ON COMPUTATION USING GIBBS SAMPLING FOR MULTILEVEL MODELS

A.E. Gelfand, B.P. Carlin, M. Trevisani

2000.6

Dipartimento di Scienze Statistiche
Università degli Studi
Via S. Francesco, 33
35121 Padova

Maggio 2000

On Computation Using Gibbs Sampling for Multilevel Models

by Alan E. Gelfand, Bradley P. Carlin, and Matilde Trevisani[1]

Abstract

Multilevel models incorporating random effects at the various levels are enjoying increased popularity. An implicit problem with such models is identifiability. From a Bayesian perspective, formal identifiability is not an issue. Rather, when implementing iterative simulation-based model fitting, a poorly behaved Gibbs sampler frequently arises. The objective of this paper is to shed light upon two computational issues in this regard. The first concerns autocorrelation in the sequence of iterates of the Markov chain. We clarify when, for *estimable* functions, autocorrelation will drop off to zero rapidly, enabling high effective sample size. The second concerns exact sampling, i.e., when, at an arbitrary iteration, the simulated value of a variable is in fact an observation from the posterior distribution of the variable. Again, for estimable functions, we clarify when the chain will produce at each iteration a sample drawn essentially from the true posterior of the function. We provide both analytical and computational support for our conclusions, including exemplification for three multilevel models having normal, Poisson, and binary responses, respectively.

Key words: Autocorrelation; estimable function; exact sampling; identifiability.

[1] Alan E. Gelfand is Professor in the Department of Statistics at the University of Connecticut, Storrs, CT, 06269. Bradley P. Carlin is Professor in the Division of Biostatistics, School of Public Health at the University of Minnesota, Minneapolis, MN, 55455. Matilde Trevisani is postdoctoral researcher, Department of Statistical Sciences at the University of Padua, Padua, Italy, 35121. The work of the first author was supported in part by NSF grant DMS 96-25383, while that of the second author was supported in part by National Institute of Environmental Health Sciences (NIEHS) Grant 1-R01-ES07750 and National Institutes of Allergy and Infectious Diseases (NIAID) Grant R01-AI41966.

1 Introduction

Multilevel models incorporating random effects at the various levels are enjoying increased popularity among practitioners, particularly as fast, inexpensive computing makes their fitting more widely accessible. The book by Goldstein (1995) has detailed the classical viewpoint including implementation.

From the Bayesian perspective, a Gibbs sampler (Gelfand and Smith, 1990) for such models is conceptually straightforward to implement since the required full conditional distributions are either standard (arising from conjugacy) or log concave (e.g., an exponential family first stage specification with canonical link). See, for example, the book of Gilks, Richardson and Spiegelhalter (1995). The BUGS software (Spiegelhalter et al., 1995) is a reliable package which is frequently used to implement the simulation-based model fitting. Comparison between Bayesian and likelihood methods for fitting multilevel models is taken up in the recent work of Browne and Draper (1999).

An implicit problem which arises under multilevel random effects models is identifiability. Upon an appropriate linear transformation, the likelihood only involves a subset of the parameters. The remainder are not *identified* in the classical sense. Fortunately, parametric functions of interest (e.g., individual means and contrasts at each level) are usually estimable (Searle, 1971) and the likelihood does identify such functions, enabling classical point and interval estimation.

From the Bayesian perspective, hierarchical models are routinely overparametrized. However, under proper priors there is no identifiability problem (Lindley, 1971); the posterior for every model unknown is proper. Dawid (1979) clarifies the Bayesian notion of unidentifiability and recent work of Poirier (1998) and Gelfand and Sahu (1999) provides further elaboration. An equivalence with classical nonidentifiabilty then follows.

Perhaps most interestingly, what emerges from all of this discussion is an informal notion of weak identifiability. For certain unknowns the posterior differs little from the prior; the data provide little Bayesian learning about the unknown. A practical consequence when implementing iterative simulation-based model fitting is a poorly behaved Gibbs sampler. When rather vague priors are used, trajectories of the Markov chain for weakly identified parameters will tend to exhibit drift to very extreme values, since there is nothing in the structure to center them. Convergence assessment is difficult; unstable computation and inaccurate inference ensue. On the other hand, very precise priors are generally unattractive, since then Bayesian learning is necessarily limited. Taken to the extreme, a degenerate prior would be specified which would be analogous to imposing restrictions or constraints, as is customarily done in the classical setting.

