Approximate pre-determined convergence properties of the Gibbs sampler

Gareth O. Roberts and Sujit K. Sahu

This article aims to provide a method for approximately *pre-determining* convergence properties of the Gibbs sampler. This is to be done by first finding an approximate rate of convergence for a normal approximation of the target distribution. The rates of convergence for different implementation strategies of the Gibbs sampler are compared to find the best one. In general, the limiting convergence properties of the Gibbs sampler on a sequence of target distributions (approaching a limit) are not the same as the convergence properties of the Gibbs sampler on the limiting target distribution. Theoretical results are given in this article to justify that under conditions, the convergence properties of the Gibbs sampler can be approximated as well. A number of practical examples are given for illustration.
KEY WORDS: EM algorithm; Gaussian distribution; Generalized linear models; Hierarchical centering; Laplace approximation; Markov chain Monte Carlo.

1 Introduction

The Gibbs sampler is the most popular Markov chain Monte Carlo (MCMC) algorithm in Bayesian computation because of its ease of implementation and universal applicability. However, analytic analyses, ideally needed for convergence assessment, of Gibbs sampler Markov chains are rarely possible, and existing techniques for providing upper bounds on *convergence times*, perhaps by means of eigenvalue estimates, (see for example Rosenthal, 1995 and Roberts and Tweedie, 1999) are not powerful enough to deal with the demands set by the ever more ambitious statistician. Indeed, it is unlikely that theoretical work on convergence of MCMC algorithm will ever be able to keep pace with increasingly complex statistical models used in practice. Therefore approximate approaches which can cover large classes of problems occurring in practice are appealing. The central methodological idea in this paper is the analysis of target density correlation structure in order to

- predict a-priori required MCMC run lengths;
- guide design of MCMC algorithms for instance in terms of parameterization, and the use of blocking schemes and alternative component update ordering schemes.

This paper will attempt to exploit synergies between the Gibbs sampler and various forms of the EM algorithm (Dempster *et. al.*, 1977). Specifically, for suitably approximately Gaussian target distributions, the convergence properties of the two algorithms are closely related. We prove a more precise statement of this result, and illustrate uses of this result in a number of examples.

The rate of convergence describes the limiting convergence properties of the MCMC algorithm. For example, the rate of convergence together with the starting value determines the number of iterations needed for convergence. Let $||\cdot||$ define a suitable divergence measure, e.g. the total variation norm, and $K^t(x,y)$ denote the t step transition kernel of the MCMC algorithm with unique equilibrium distribution $\pi(y)$. The rate of convergence, denoted by ρ, is defined as

$$\rho = \inf_{0<\xi\leq 1} \left\{\xi;\ \exists \text{ a function } V(\cdot) \text{ such that } ||K^t(x,\cdot)-\pi(\cdot)|| \leq V(x)\xi^t, \ \text{ for all } t>0\right\}.$$

The requirement that $V(x)\rho^t < \epsilon$, for some small ϵ dictates that the algorithm should be run a suitable multiple of $-(\log \rho)^{-1}$ number of iterations in order to achieve the given accuracy ϵ. The multiplier may depend on the starting point x, the function $V(\cdot)$ and the desired accuracy ϵ.

The efficiency in estimating expectations under the target distribution is also related to ρ. Suppose that $\bar{g}_N$, based on N iterates of the Markov chain X_n, is used to estimate $\mathbb{E}_\pi[g]$ for some integrable function g. The variance of $(\bar{g}_N)$ can be bounded as follows,

$$\text{Var}\ (\bar{g}_N) \leq \frac{\sigma^2}{N}\frac{1+\rho}{1-\rho}, \tag{1}$$

where σ^2 is the variance of $g(\mathbf{X})$ under π, see e.g. Polson (1996). In fact the inequality (1) is tight in the case where $g(\mathbf{X}_n)$ is an approximate first order autoregressive process, see e.g. Tierney (1994). Hence, the higher the rate of convergence, the more iterates are needed to reduce the Monte Carlo variance (though see Green and Han (1992) for a fuller understanding of the difference between the speed of convergence and efficiency in estimation).

It would of course be preferable to compare and assess samplers directly in terms of the chosen divergence measure. However such an approach is complicated, not least by the fact that such a divergence typically depends on the starting value of the chain, x. Instead we shall use ρ as a surrogate for assessing convergence performance. This has two crucial advantages. As well as being independent of x, convergence rates are readily accessible in the Gaussian case which we shall use to approximate the convergence properties of the sampler.

A large proportion of posterior densities in Bayesian analyses in practice are unimodal. For example, Sahu (2000) points out that if all the one dimensional conditional distributions are log-concave (i.e. unimodal) then the associated marginal distributions are also log-concave. This general result certainly covers the popular generalized linear models since the conditional distributions for such models are log-concave, see e.g. Dellaportas and Smith (1993).

For unimodal posterior densities results on the asymptotic normality in Bayesian inference are well known, see e.g. Bernardo and Smith, (1994, page 291). These are particularly useful for studying the convergence properties of the Gibbs sampler by studying the sampler for an approximate Gaussian distribution. This is because convergence properties of the Gibbs sampler are invariant under location and scale changes along coordinate directions, and are therefore only dependent on the correlation structure of the Gaussian target distribution. However, it is precisely this which is held fixed asymptotically for posterior distributions as the number of data points increase.

