Roberts, G.O., Gelman, A., and Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7, 110–120.

Rosenthal, J.S. (1995a). Rates of convergence for Gibbs sampler for variance components models. *Ann. Statist.*, 23, 740–761.

Rosenthal, J.S. (1995b). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.*, 90, 558–566. [Correction, p. 1136.]

Rosenthal, J.S. (1996a). Convergence of Gibbs sampler for a model related to James-Stein estimators. *Statist. Comput.*, 6, 269–275.

Rosenthal, J.S. (1996b). Markov chain convergence: From finite to infinite. *Stochastic Process. Appl.*, 62, 55–72.

Rossky, P.J., Doll, J.D., and Friedman, H.L. (1978). Brownian dynamics as smart Monte Carlo simulation. *J. Chem. Phys.*, 69, 4628–4633.

Schervish, M.J., and Carlin, B.P. (1992). On the convergence of successive substitution sampling, *J. Comput. Graph. Statist.*, 1, 111–127.

Sinclair, A. (1992). Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combin. Probab. Comput.*, 1, 351–370.

Sinclair, A. (1993). *Algorithms for Random Generation and Counting: A Markov chain approach.* Birkhäuser, Boston.

Smith, A.F.M., and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B*, 55, 3–24.

Smith, R.L., and Tierney, L. (1996). Exact transition probabilities for the independence Metroplis sampler. Preprint, Dept. of Statistics, University of North Carolina at Chapel Hill.

Tanner, M.A., and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, 82, 528–550.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, 22, 1701–1762.

Tweedie, R.L. (1996). Truncation approximations of invariant measures for Markov chains. Preprint, Colorado State University.

Vattulainen, I., Ala-Nissila, T., and Kankaala, K. (1994). Physical tests for random numbers in simulations. Technical Report, Research Institute for Theoretical Physics, University of Helsinki, Finland.

Received 27 February 1997
Revised 13 August 1997
Accepted 13 August 1997

Statistical Laboratory
University of Cambridge
Cambridge, United Kingdom
CB2 1SB

email: G.O.Roberts@statslab.cam.ac.uk.

Department of Statistics
University of Toronto
Toronto, Ontario
Canada M5S 3G3

email: jeff@utstat.toronto.edu.

Discussion*

Hemant ISHWARAN

University of Ottawa

As we all know, there has been tremendous recent interest in applying MCMC methods in statistics, with an equally tremendous amount of literature written on the same topic. Simply trying to keep up with the new methods and technology is difficult enough, and

---

*This research was partially supported by grant funds from the Natural Sciences and Engineering Research Council of Canada.

it seems that there is very little time left over for the study and development of theory for guiding its use. But as the authors of the present paper have clearly indicated, a deeper understanding of MCMC theory is essential in selecting between competing methods and in using those methods efficiently and properly. I congratulate Roberts and Rosenthal on this point and will expand upon it by considering in detail the use of a hybrid Gibbs sampler in a specific Bayesian analysis. As we shall see, following some of the practical guidelines suggested by the authors can help in navigating through some of the MCMC problems in this particular example. The analysis, I hope, will also point to areas for further research.

The example that I will look at is the Rasch model for binary outcomes. My interest in the model will be from within the Bayesian paradigm, but I will follow the motivation and description of the model given in Lindsay *et al.* (1991). The interested reader will find a nice overview of the model in that paper as well as further references to other articles. Briefly, though, the Rasch model is an exponential model used in modeling 0-1 binary outcomes. One of its important uses is in item response studies where each individual i gives a binary response to each of $J \geq 2$ different items or questions. For example, the data which will be studied here are based on $n = 216$ individuals who gave 0-1 binary responses to a set of $J = 4$ questions involving role conflict [Stouffer and Toby (1951), and also analyzed in Lindsay *et al.* (1991, Table 1, column A)]. If $X_{i,j}$ is the 0-1 binary response for individual i to question j, then $X_{i,j}$ has the conditional density

$$f(x_{i,j}|\boldsymbol{\theta}, y_i) = \frac{\exp\{(\theta_j + y_i)x_{i,j}\}}{1 + \exp(\theta_j + y_i)} \qquad \text{for} \quad i = 1,\ldots,n \text{ and } j = 1,\ldots,J. \tag{1}$$

Here $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_J)$ is the vector of item response parameters measuring item difficulty, while y_i is the unique ability parameter for individual i. Notice that (1) implicitly implies that the $X_{i,j}$ are conditionally independent given the item difficulty parameter $\boldsymbol{\theta}$ and the individual parameter y_i.