The objective of this paper is to shed light upon two computational issues in this regard. Possibly unexpected implications for using Gibbs sampling to fit multilevel models result. That is, the behavior of the Gibbs sampler with regard to certain parameters may improve as prior specifications are made increasingly vague. The first issue concerns autocorrelation in the sequence of iterates of the Markov chain. In particular, autocorrelations which drop off to zero rapidly enable, following a diagnosis of convergence, high effective sample size thus avoid thinning of this output. We show, using both analytical and computational evidence, that if the variance components associated with the random effects at the highest level are made increasingly larger, then post-convergence, the lag-one autocorrelation for any estimable parameter (or any function of an estimable parameter) tends to 0. Moreover, if all of the variance components in the model are made increasingly larger, the overall posterior tends to impropriety. Following Gelfand and Sahu (1999), in this case there is a unique proper embedded posterior associated with the estimable parameters. Remarkably, the post-convergence simulation behavior from this posterior is improving with regard to autocorrelation.

The second issue concerns exact sampling. In the Markov chain Monte Carlo context, exact sampling for a particular parameter means that, regardless of iteration, the simulated value at that iteration is, in fact, an observation from the posterior distribution of the parameter. Again we show, using both analytical and computational evidence, that if all of the variance components in the model are made increasingly larger, for any estimable parameter (or function of the parameter) we tend to exact sampling. In other words, while the overall posterior tends to impropriety so that sampling its full conditional distributions cannot lead to meaningful convergence for the full parameter vector, we tend to exact sampling of the unique proper embedded posterior. This result provides clarification and extension of a result in Section 5 of Gelfand and Sahu (1999).

The format of the paper is as follows. In Section 2 we present an elementary example which is useful in illustrating all of our ideas. Section 3 presents the formal technical work in the form of two theorems. Section 4 offers empirical clarification in the Gaussian case with unknown variance components but familiar prior specifications for these components. Section 5 provides empirical support for the non-Gaussian case where analytic work is infeasible; illustration is given using a two-level Poisson spatial model. Finally, Section 6 analyzes a three-level binary response model for data concerning plant health and the presence of certain species of fungi at fine root apexes.

2 An Elementary Example

A simple illustrative example may be helpful to appreciate the general results of the next section. Suppose $Y \sim N(\theta+\phi, 1)$ with $\theta \sim N(0, \sigma_\theta^2)$, $\phi \sim N(0, \sigma_\phi^2)$ and let $\eta = \theta+\phi$. By routine calculation,

$$f(\theta \mid \phi, Y) = N(\epsilon_\theta(Y-\phi), \epsilon_\theta) \text{ and } f(\phi \mid \theta, Y) = N(\epsilon_\phi(Y-\theta), \epsilon_\phi)$$

where $\epsilon_\theta = \sigma_\theta^2/(\sigma_\theta^2+1)$ and $\epsilon_\phi = \sigma_\phi^2/(\sigma_\phi^2+1)$. Also

$$f\left(\begin{pmatrix}\theta \\ \\ \phi\end{pmatrix} \Bigg| Y\right) = N\left(\begin{pmatrix}\sigma_\theta^2 Y/a \\ \\ \sigma_\phi^2 Y/a\end{pmatrix}, \begin{pmatrix}\sigma_\theta^2(1+\sigma_\phi^2)/a & -\sigma_\theta^2\sigma_\phi/a \\ \\ -\sigma_\theta^2\sigma_\phi^2/a & \sigma_\phi(1+\sigma_\theta^2)/a\end{pmatrix}\right),$$

and $f(\eta \mid Y) = N\left(\frac{\sigma_\theta^2+\sigma_\phi^2}{a}Y,\ \frac{\sigma_\theta^2+\sigma_\phi^2}{a}\right)$ where $a = \sigma_\theta^2 + \sigma_\phi^2 + 1$.

Suppose we implement a Gibbs sampler updating θ and then ϕ, i.e.,

$$f(\theta^{(t+1)}, \phi^{(t+1)} \mid \theta^{(t)}, \phi^{(t)}, Y) = f(\phi^{(t+1)} \mid \theta^{(t+1)}, Y) \cdot f(\theta^{(t+1)} \mid \phi^{(t)}, Y). \tag{1}$$

Then directly, $cov(\theta^{(t+1)}, \theta^{(t)}|Y) = -\epsilon_\theta cov(\phi^{(t)}, \theta^{(t)}|Y)$ from which, at convergence, we have that $corr(\theta^{(t+1)}, \theta^{(t)}|Y) = \epsilon_\theta\epsilon_\phi$. Similarly, $cov(\phi^{(t+1)}, \phi^{(t)}|Y) = -\epsilon_\phi cov(\theta^{(t+1)}, \phi^{(t)}|Y) = \epsilon_\phi\epsilon_\theta var(\phi^{(t)}|Y)$ from which, at convergence, $corr(\phi^{(t+1)}, \phi^{(t)}|Y) = \epsilon_\theta\epsilon_\phi$. Hence, severe autocorrelation occurs when both σ_θ^2 and $\sigma_\phi^2 \to \infty$.