The primary objective of the current article is to use asymptotic normality of the posterior distributions to study convergence properties of the associated Markov chains. It is easy to envisage that if a sequence of target densities approaches a certain limit, then the associated sequence of Markov transition kernels will approach the limiting transition kernel for which the invariant distribution is the limiting target distribution. However, *this is not true in general* and it is easy to contrive examples illustrating this. A theoretical result obtained in Section 4 shows that for Gibbs sampler Markov chains the sequence of transition kernels do approach the limiting transition kernel under rather general conditions.

Practical implications of the above result are wide ranging. It provides theoretical justification for extrapolating the results obtained in Roberts and Sahu (1997) to non-Gaussian cases. Many practical issues relating to the design of the Gibbs sampler, e.g. parameterization, blocking and updating schemes can be decided by studying the correlation structure of the target distribution. Observe that, this is to be done before running the Gibbs sampler. In this sense, the proposed methodology does not directly fall in the suite of convergence diagnostics, rather it is a technique for pre-determining approximate convergence properties, and thus for guiding sampler design as a result.

The convergence properties of the Gibbs sampler are strongly linked with the statistical properties of the distribution to be explored. Hills and Smith (1993) and Gelfand *et al.* (1996) noted that reducing correlations between components in the target distribution by re-parameterization could have dramatically beneficial consequences for the speed of convergence of the algorithm. Roberts and Sahu (1997) perform a detailed analysis of the Gibbs sampler on Gaussian target densities. Because of the need for relevance to non-Gaussian densities, they concentrate on qualitative results which might be of use in much more general contexts. Specifically, they give comparison results on convergence rates between the different samplers according to the correlation structure of the target density, and results regarding random updating of components and the choice of blocking strategy to be adopted.

The issue of adequacy of the sample size required for obtaining good Gaussian approximations for the purpose of estimating the rate of convergence is not at all pursued in this paper, though it is clearly important. There are methods available for checking the validity of such approximations and finding re-parameterization schemes which lead to the posterior being closer to normality, see e.g. Hills and Smith (1993).

This paper considers only continuous unimodal posterior distributions. The proposed method clearly is not applicable in the case of discrete distributions which arise in many statistical problems. Also, it is not suitable for the Metropolis-Hastings algorithm. Further, it will not detect multi-modality if this exists in the target distribution. However, accurate analytic estimation of the target correlation structure may indicate such problems.

Finally, this paper demonstrates the proposed method with four examples, all based on the generalized linear models. The first two examples aim to provide guidelines for determining the Gibbs sampler burn-in period. Convergence assessment approaches used by other authors on the same examples are also compared. The third and the fourth example (based on high dimensional target distributions) demonstrate the benefits of re-parameterization schemes. Our methodology obtains the approximate target correlation structure and the rates of convergence of the Gibbs samplers for different parameterizations.

The plan of the remainder of this article is as follows. In Section 2 the proposed methodology is described in detail. Section 3 considers four examples to illustrate the effectiveness of the method. To enhance readability, the main technical result and its proof is placed in Section 4. We conclude with a few summary remarks in Section 5.

2 Proposed Methodology

2.1 Approximate Correlation Structure

In order to develop our methodology we require a reasonably easily implementable method for approximating the asymptotic correlation structure of the posterior distribution. Here we outline a method (see e.g. Bernardo and Smith, 1994, page 291) how such approximations can be obtained.

Let $\pi_n(\mathbf{x})$ denote the posterior (target) distribution based on a random sample of size n and let $L_n(\mathbf{x}) = \log \pi_n(\mathbf{x})$. Further, let $\mathbf{m}_n$ denote the posterior mode and

$$Q_n = -L_n''(\mathbf{m}_n) \tag{2}$$

where $L_n''(\mathbf{m}_n)$ is the Hessian matrix with elements

$$[L_n''(\mathbf{m}_n)]_{ij} = \left(\partial^2 L_n(\mathbf{x})/\partial x_i \partial x_j\right)|_{\mathbf{x}=\mathbf{m}_n}.$$

Under suitable regularity conditions, see e.g. Proposition 5.14 in Bernardo and Smith $\boldsymbol{\phi}_n = Q_n^{-1/2}(\mathbf{x}-\mathbf{m}_n)$ converges in distribution to a standard normal random variable.

In many applications, the Hessian matrix Q (henceforth we drop the subscript n for notational convenience) may be easy to obtain in analytic algebraic form (for example in parametric exponential family models) by simple differentiation. However, it is usually non-trivial to find the posterior mode, $\mathbf{m}$ using analytic arguments directly. Following Sahu and Roberts (1999) we propose to use the EM algorithm (Dempster *et al.*, 1977, Meng and vanDyk, 1997) to find the mode and hence the Hessian matrix Q.