Without some additional structure on the y_i-values, we would have too many parameters in the model (in total, $n+J$). One useful method to resolve this problem, used in Lindsay *et al.* (1991), is to assume that the Y_i are independent random variables with some unknown finitely discrete distribution G_0 (a finite discrete distribution is a distribution with a finite number of support points $K < \infty$). Such an assumption implies that the response values have a finite semiparametric mixture density. If $\mathbf{X}_i = (X_{i,1},\ldots,X_{i,J})$ is the vector of binary responses for individual i, then $\mathbf{X}_i$ has the Rasch semiparametric mixture density

$$f(\mathbf{x}_i|\boldsymbol{\theta}, G_0) = \int \prod_{j=1}^{J} f(x_{i,j}|\boldsymbol{\theta}, y)\, dG_0(y) = \sum_{k=1}^{K} p_{0,k} \prod_{j=1}^{J} f(x_{i,j}|\boldsymbol{\theta}, y_{0,k}). \tag{2}$$

A simple method to ensure that $\boldsymbol{\theta}$ is identified in this model is to constrain $\theta_J = 0$ to act as a baseline, although this will not guarantee identification for the mixing distribution G_0. Nevertheless, even though G_0 is nonidentified, Lindsay *et al.* (1991) show that one can still apply nonparametric maximum likelihood methods to properly estimate $\boldsymbol{\theta}$, and to a lesser extent to recover partial information about the unknown mixing distribution. The nonidentification of G_0 plays an important role in this analysis, and as we shall soon see, it also plays an important role in the application of our hybrid Gibbs sampler.

If the data $\mathbf{X} = (\mathbf{X}_1,\ldots,\mathbf{X}_n)$ are i.i.d r.v.'s with the finite mixture density (2), then the following Bayesian model can also be used to properly study $\boldsymbol{\theta}$. Assume that the prior

$\pi_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is independent of the prior $\pi_{\mathbf{Y}^n}$ for $\mathbf{Y}^n = (Y_1, \ldots, Y_n)$ and that the data can be represented in the hierarchical fashion

$$\begin{aligned} &\mathbf{X}_i|\boldsymbol{\theta}, Y_i \overset{\text{ind}}{\sim} f(\cdot|\boldsymbol{\theta}, Y_i), \qquad i = 1, \ldots, n, \\ &(\boldsymbol{\theta}, \mathbf{Y}^n) \sim \pi = \pi_{\boldsymbol{\theta}} \times \pi_{\mathbf{Y}^n}. \end{aligned} \tag{3}$$

If $\pi_{\boldsymbol{\theta}}$ is a continuous positive density and $\mathbf{Y}^n$ a sample from a Dirichlet process prior with a continuous positive concentration density, then the posterior for $\boldsymbol{\theta}$ will concentrate on each open neighbourhood of the true $\boldsymbol{\theta}$ with exponentially high probability (Ishwaran 1997). To remind the reader: a sample $\mathbf{Y}_n$ from a Dirichlet process prior with precision parameter $A > 0$ and concentration density h has a distribution which satisfies

$$\pi_{\mathbf{Y}^n}(dy_1, \ldots, dy_n) = h(y_1)\, dy_1 \prod_{i=2}^{n} \left(\alpha_i h(y_i)\, dy_i + \frac{1 - \alpha_i}{i - 1} \sum_{s=1}^{i-1} \delta(y_s, dy_i) \right), \tag{4}$$

where $\delta(y, \cdot)$ is the unit measure concentrated at y and $\alpha_i = A/(A + i - 1)$.

