j is PD and 0 otherwise. We followed the Bayesian model with an improper uniform prior distribution on the hyper parameters $\phi = (\sigma_y^2, \sigma_\alpha^2, \lambda, \mu, \beta, \tau)$as given by Gelman and Rubin (1996).

Posterior Modes using Expectation Conditional Maximization (ECM) Algorithm

We draw 100 points at random from the distribution and use each as a starting point for the ECM algorithm to search for modes as given by Gelman and Rubin (1996). We also approximated the posterior distribution by a multivariate t distribution centered at the major mode of ECM with covariance matrix as the inverse of negative of the second derivative matrix of the log posterior density. We have drawn 1000 independent samples and importance resample a subset of 10, which was used as a starting point for independent Gibbs samplers.

The Table I displays the posterior inferences and potential scale reduction factor for selected parameters after 50 iterations and 200 iterations. After 200 iterations, the potential scale reduction factor was approximately 1 for all parameters in the model.

Table I: Posterior quantiles and estimated potential scale reduction factors for parameters.

Parameter	After 50 iterations				After 200 iterations			
	2.5%	50%	97.5%	$\sqrt{R}$	2.5%	50%	97.5%	$\sqrt{R}$
λ	0.11	0.23	0.45	2.2	0.15	0.19	0.24	1.02
τ	0.62	0.91	1.37	1.8	0.77	0.89	1.22	1.00
β	0.23	0.45	0.67	1.4	0.32	0.45	0.64	1.01

Discussion

The existing MCMC methods provide a powerful statistical tool and have revolutionized Practi Bayesian statistics over the past few years. The Ability to fit complicated models with little programm effort is in fact a key advantage of MCMC methods. The MCMC simulation should be undertaken a the problem has been approximated and explore using simple methods. There are a variety of methods constructing efficient MCMC algorithms. However the implimentation of many of these methods requ some expertise. The main problem is ascertaining the proximity of the distribution of any given Mark chain output to the target distribution. Even though the recent works focus on construction of samp without this problem using exact samples. Considerable work is still needed on the implimentation issu

References:

Barker (1965) : Monte Carlo calculations of radial distribution functions for proton- electron plasma. Australian Journal of Physics 18, 119-133.

Berry, D.A and Stangle, D.K (1996) : Biostatistics, Marcel Dekker, New York.

Besag J and York J.C (1989): Bayesian restoration of images. In analysis of statistical information (ed T.Matsunawa) 491-507Tokyo, Institute of statistical mathematics

Besag, J., and Green, P.J (1993): Spatial Statistics and Bayesian Computation (with discussion). Journal of the Royal Statistical Society, Ser B, 55,72-37.

Besag.J., Green P.J., Migdon.D & Mengersen,K. (1995): Bayesian computation and stochastic systems. Statistical Science 10,3-41.

Bharathi. P, Prema Sembulingam, Sembulingam K, Srinivasan T.N, Venkatesan P, and Namasivayam .A (1994): R-R variance during CO_2 –induced Panic States. Biomedicine, 14, 113-116.

Boscardin, W.J. (1996): Bayesian Analysis for Some Hierarchical linear Models Unpublished Ph.D. Thesis, University of California-Berkeley, Dept. of Statistics.

Carlin B.P, Gelfand A.E and Smith A.F.M (1992): Hierarchical Bayesian analysis of change point problems. Journal of the Royal Statistical Society 41,389-405.

Cowles, M.K., and Carlin, B.P (1996): Markov Chain Monte Carlo Convergence Diagnostics: A comparative Review. Journal of the American Statistical Association, 91,883-904.

Gelfand A.E and Smith A.F.M (1990): Sampling- based approaches to calculating marginal densities. Journal of the American Statistical Association 85,398-409.

Gelfand A.E, Hills 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.

Gelfand A.E (1992): Discussion to the paper of Gelman and Rubin and of Geyer. Statistical Science 4, 486-487.