On the other hand, $cov(\eta^{(t+1)}, \eta^{(t)}|y) = cov(\theta^{(t+1)} + \phi^{(t+1)}, \theta^{(t)} + \phi^{(t)}|Y) = (1 - \epsilon_\phi) \times cov(\theta^{(t+1)}, \theta^{(t)} + \phi^{(t)}|Y) = -\epsilon_\theta(1-\epsilon_\phi)\{cov(\phi^{(t)}, \theta^{(t)}|Y) + var(\phi^{(t)}|Y)\} = -\epsilon_\theta\epsilon_\phi/a$ at convergence. Hence, if either σ_θ^2 or σ_ϕ^2 grows large, a grows large and the posterior association between $\eta^{(t+1)}$ and $\eta^{(t)}$ tends to 0. This illustrates the primary result in Subsection 3.1.

Next, from (1), $f(\theta^{(t+1)}, \phi^{(t+1)} \mid \theta^{(t)}, \phi^{(t)}, Y)$ is bivariate normal, so $f(\eta^{(t+1)} \mid \theta^{(t)}, \phi^{(t)}, Y)$ is normal. In fact, $\eta^{(t+1)} \mid \theta^{(t)}, \phi^{(t)}, Y \sim N(\epsilon_\phi Y + (1-\epsilon_\phi)\epsilon_\theta(Y - \phi^{(t)}), (1-\epsilon_\phi)^2\epsilon_\theta + \epsilon_\phi)$. As $\sigma_\theta^2 \to \infty$ and $\sigma_\phi^2 \to \infty$, this distribution tends to $N(Y, 1)$. But note that, in this case, $f(\eta \mid Y) = N(Y, 1)$. In the limit, at each iteration of the Gibbs sampler, we have exact sampling from the posterior of η. This illustrates the primary result in Subsection 3.2.

Lastly, it is apparent that, if we reverse the order of updating, drawing ϕ first then θ, all of the foregoing conclusions still hold.

3 Technical Results

Our general analytic results presume a Gaussian specification for the data. In addition, all variance components are assumed known. Evidently, priors can be placed on the variance components to, for instance, encourage one or more of them to be large. A non-Gaussian first stage model precludes analytic investigation since the resultant full conditional distributions associated with the Gibbs sampler are not standard. However, we argue below that, when the likelihood is approximately normal, our analytic results still apply. As a result, in Section 5 we present a numerical illustration using a Poisson first stage, while in Section 6 we consider a three-stage random effects model with a binary first stage.

We introduce some notation. Consider the linear model

$$\mathbf{Y} = (X_0 \quad X_0\Delta_1 \quad \cdots \quad X_0\Delta_b)\begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_b \end{pmatrix} + \epsilon \tag{2}$$

where Y is $n \times 1$, X_0 is $n \times r_0$ with full column rank, Δ_i is $r_0 \times r_i$, $\boldsymbol{\beta}_i$ is a parameter vector of length r_i, and $\epsilon \sim N(\mathbf{0}, I_n)$. For analytic investigation we can set the error variances to 1 without loss of generality, since we will be looking at the variance components associated with the $\boldsymbol{\beta}_i$ tending to ∞. In practice, we will require priors making other components large relative to the error variances.

Letting X denote the design matrix in (2) and $\boldsymbol{\beta}$ the concatenated vector of $\boldsymbol{\beta}_i$'s in (1), $E(\mathbf{Y}) \equiv \eta = X\boldsymbol{\beta} = X_0(\boldsymbol{\beta}_0 + \sum_{i=1}^{b} \Delta_i\boldsymbol{\beta}_i)$. Also

$$X^TX = \begin{pmatrix} X_0^TX_0 & X_0^TX_0\Delta_1 & \cdots & X_0^TX_0\Delta_b \\ \Delta_1^TX_0^TX_0 & \Delta_1^TX_0^TX_0\Delta_1 & \cdots & \Delta_1^TX_0^TX_0\Delta_b \\ \vdots & & \ddots & \vdots \\ \Delta_b^TX_0^TX_0 & \Delta_b^TX_0^TX_0\Delta_1 & \cdots & \Delta_b^TX_0^TX_0\Delta_b \end{pmatrix}. \tag{3}$$

It is immediate that the design matrix X arises in any ANOVA specification which is fully nested, hence any multilevel model with no quantitative regressors. In fact, it includes general multilevel models with quantitative covariates and general Laird-Ware (1982) models, as we clarify below following Remark 2. It also includes any main effects model which incorporates an interaction involving all of the main effects. Also, the Δ_i need not be distinct as long as when $\Delta_i = \Delta_j$ the prior covariance matrix for $\boldsymbol{\beta}_i$ is distinct from that for $\boldsymbol{\beta}_j$. In this way we can accommodate, e.g., both spatial and heterogeneity effects (see e.g. Clayton and Bernardinelli, 1992; Bernardinelli et al., 1995; c.f. Section 5 below) at a given level of the model. In any event, we shall see that X_0 is always the portion of the design matrix associated with the parameters at the highest level.