The exponential posterior consistency for $\boldsymbol{\theta}$ is comforting, but there are several practical issues that need to be resolved in computing the posterior for (3). Firstly, we will need to rely on MCMC methods for sampling the posterior of $\boldsymbol{\theta}$ and $\mathbf{Y}^n$ due to its intractability. This means that we need to resolve Practical Implication 1 of the paper: which sampler do we choose? Another serious consideration is the lack of identification for G_0, which can force the likelihood surface for the mixture model to be infinitely multimodal (indeed, this is what happens in the data that will be considered). This multimodality can have dire consequences for an MCMC-based method. However, we should remember that the posterior of interest may look vastly different than the likelihood because of the effect of the prior. Furthermore, the posterior under consideration is a function of the nuisance $\mathbf{Y}^n$ values and not the unknown mixing distribution G_0, as is the case for the frequentist likelihood surface. Nevertheless, the implications of a multimodal likelihood surface have to be seriously considered in an MCMC analysis.

Because of the large number of Y_i-values ($n = 216$), it seemed wise to use MacEachern's (1994) Gibbs sampling method for coarsening the state space (also see Müller *et al.* 1996) rather than using the more conventional method described in Escobar (1994). The key idea behind MacEachern's approach is that a sample $\mathbf{Y}^n$ from a Dirichlet process prior will have relatively few distinct values $\mathbf{Y}^* = (Y_1^*, \ldots, Y_L^*)$ (the expected number of distinct values $L \leq n$ depends directly upon the precision parameter A). Therefore, rather than working directly with the $\mathbf{Y}^n$-values in the Gibbs sampler, MacEachern's (1994) approach is to work with the fewer distinct values $\mathbf{Y}^*$ and a classification vector $\mathbf{C} = (C_1, \ldots, C_n)$ which keeps track of the cluster membership of each Y_i-value: $C_i = j$ if and only if $Y_i = Y_j^*$. This coarsens the state space of the sampler, because fewer $\mathbf{Y}$-values are involved, while the state space for $\mathbf{C}$ is a subset of the finite space $\{1, \ldots, n\}^n$. Coarsening the state space allows the sampler to mix more quickly, which seems to be a reasonable reason for using this method as directed by Practical Implication 1.

Following this reasoning, a hybrid Gibbs sampler based on MacEachern's (1994) method was used in order to investigate the posterior for $(\boldsymbol{\theta}, \mathbf{Y}^*, \mathbf{C}|\mathbf{X})$ from (3). The sampler that was used started with an initial value for $(\boldsymbol{\theta}, \mathbf{Y}^*, \mathbf{C})$ and then successively sampled from the conditional distributions

$$(\boldsymbol{\theta}\,|\mathbf{Y}^*, \mathbf{C}, \mathbf{X}) \propto f(\mathbf{X}|\boldsymbol{\theta}, \mathbf{Y})\, \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \tag{5}$$

$$(\mathbf{Y}^*|\boldsymbol{\theta},\mathbf{C},\mathbf{X}) \propto \prod_{l=1}^{L} h(Y_l^*) \prod_{\{i:C_i=l\}} f(\mathbf{X}_i|\boldsymbol{\theta},Y_l^*) \tag{6}$$

and the conditional distribution $(C_i|\boldsymbol{\theta},\mathbf{Y}^*,\mathbf{C}^{(i)},\mathbf{X})$ for $i = 1,\ldots,n$, where $\mathbf{C}^{(i)}$ is the classification vector $\mathbf{C}$ with the ith coordinate removed. A $\mathsf{N}(0,\sigma_{\boldsymbol{\theta}}^2)$ density was used for $\pi_{\boldsymbol{\theta}}$, and a $\mathsf{N}(0,\sigma_Y^2)$ density was used for the concentration density h. The choice of normal densities yields conditional distributions (5) and (6) that are not easy to simulate from, and so a Metropolis-Hastings step was used in order to sample from each of them. Let $\mathbf{Y}^{(i)n}$ denote the vector of $\mathbf{Y}^n$-values with the value Y_i removed, and let $Y_1^*,\ldots,Y_{L_i}^*$ be the L_i distinct values in $\mathbf{Y}^{(i)n}$ and $n_{i,l}$ the number of times these distinct values occur, $l = 1,\ldots,L_i$. Then the conditional distribution for C_i is discrete with sample space $\{1,\ldots,L_i+1\}$, where

$$\mathbb{P}\{C_i = l|\boldsymbol{\theta},\mathbf{Y}^*,\mathbf{C}^{(i)},\mathbf{X}\} \propto \begin{cases} n_{i,l} f(\mathbf{X}_i|\boldsymbol{\theta},Y_l^*), & l = 1,\ldots,L_i, \\ A f(\mathbf{X}_i|\boldsymbol{\theta},Y^*), & l = L_i+1, \end{cases}$$

and Y^* is an independent variable with the concentration density h [the many details omitted here can be found in MacEachern (1994)].