Gelfand A.E., Smith A.F.M and Lee T.M (1992): Bayesian Analysis of Constrained parameter and truncated data problems. Journal of the American Statistical Association 87,523-532.

Gelfand, A.E., Sahu, S.K., and Carlin, B.P. (1995): Efficient parameterization for normal linear mixed models. Biometrika, 82, 479-488.

Gelfand, A.E., Sahu, S.K., and Carlin, B.P.(1996): Efficient parameterizations for generalized linear mixed models (with discussion). In Bayesian Statistics 5(eds. J.M. Bernardo, J.O.Berger, A.P. Dawid, and A.F.M. Smith) Oxford: Oxford University Press, pp. 165 –180.

Gelman, A., and Rubin, D.B (1992): Inference from Iterative Simulation using Multiple Sequences" (with discussion). Statistical Science, 7, 457-511.

Gelman. A. (1996) : Inference and Monitoring Convergence. In Markov Chain Monte Carlo in Practice, (eds W.R.Gilks, S. Richardson, and D.J. Spiegelhalter)London: Chapman and Hall, pp. 131-143.

Gelman.A and Rubin D.B (1996) : Markov Chain Monte Carlo Methods in Biostatistics, Statistical methods in Medical Research 5,339-355.

Geman.S and Geman.D (1984): Stochastic relaxation Gibbs distributions and the Bayesian restoration of images. IEEE transactions on Pattern Analysis and Machine Intelligence 6, 721-741

Geyer C.J (1992): Practical Markov Chain Monte Carlo. Statistical Science 7, 473-511.

Geyer, C.J., and Thompson, E.A. (1995): Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. Journal of the American Statistical Association, 90, 909-920.

Gilks, W.R. Richardson, S. and Spiegel halter D.J Eds. (1996): Markov Chain Monte Carlo in Practice. Chapman & Hall, London.

Hastings W.K (1970): Monte Carlo Sampling Methods using Markov Chains and their applications, Biometrika 57,97-109.

Kass, R.E, Carlin B. P, Gelman. A and Neal. R.M (1998): Markov chain Monte Carlo in practice; A round table discussion. The American Statistician 52, 93-100.

Li .K.M (1988): Imputation using Markov Chain. Journal Of Statistical computation and Simulation 30,57-79.

Metropolis .N, Rosenbluth A.W., Rosenbluth, M.N, Teller A.H and Teller.E (1953): Equation of State Calculations by fast computing Machine. Journal of chemical physics 21.1087-091.

Muller – Krumbhaar.H and Binder.K (1973): Dynamic properties of the Monte Carlo method in Statistical Mechanics. Journal of Statistical Physics 8,1-24.

Muller.P (1991) : A generic approach to posterior integration and Gibbs sampling. Technical Report 91-09, Dept. of Statistics, Purdue University.,

Peskun.P.H. (1973): Optimum Monte Carlo Sampling using Markov chains; Biometrika 60, 607-612.

Ripley, B.D. (1987): Stochastic Simulation. John Wiley, New York.

Rubin D.B (1987): comment on " The calculation of posterior distributions by data augmentation" by M.A Tanner and W.H.Wong. Journal of American Statistical Association 82, 543-546.

Smith A.F.M and Gelfand A.E (1992): Bayesian Statistics without tears. American Statistician 46, 84-88. Spiegelhater, D. J., Thomas, A., Best, N., and Gilks, W.R. (1995): BUGS: Bayesia Inference using Gibbs sampling, Version 0.50, Technical Report, Cambridge University, Medical Research Council Biostatistics Unit, Institute of Public Health.

Tanner M.A and Wong W.H (1987): The Calculations of Posterior distributions by dat augmentation, Journal of American Statistical Association 82, 528-540.

Tierney.L (1991) Markov Chains for exploring posterior distributions. Technical Report, School c Statistics, University of Minnessoter.