2.2 Rate of Convergence

Let π denote the limiting Gaussian density with mean $\boldsymbol{\mu}$ and variance-covariance matrix Σ. Also let $Q = \Sigma^{-1}$ and define Q_L and Q_U by $Q = Q_L - Q_U$ where Q_L be the block lower triangular matrix with blocks in the lower triangle being identical to those of Q and Q_U is block upper triangular with null diagonal blocks. That is Q_L contains the diagonal blocks of Q. Further, let $\mathbf{b} = (I - B)\boldsymbol{\mu}$ where

$$B = Q_L^{-1} Q_U. \tag{3}$$

Roberts and Sahu (1997) show that the Markov chain induced by the Gibbs sampler has a normal transition density with mean $E(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}) = B\mathbf{x}^{(t)} + \mathbf{b}$ and dispersion $\Sigma - B\Sigma B'$. Thus $\{\mathbf{x}^{(t)}\}$ is a multivariate AR(1) process. It follows that the rate of convergence of the Gibbs sampler is $\rho(B)$, the spectral radius of the matrix B. See Roberts and Sahu (1997) for further discussion about this rate of convergence and bibliographical notes about the matrix B.

2.3 Implementation

The proposed methodology is implemented as follows. First, we obtain the approximate correlation structure following the methods outlined above. Then we use this approximate correlation structure of the target distribution to obtain the associated approximate rate of convergence.

```
STEP I:    Find the posterior mode (e.g.  using the EM algorithm).
STEP II:   Obtain the Hessian matrix Q using (2) evaluated at the mode.
STEP III:  Obtain the matrix B using (3).
STEP IV:   Obtain the approximate rate of convergence ρ, the spectral radius of the
           matrix B.
```

The approximate rate of convergence so obtained can be used in several ways. Firstly, $-(\log \rho)^{-1}$ provides an indication of the required burn-in period. Secondly, this also helps to obtain guidelines for the number of iterates to be used for forming the ergodic averages. For example, $N \times \frac{1+\rho}{1-\rho}$ of iterates are approximately equivalent to N i.i.d. samples from the target distribution, see equation (1). Lastly, the rates of convergence for different competing parameterizations, updating schemes and blocking can be used to find the optimal strategy. In fact in step II, it may be possible in some situations to re-parameterize the target density to approximate independence, leading to an extra step between II and III. However, this often destroys any conjugacy present in the problem and consequently the implementation of the Gibbs sampler becomes difficult.

3 Examples

3.1 Pump Failure Data Example

This example demonstrates a case where the rate of convergence of the Gibbs sampler by Gaussian approximation is very close to the true rate of convergence. We compare this approximate rate of convergence with the rate of convergence obtained by Garren and Smith (2000). They obtain the rate of convergence by using statistical techniques to estimate the principal eigenvalue of the induced Markov chain from the output of the Gibbs sampler.

This example, relating to 10 power plants, has been studied by Gelfand and Smith (1990) in the Gibbs sampling context using data from Gaver and O'Muircheartaigh (1987). The number of failures, y_i, follows a Poisson distribution with mean $\theta_i t_i$ where θ_i is the failure rate for pump $i, i = 1, \ldots, 10$ and t_i is the length of operation time of the pump (in 1000s of hours). A conjugate gamma prior distribution $\Gamma(\alpha, \beta)$ with density $\beta^\alpha \theta_i^{\alpha-1} \exp\{-\beta\,\theta_i\}/\Gamma(\alpha)$ is adopted for each θ_i. The

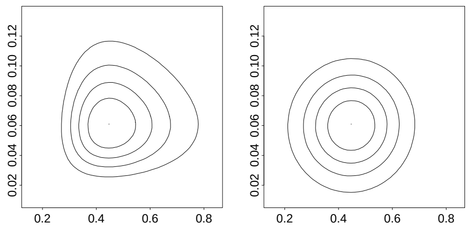

Figure 1: Contour plots of marginal joint distribution of θ_1 and β in the pump failure example. Left panel: exact contours; right panel: approximating contours.

parameter β is assigned a $\Gamma(0.01, 1)$ prior distribution. Moreover, the parameter α is fixed at 1.802 which is its method of moments estimator.

Here, it is straightforward to find the posterior mode and the Hessian matrix evaluated at the mode. Observe that the θ_i's are conditionally independent given β. Figure 1 shows the exact and approximate bivariate marginal contour plot of θ_1 and β. The plotted contours agree very closely near the mode but the exact one (left panel of Figure 1) has heavier tails than its normal approximation.

Though the Gaussian approximation is not entirely satisfactory, we nevertheless calculate the rate of convergence for this example. We report that the approximate rate of convergence for this problem is 0.299. This is almost equivalent to the approximate rate of convergence 0.3 obtained by Garren and Smith (2000). Rosenthal (1995), using detailed analysis of the Gibbs sampler for this case, obtains a theoretical upper bound of 0.976 on the rate of convergence.