To complete the specifications of the model, the variances $\sigma_{\boldsymbol{\theta}}^2$ and $\sigma_{\mathbf{Y}}^2$ in the normal densities $\pi_{\boldsymbol{\theta}}$ and h were both set equal to 100. A large variance such as 100 is used to reflect uncertainty in the values for the parameters, and in the case of $\sigma_{\mathbf{Y}}^2$ ensures that the posterior for $\mathbf{Y}^n$ will contain a broad range of values with high probability. This leaves only the choice for the value of the precision parameter, A, still to be decided. A very small value of A will mean few distinct $\mathbf{Y}$-values, which can force the posterior to resemble a parametric likelihood surface. Too large a value for A, such as n, will mean almost as many clusters as observations (actually $\simeq n\log 2$ on average) and represents a Bayesian parametric approach. Therefore, in order to retain a nonparametric flavour, it was decided that A would be set at $\sqrt{n} \simeq 15$, which would force approximately 41 distinct values on average $[\sum_{i=1}^{n} A/(A+i-1)$: see Antoniak (1974, p. 1161)].

With a large $\sigma_{\mathbf{Y}}^2$ variance and many expected $\mathbf{Y}$-clusters, there is a possibility for the sampler to become stuck in a region of the parameter space for many iterations. Therefore, it is important to carefully choose the type of transition distribution used in the Metropolis steps for (5) and (6). My decision was to use a $\mathsf{N}(0,\tau_{\boldsymbol{\theta}}^2)$ and a $\mathsf{N}(0,\tau_{\mathbf{Y}}^2)$ transition distribution for each of the $\boldsymbol{\theta}$ and $\mathbf{Y}$ coordinates in (5) and (6), respectively. The value for $\tau_{\boldsymbol{\theta}}^2$ was set at the average variance of 0.25^2 observed for $\boldsymbol{\theta}$ using an EM algorithm for a two-point mixture (see Table 1). The likelihood surface at $\boldsymbol{\theta}$ looks similar for any mixture with at least two support points, and because we expect many $\mathbf{Y}$-cluster values it seemed appropriate to use the variance observed from the two-point mixture estimate. Choosing the value for $\tau_{\mathbf{Y}}^2$ is much more critical in assuring proper mixing. Too small a value, and the sampler can spend too much time moving between competing choices for $\mathbf{Y}^n$. Too large a value, and the sampler will quickly find a large mode and sit there for a very long time. Therefore, as a compromise, $\tau_{\mathbf{Y}}^2$ was also set at 0.25^2, which was approximately the variance seen in the estimates of the support points using our EM algorithm. This seemed somewhat *ad hoc*, but with such a complicated problem there did not seem to be a simple rule for choosing the scaling as suggested by Practical Implication 5.

For lack of space, I do not reproduce the data that are analyzed here; the interested reader will find them given in Lindsay *et al.* (1991, Table 1, column A). In Table 1 here are the results from fitting an EM algorithm to these data using one-, two-, three-

TABLE 1: Parameter estimates for $\boldsymbol{\theta}$ and log-likelihood values using an EM algorithm (Aitkin and Rubin 1985) in fitting one-, two-, three- and four-point Rasch semiparametric mixtures. We set $\theta_4 = 0$ to ensure identification for $\boldsymbol{\theta}$.