3.2 Binary Regression Example

The main objectives of this example are to obtain guidelines about the Gibbs sampler burn-in period and to study the convergence behavior of the Gibbs sampler when the precision of the prior distribution increases. The example, originally considered by Knuiman and Speed (1988) and reanalyzed by Dellaportas and Smith (1993), is a quadratic logistic regression model. The data given in Table 1 of Dellaportas and Smith consist of 8 groups of durations of diabetes and a division of the number of individuals into two categories, with and without Retinopathy. The basic model formulation is as follows. Let $x_{1j}, x_{2j}, j = 1, \ldots, 8$ denote the current number of individuals with and without Retinopathy for the jth duration group. Let z_j denote the mid-duration for the jth group. Let π_{1j}, π_{2j} be the probabilities of being with or without Retinopathy for the individuals in the jth duration group. Then the following quadratic regression model

$$\log\left(\frac{\pi_{1j}}{\pi_{2j}}\right) = \beta_1 + \beta_2 z_j + \beta_3 z_j^2 = \eta_j$$

is suggested with a prior for $\boldsymbol{\beta}$ being $N(\boldsymbol{\beta}_0, D_0)$, where

$$\boldsymbol{\beta}_0 = \begin{pmatrix} -3.17 \\ 0.33 \\ -0.007 \end{pmatrix}, \; D_0 = \frac{1}{n_0} \times \begin{pmatrix} 638 & & \\ -111 & 24.1 & \\ 3.9 & -0.9 & 0.04 \end{pmatrix}$$

and n_0 is the prior precision. The joint posterior density has the form

$$\pi(\beta|\text{data}) \propto \exp\left[-\frac{1}{2}(\beta-\beta_0)D_0^{-1}(\beta-\beta_0)+\sum_{j=1}^{8}\{x_{1j}\eta_j-(x_{1j}+x_{2j})\log(1+e^{\eta_j})\}\right].$$

Observe that the above posterior density can be well approximated by a Gaussian distribution, because the prior on β is normal and the likelihood contributions can be approximated by a normal distribution due to the asymptotic normality of the maximum likelihood estimates (mle).

We find the variance-covariance estimates of the mle and then we adjust this by the prior variance-covariance. More particularly, we simply add up the precision matrix of the mle and the precision matrix of the prior together to obtain the posterior precision matrix. We use the resulting dispersion matrix to estimate the rate of convergence. Note that we are unable to use the posterior dispersion matrices given by Dellaportas and Smith (1993) directly because they are not positive definite possibly due to numerical problems.

Table 1 gives the rates of convergence for different values of n_0. As expected the rates go down as prior information become more precise. Dellaportas and Smith use $n_0 = 10^4$. For this choice of n_0 we obtain an approximate rate of convergence of 0.956 (see Table 1) for the deterministic updating scheme. Hence if we want a precision of 0.0001 we need to run the Gibbs sampler approximately $\log(0.0001)/\log(0.956) = 205$ number of iterations before starting to make ergodic averages. Note that Dellaportas and Smith suggest a burn-in period of 150 iterations for this example.

	$\log_{10} n_0$				
	2	3	4	5	6
ρ	0.964	0.963	0.956	0.933	0.920
$\frac{-1}{\log\rho}$	27.3	26.5	22.2	14.4	12.0
$\frac{1+\rho}{1-\rho}$	54.6	53.1	44.5	28.9	24.0

Table 1: Approximate convergence rates for the binary regression example.

3.3 Seed Data Example

In this example we find the rate of convergence of the Gibbs sampler by first running the EM-algorithm (see e.g. Sahu and Roberts, 1999). The rates of convergence for different parameterizations are compared to find the optimal one. We consider the seed germination data of Crowder (1978, Table 3). The probability of seed germination p_i in the ith plate is modeled as

$$\text{logit}[p_i] = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \alpha_{12} x_i z_i + b_i,\ i = 1, \ldots, 21, \tag{4}$$

where x_i and z_i are seed and extraction indicators, respectively. Further, we specify that $b_i \sim N(0, \sigma^2)$. Following the implementation in BUGS, Spiegelhalter *et al.*, (1996) we assume that σ^{-2} has a gamma distribution $\Gamma(0.001, 0.001)$. We label this setup NCEN (not-centered).

We consider two other parameterizations for this example. Since the model (4) does not allow full hierarchical centering as proposed by Gelfand *et al.* (1996) we consider the following partial centering. We assume that

$$\text{logit}[p_i] = \alpha_1 x_i + \alpha_2 z_i + \alpha_{12} x_i z_i + b_i,\ i = 1, \ldots, 21,$$

where $b_i \sim N(\alpha_0, \sigma^2)$. We label this parameterization HCEN (hierarchically centered).

When fitting the above type of linear models it is sometimes desirable to constrain the random effects to sum to zero, i.e., b_i's are constrained to satisfy $\sum b_i = 0$, although this changes the posterior distribution slightly. The BUGS implementation for this problem also considers this parameterization. All other assumptions are the same as in the NCEN case. We label this SUMZ (sum to zero). Using the rates of convergence for all three parameterizations we obtain the best one as follows.

We first obtain the mode using an approximate EM-algorithm of Sahu and Roberts (1999). Now we evaluate three Hessian matrices under three different parameterizations (after suitably transforming the modal estimates). The rates of convergence are given in Table 2. Several comments are in order.

1. The NCEN parameterization is the worst. The other two are about three times better than the NCEN. This confirms theoretical results of Roberts and Sahu (1997) for similar Gaussian problems. This also highlights the worth of considering parameterizations other than the default (NCEN) for similar problems.

2. The HCEN parameterization is perhaps not the best one. It loses to SUMZ because full hierarchical centering (as could be done for a fully nested model) is not possible for this problem.

3. The SUMZ parameterization turns out to be slightly better than the HCEN because: (i) it enhances the identifiability of the parameters as in usual linear models; (ii) it reduces the dimension of the problem (since $\sum b_i = 0$).

4. Since the rates of convergence of the last two parameterizations are almost equivalent there would not be much difference in the Gibbs sampler performances under the two.

	ρ	$\frac{-1}{\log \rho}$	$\frac{1+\rho}{1-\rho}$
NCEN	0.9697	33	65
HCEN	0.9139	12	22
SUMZ	0.9068	11	20

Table 2: Rates of convergence for the seed data example.

3.4 Epilepsy Data Example

Finally we consider a Poisson generalized linear mixed model to demonstrate the effectiveness of parameterization in improving the performance of the Gibbs sampler. Breslow and Clayton (1993) analyze data initially provided by Thall and Vail (1990) concerning seizure counts in a randomized trial of anti-convulsion therapy in epilepsy. We consider the model III postulated there. The seizure count y_{jk} for the patient $j (j = 1, \ldots, 59)$ on the kth $(k = 1, \ldots, 4)$ visit is modeled as follows.

$$
\begin{aligned}
y_{jk}|\mu_{jk} &\sim \text{Poisson } (\mu_{jk}) \\
\log(\mu_{jk}) &= \alpha_0 + \alpha_{Base} \log(\text{Base}_j/4) + \alpha_{Trt} \text{ Trt}_j + \alpha_{BT} \log(\text{Base}_j/4) \times \text{ Trt}_j + \alpha_{Age} \log(\text{Age}_j) \\
&\quad +\alpha_{V4}\text{V4}_k + b1_j + b_{jk} \\
b_{jk}|\sigma_b^2 &\sim \text{Normal } (0, \sigma_b^2) \\
b1_j|\sigma_{b1}^2 &\sim \text{Normal } (0, \sigma_{b1}^2)
\end{aligned}
$$

Priors on all the α parameters are assumed to be normally distributed with mean 0 and variance 10^4. The parameters σ_{b1}^{-2} and σ_b^{-2} are assumed to have Gamma distribution $\Gamma(0.001, 0.001)$ as in the BUGS implementation. Figure 2 gives a graphical representation of this model. We label this parameterization NCEN.

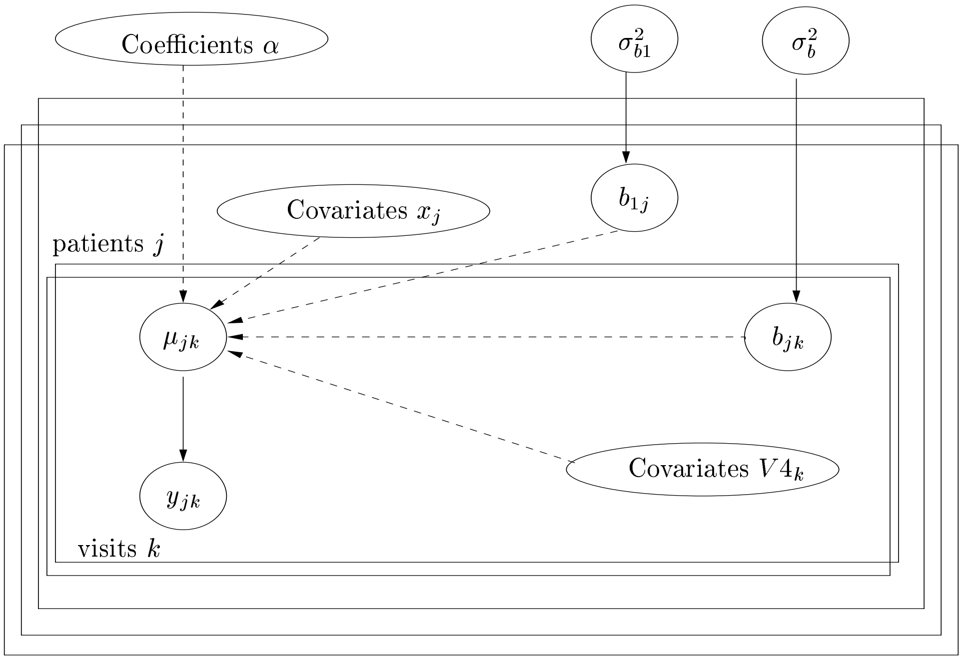

Figure 2: Graphical model for the Epilepsy data. Standard parameterization.