No. of support points	θ_1	θ_2	θ_3	Log likelihood
1	2.13	0.799	0.855	−543.65
2	2.87	1.25	1.32	−504.56
3	2.94	1.24	1.32	−503.65
4	2.94	1.24	1.32	−503.65

and four-point semiparametric mixtures of the form (2). As can be seen, the parameter estimates and values for the log likelihood are the same for the three- and four-point mixtures and almost the same for the two- and three-point mixture models. The equality between the three- and four-point mixtures is no coincidence: the same pattern persists in models with even more support points. Indeed, Lindsay *et al.* (1991) show that in this example there are an infinite number of nonparametric MLE solutions containing more than two support points and having the same $\boldsymbol{\theta}$-values. With such a multimodal likelihood surface, the EM algorithm becomes a useful method for checking whether the $\boldsymbol{\theta}$-values from the sampler are in the right part of the parameter space. It can also be used in monitoring convergence of the sampler using Gelman and Rubin's (1992) multiple-chain method. I used this method by starting four different chains using values for $\boldsymbol{\theta}$ generated by an overdispersed multivariate t-density (three degrees of freedom). The mean and variance for the overdispersed distribution were obtained from the EM algorithm's estimate for $\boldsymbol{\theta}$ from a two-point mixture. Starting values for the $\mathbf{Y}^n$ were generated randomly from (4) using our choice for A and concentration density h. These values were then used to start the four chains, which were first run for 75,000 iterations of burn-in, before being monitored every 40th iteration. Convergence for $\boldsymbol{\theta}$ was measured by evaluating Gelman and Rubin's (1992) potential scale reduction factor (with corrected degrees of freedom). In this case the scale reduction values were almost identically equal to 1, indicating convergence. Figure 1 contains the history of the $\boldsymbol{\theta}$-iterations and also indicates convergence in $\boldsymbol{\theta}$. The superimposed bivariate normal contours demonstrate the multivariate normality of the posterior for $\boldsymbol{\theta}$, a critical assumption in Gelman and Rubin's method.

It is also important to monitor the convergence of $\mathbf{Y}^n$, but Gelman and Rubin's method is inappropriate because of the nonnormality of the posterior for $\mathbf{Y}^n$. Furthermore, it is computationally too burdensome to monitor the scale reduction for so many values ($n = 216$). Unfortunately, there seems to be no other general, simple method for monitoring the behaviour of the $\mathbf{Y}^n$-values, which seems surprising given the interest in the Dirichlet process prior. Therefore, I was forced to resort to an *ad hoc* method, which consisted in plotting the cluster values and number of values in each cluster for clusters containing a substantial fraction of the data ($\geq 10\%$). Figure 2 contains this information for each of the four chains, and seems to indicate that the sampler is moving between clusters of size 1, 2 and 3 with $\mathbf{Y}$-values predominately selected between -3 and -0.5. The experience of the four chains looks similar which provides some evidence for convergence, but could be misleading as a diagnostic tool in general. The ideal solution would have been to derive quantitative rates of convergence, but I suspect this is very hard because of the complexity of the model. This leaves convergence diagnostics as the only other possibility, where