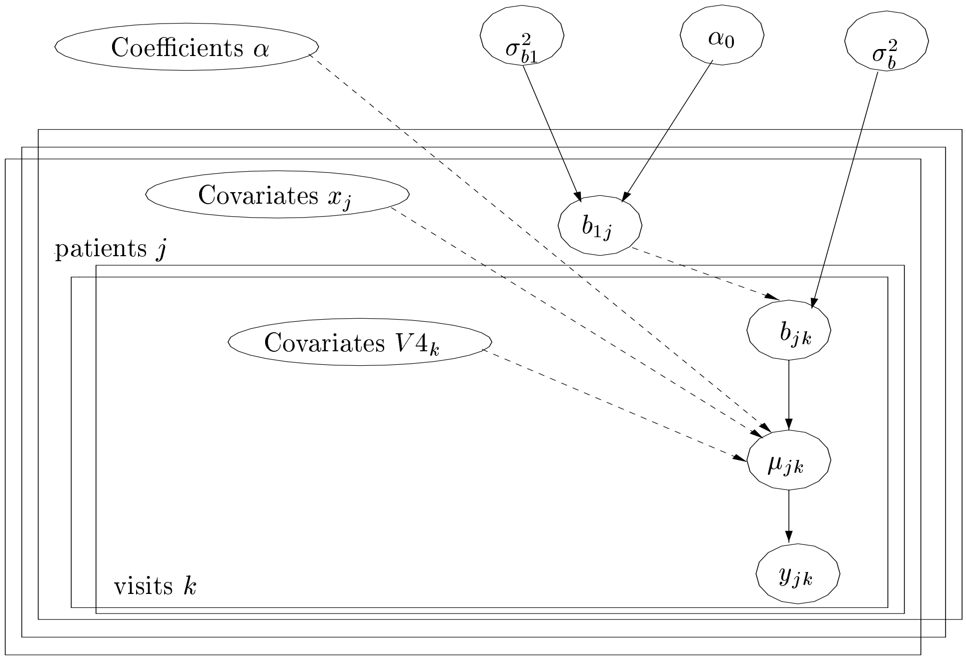

Figure 3: Graphical model for the Epilepsy data. Hierarchically centered parameterization.

The Gibbs sampler for this problem is challenging and runs on 303 dimensions. As noted in the `BUGS` manual, running time and convergence of the Gibbs sampler for this problem are extremely slow. Further, without centering the covariates it is almost impossible to to obtain convergence. Therefore, we center all covariates corresponding to the α parameters except for the intercept α_0. Hence we assume that the covariates in the above presentation are centered, i.e., for example Trt_j above have already been replaced by $\mathrm{Trt}_j - \mathrm{mean}(\mathrm{Trt})$.

We consider the following partial *hierarchically centered* parameterization (see e.g. Gelfand *et al.*, 1996) since the full model can not be written in a nested fashion.

$$
\begin{aligned}
y_{jk}|\mu_{jk} &\sim \text{Poisson }(\mu_{jk}) \\
\log(\mu_{jk}) &= \alpha_{Base}\log(\mathrm{Base}_j/4) + \alpha_{Trt}\ \mathrm{Trt}_j + \alpha_{BT}\log(\mathrm{Base}_j/4)\times\ \mathrm{Trt}_j + \alpha_{Age}\log(\mathrm{Age}_j) \\
&\quad +\alpha_{V4}\mathrm{V4}_k + b_{jk} \\
b_{jk}|b1_j,\sigma_b^2 &\sim \text{Normal }(b1_j,\sigma_b^2) \\
b1_j|\alpha_0,\sigma_{b1}^2 &\sim \text{Normal }(\alpha_0,\sigma_{b1}^2)
\end{aligned}
$$

All other model (prior) assumptions are the same as in the NCEN case. Figure 3 gives the graphical representation of this model. Comparing this with Figure 2 we see that the centered parameterization induces more conditional independence among the parameters. This enhances the identifiability of individual parameters in the posterior surface. On the other hand, such clear identification is not possible in the case of the standard parameterization. We label this HCEN.

The rates of convergence for the two parameterizations are given in Table 3. It is seen that the HCEN provides an order of magnitude improvement over the NCEN. For example, it may be concluded that HCEN is more than four times more efficient than the NCEN. The last column shows that under NCEN, about 2000 iterates are equivalent to one exact sample from the target distribution.

	ρ	$\frac{-1}{\log\rho}$	$\frac{1+\rho}{1-\rho}$
NCEN	0.9990	1000	1999
HCEN	0.9954	217	434

Table 3: Rates of convergence for the Epilepsy data example.

4 Theoretical Justification

Let $\pi_n(\mathbf{x})$ denote the posterior density of the k-dimensional vector of parameters $\mathbf{x}$ based on an observed data set of size n. Let $P_n(\cdot,\cdot)$ denote the Gibbs sampler transition kernel with unique stationary density π_n. Suppose that

$$\lim_{n\to\infty}\pi_n(\mathbf{x}) = \pi(\mathbf{x}) \tag{5}$$

for all continuity points $\mathbf{x}$ of π. Let $P(\cdot,\cdot)$ denote the Gibbs sampler transition kernel with positive transition density, so that the Markov chain has a unique stationary density $\pi(\mathbf{x})$ (up to modification on null sets), see Roberts and Smith (1994).