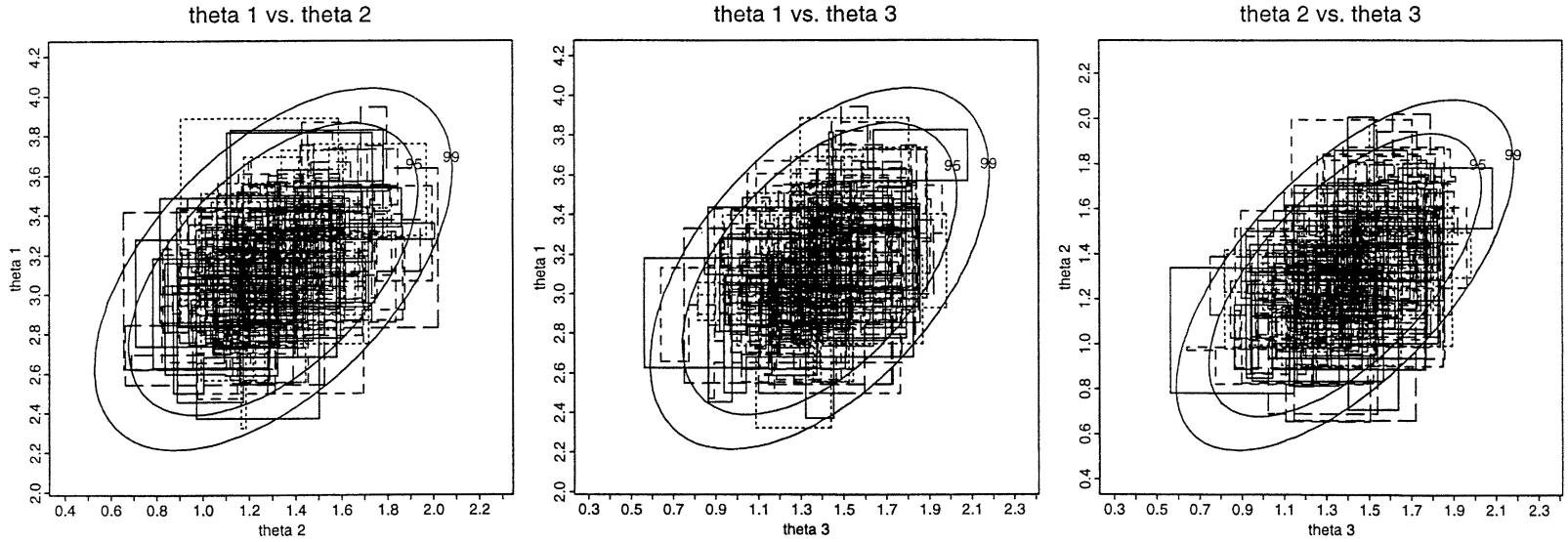

FIGURE 1: History of 2500 $\boldsymbol{\theta}$-values generated from the hybrid Gibbs sampler for four different chains, with each chain started from an overdispersed starting distribution. Steps are given every 40th iteration after a burn-in sample of 75,000. The 95% and 99% bivariate normal contours for the $\boldsymbol{\theta}$-values are superimposed. Note that the modes agree closely with the estimates for $\boldsymbol{\theta}$ from Table 1 for the two-, three- and four-point mixtures.