One may expect that the convergence properties of P_n and P are similar because of (5). However, in general this is not true. The condition (5) alone does not guarantee that P_n converges to P, and it is straightforward to contrive examples which demonstrate this.

Let $\mathbf{x}_{-i}$ denote the vector $\mathbf{x}$ with the ith coordinate deleted, and write $\pi_n^{(i)}$ for the conditional density:

$$\pi_n^{(i)}(x_i|\mathbf{x}_{-i}) = \frac{\pi_n(\mathbf{x})}{\int_{\mathbb{R}}\pi_n(\mathbf{x})dx_i}.$$

We need the following set of conditions to establish a theorem that the convergence properties of P_n can be well approximated by the convergence properties of P.

(A1) $\{\pi_n,\ n \in \mathbb{N}\}$ is a sequence of positive continuous $\mathbb{R}^k$-valued probability densities.

(A2) $\lim_{n\to\infty} \pi_n = \pi$ exists point-wise.

(A3) For all $1 \le i \le k$, and for almost all $\mathbf{x}_{-i}$

$$\lim_{n\to\infty} \pi_n^{(i)}(\cdot|\mathbf{x}_{-i}) = \pi^{(i)}(\cdot|\mathbf{x}_{-i})$$

exists point-wise as well as in L^1. This is very natural in a Bayesian context. Essentially this assumes that nested models within the larger k-dimensional model also satisfy asymptotic normality. Usually (A3) holds whenever (A2) does.

Recall that the total variation distance between two probability measures, ν_1 and ν_2 on a measure space $(E, \mathcal{E})$ is:

$$\|\nu_1 - \nu_2\| = 2 \sup_{A \in \mathcal{E}} (\nu_1(A) - \nu_2(A)) \ .$$

Let $T = T(\epsilon, \mathbf{x})$ be the smallest integer such that

$$\|P^T(\mathbf{x}, \cdot) - \pi(\cdot)\| < \epsilon$$

for all $\epsilon > 0$ and $T_n = T_n(\epsilon, \mathbf{x}),\ n \in \mathbb{N}$ be a collection of integers such that T_n is the smallest integer with

$$\|P_n^{T_n}(\mathbf{x}, \cdot) - \pi_n(\cdot)\| < \epsilon.$$

Further let, $S = S(\epsilon, \mathbf{x})$ be the smallest integer such that

$$\|P^S(\mathbf{x}, \cdot) - \pi(\cdot)\| \le \epsilon.$$

Since P is a transition kernel with positive density, $\|P^m(\mathbf{x}, \cdot) - \pi(\cdot)\| \to 0$ monotonically as $m \to \infty$, so that $|T - S| = 0$ or 1.

The fundamental theoretical result of this paper is therefore:

Theorem 1 *Suppose that (A1)–(A3) hold. Then for almost all* $\mathbf{x} \in \mathbb{R}^k$*,*

$$T - 1 \le S \le \liminf_{n\to\infty} T_n \le \limsup_{n\to\infty} T_n \le T.$$

Proof: The proof proceeds by constructing an appropriate probability space, supporting Markov chains started at $\mathbf{x}$, $\mathbf{X}$ and $\{\mathbf{X}^n,\ n \in \mathbb{N}\}$ with respective dynamics P and $\{P_n,\ n \in \mathbb{N}\}$ on which $\mathbb{P}(|\mathbf{X}_t - \mathbf{X}_t^n|) \to 0$ as $n \to \infty$. This is possible for almost all $\mathbf{x}$ by the standard coupling construction (see for example Jacka and Roberts, 1997, and Roberts *et al.*, 1998 for an example closer to the present context). This construction produces a probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$ on which for all $N \in \mathbb{N}$ and for all $\omega \in \Omega$, there exists an integer $R = R(\omega)$ with $X_t^n(\omega) = X_t(\omega)$ for all $n \ge R$ and for all $0 \le t \le N$. Now

$$\|P_n^T(\mathbf{x}, \cdot) - \pi_n(\cdot)\| \le \|P^T(\mathbf{x}, \cdot) - \pi(\cdot)\| + \|P_n^T(\mathbf{x}, \cdot) - P^T(\mathbf{x}, \cdot)\| + \|\pi_n(\cdot) - \pi(\cdot)\|,$$

so that by the coupling inequality, the second term on the right hand side of the above converges to 0 and the last term goes to 0 by assumption (A2). Hence,

$$\lim_{n\to\infty} \|P_n^T(\mathbf{x}, \cdot) - \pi_n(\cdot)\| < \epsilon$$

for almost all $\mathbf{x}$. Therefore, $\limsup_{n\to\infty} T_n \le T$. Conversely, we have

$$\|P^T(\mathbf{x}, \cdot) - \pi(\cdot)\| \le \|P^T(\mathbf{x}, \cdot) - P_n^T(\mathbf{x}, \cdot)\| + \|P_n^T(\mathbf{x}, \cdot) - \pi_n(\cdot)\| + \|\pi_n(\cdot) - \pi(\cdot)\|.$$

Hence using similar arguments as above,

$$\liminf_{n\to\infty} \|P^{T_n}(\mathbf{x},\cdot) - \pi(\cdot)\| \leq \epsilon$$

and the result follows. □

In the Bayesian inference setup under suitable conditions π_n has the Gaussian limit π. Hence the relevance of the above result can be summarized in the following. Under the conditions required for the application of Theorem 1 and for large enough data sets, the convergence time for the Gibbs sampler can be approximated by the corresponding convergence time for the limiting Gaussian density.