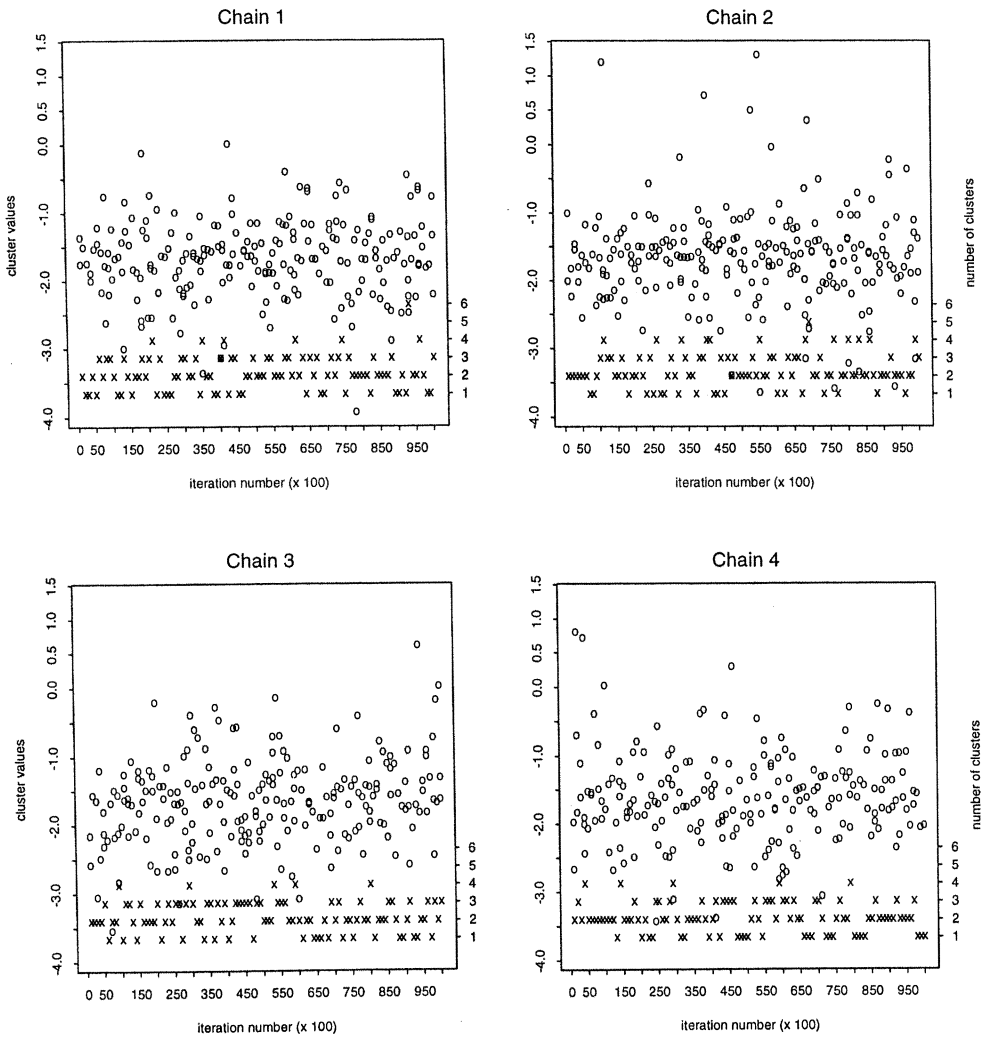

FIGURE 2: The trimmed cluster values (left axis and indicated by o) and number of values in each trimmed cluster (right axis and indicated by ×) for **Y** are given for the four chains in Figure 1 every 1000th iteration after the same 75,000-iteration burn-in. Only clusters containing at least 10% of the **Y**-values are shown. For clarity, some large cluster values are hidden.

there is still room for much work, especially for MCMC methods involving the Dirichlet process prior.

REFERENCES

Aitkin,

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, 2, 1152–1174.

Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.*, 89, 268–277.

Gelman, A., and Rubin D.B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.*, 7, 457–511.

Ishwaran, H. (1997). Exponential posterior consistency via generalized Polya urn schemes in finite semiparametric mixtures. Technical Report 306, Laboratory for Research in Statistics and Probability, Carleton University–University of Ottawa.

Lindsay, B., Clogg, C.C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Amer. Statist. Assoc.*, 86, 96–107.

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Commun. Statist.—Simula.* 23, 727–741.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83, 67–79.

Stouffer, S.A., and Toby, J. (1951). Role conflict and personality. *Amer. J. Sociol.*, 56, 295–306.

Department of Mathematics and Statistics
University of Ottawa
Ottawa, Ontario
Canada K1N 6N5

email: ishwaran@capresso.mathstat.uottawa.ca

Discussion*

Neal MADRAS

York University

Gareth Roberts and Jeff Rosenthal have given us a pretty clear picture of how some recent theoretical results relate to practical issues in the design and implementation of Monte Carlo studies. They admit there there is still a long way to go before the theory can handle the complexity of many typical statistical simulations, but they do have some specific, though modest, guidelines based on theory.

Let me first address the example that they used to illustrate geometric convergence: the independence sampler whose target distribution is exponential with parameter 1, and whose proposals are exponential with parameter k. The present paper shows that the central limit theorem fails when $k > 2$. Would it be fairer to blame this on the fact that the asymptotic variance σ_g^2 is infinite? It is possible to have chains which are not geometrically ergodic but in which the central limit theorem does hold. In principle, they could be pretty efficient, but I don't know how common such chains are in practice.

Also, as pointed out by Liu (1996), the independence sampler is very closely related to the more traditional Monte Carlo methods of importance sampling and rejection sampling. If we tried to do importance sampling by generating exponentials with parameter k and reweighting them to estimate the mean of an exponential distribution with parameter 1, then the central limit theorem would fail for $k \geq 2$, for the simple reason that a single observation would have infinite variance. At the other extreme, when k is small, the efficiency would be roughly proportional to k, as is the case with the independence sampler, and with rejection. Liu gave some relations in the form of inequalities, but there is no full correspondence among these methods in general. It would be interesting to try to refine Liu's work for some broad classes of realistic examples.

---

*This work was supported in part by a Research Grant from NSERC, and by strike funds from the York University Faculty Association and the Canadian Association of University Teachers Defense Fund.