We must note the limitations of Theorem 1, however. It is extremely difficult to quantify the size of the data set needed for the Gaussian approximation to be reasonable. In practice therefore, some empirical confirmation of approximate normality is required.

5 Discussion

We have proposed a method for assessing convergence of Gibbs samplers, and for designing improved algorithms. The method can be applied to large classes of problems (though not on discrete state spaces), and appears to be effective in the examples we considered.

There are a number of important outstanding issues that we have barely touched on though. It is unclear how far from normality we can be allowed to stray before our approximations break down. Moreover, any problems of this type are likely to be exacerbated in high-dimensional examples, and it is unclear to what extent the various approximations we make are robust in high dimensions. (The largest problem we consider is only 303 dimensional, which is still relatively small in some MCMC contexts.)

We have concentrated largely on one way of implementing our methodology by obtaining the Hessian matrix of the posterior distribution at the mode. Other variations are possible. For instance it is possible to estimate the rate of convergence of the EM algorithm directly.

Acknowledgements

We thank two anonymous referees and an associate editor for their helpful comments and suggestions which improved the presentation of the paper.

REFERENCES

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory.* Chichester: John Wiley & Sons.

Breslow, N. E. and Clayton, D. G. (1993) "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9–25.

Crowder, M. J. (1978) "Beta-binomial ANOVA for proportions," *Applied Statistics*, 27, 34–37.

Dellaportas, P. and Smith, A. F. M. (1993) "Bayesian inference for generalised linear and proportional hazards models via Gibbs sampling," *Applied Statistics*, 42, 443–459.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) "Maximum likelihood from from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society*, B, 39, 1–38.

Garren, S. and Smith, R. L. (2000) "Estimating the Second Largest Eigenvalue of a Markov Transition Matrix," *Bernoulli*, 6, ?–?.

Gaver, D. P. and O'Muircheartaigh, I. G. (1987) "Robust empirical Bayes analysis of event rates," *Technometrics*, 29, 1–15.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1996) "Efficient Parametrizations for Generalised 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.

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.

Green, P. J. and Han, X-L. (1992) "Metropolis methods, Gaussian proposals, and antithetic variables," in *Stochastic models, Statistical Methods and Algorithms in Image Analysis. Lect. Notes Statist.,* 74, 142–164, Springer-Verlag, Berlin.

Hills, S. E. and Smith, A. F. M. (1993) "Diagnostic plots for improved parametrization in Bayesian inference," *Biometrika*, 80, 61–74.

Jacka, S. D. and Roberts, G. O. (1997) "On Strong Forms of Weak Convergence," *Stochastic Processes and Their Applications*, 67, 41–53.

Knuiman, M. W., and Speed, T. P. (1988) "Incorporating prior information into the analysis of contingency tables," *Biometrics*, 44, 1061–1071.

Meng, X-L. and van Dyk, D. (1997) "The EM algorithm–an old folk-song sung to a fast new tune (with discussion)," *Journal of the Royal Statistical Society*, B, 59, 511–567.

Roberts, G. O., Rosenthal, J. S. and Schwartz, P. O. (1998) "Convergence properties of perturbed Markov chains," *Journal of Applied Probability*, 35, 1–11.

Roberts, G. O. and Sahu, S. K. (1997) "Updating Schemes, Correlation Structure, Blocking and Parameterisation for the Gibbs Sampler," *Journal of the Royal Statistical Society*, B, 59, 291–317.

Roberts, G. O. and Smith, A. F. M. (1994) "Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms," *Stochastic Processes and Their Applications*, 49, 207–216.

Roberts, G. O. and Tweedie, R. L. (1999) "Bounds on Regeneration Times and Convergence Rates for Markov Chains," *Stochastic Processes and Their Applications*, 80, 211–229.

Rosenthal, J. S. (1995) "Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo," *Journal of the American Statistical Association*, 90, 558–566.

Sahu, S. K. (2000) "Discussion of the paper by Durbin and Koopman," *Journal of the Royal Statistical Society*, B, 62, 35–36.

Sahu, S. K. and Roberts, G. O. (1999) "On Convergence of the EM Algorithm and the Gibbs Sampler," *Statistics in Computing*, 9, 55—64.

Spiegelhalter, D. J., Thomas, A. and Best, N. G. (1996) "Computation on Bayesian graphical models," in *Bayesian Statistics 5*, (Eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press, pp. 407–426.

Thall, P. F. and Vail, S. C. (1990) "Some Covariance Models for Longitudinal and Count Data with Overdispersion," *Biometrics*, 46, 657–671.

Tierney, L. (1994) "Markov chains for exploring posterior distributions (with discussion)," *Annals of Statistics*, 22, 1701–1762.