The prior specification is

$$f(\beta \mid \sigma^2) \propto \exp\left(-\frac{1}{2}\sum_{i=0}^{b}\boldsymbol{\beta}_i^T V_{\sigma_i^2}\boldsymbol{\beta}_i\right) = \exp\left(-\frac{1}{2}\boldsymbol{\beta}^T V_{\boldsymbol{\sigma}^2}\boldsymbol{\beta}\right) \tag{4}$$

where V_{σ^2} is block-diagonal with i^{th} block $V_{\sigma_i^2}$. In (4), the $V_{\sigma_i^2}$ need not be full rank. The prior for β_i need not be proper. In fact, we will assume that $V_{\sigma_i^2} = V_i/\sigma_i^2$ so that σ_i^2 can be thought of as a variance component and $V_{\sigma_i^2} \to 0$ as $\sigma_i^2 \to \infty$. The priors are "zero-centered" for simplicity in the ensuing calculations, and as is typically the case in practice.

3.1 An autocorrelation result for estimable parameters

Following Lindley and Smith (1972),

$$\beta \mid \mathbf{Y} \sim N\left((X^TX + V_{\boldsymbol{\sigma}^2})^{-1}X^T\mathbf{Y}\ ,\ (X^TX + V_{\boldsymbol{\sigma}^2})^{-1}\right), \tag{5}$$

provided the inverse exists. For instance, if $b = 1$,

$$X^T X + V_{\boldsymbol{\sigma}^2} = \begin{pmatrix} X_0^T X_0 + V_{\sigma_0^2} & X_0^T X_0 \Delta_1 \\ \Delta_1^T X_0^T X_0 & \Delta_1^T X_0^T X_0 \Delta_1 + V_{\sigma_i^2} \end{pmatrix},$$

whence $\left|X^T X + V_{\sigma^2}\right| = \left|X_0^T X_0 + V_{\sigma_0^2}\right| \left|\Delta_1^T X_0^T X_0 \Delta_1 + V_{\sigma_1^2} - \Delta_1^T X_0^T X_0 (X_0^T X_0 + V_{\sigma_0^2})^{-1} X_0^T X_0 \Delta_1\right|$. If $\sigma_0^2 \to \infty$ (as we will want below), this simplifies to $\left|X_0^T X_0\right| \left|V_{\sigma_1^2}\right|$ so $\boldsymbol{\beta} \mid \mathbf{Y}$ is proper only if V_1 is nonsingular.

We require some further notation. Let $\Omega_{\sigma_i^2} = \Delta_i^T X_0^T X_0 \Delta_i + V_{\sigma_i^2}$, $i = 1, \ldots, b$ with $\Omega_{\sigma_0^2} = X_0^T X_0 + V_{\sigma_0^2}$. Also let $\mathbf{Y}_{(i)} = \mathbf{Y} - X_0(\boldsymbol{\beta}_0 + \sum_{j \neq i} \Delta_j \boldsymbol{\beta}_j) = \mathbf{Y} - X\boldsymbol{\beta} + X_0 \Delta_i \boldsymbol{\beta}_i$, $i = 1, \ldots, b$ with $\mathbf{Y}_{(0)} = \mathbf{Y} - X_0 \sum_{i=1}^b \Delta_i \beta_i = \mathbf{Y} - X\boldsymbol{\beta} + X_0 \boldsymbol{\beta}_0$. Then the full conditional distributions for the $\boldsymbol{\beta}$'s are:

$$\begin{aligned} &\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_j,\ j \neq i,\ \mathbf{Y} \sim N(\Omega_{\sigma_i^2}^{-1} \Delta_i^T X_0^T \mathbf{Y}_{(i)},\ \Omega_{\sigma_i^2}^{-1}),\ \ i \neq 0, \text{ and} \\ &\boldsymbol{\beta}_0 \mid \boldsymbol{\beta}_j,\ j \neq 0,\ \mathbf{Y} \sim N(\Omega_{\sigma_0^2}^{-1} X_0^T \mathbf{Y}_{(0)},\ \Omega_{\sigma_0^2}^{-1})\,. \end{aligned} \tag{6}$$

Using standard matrix identities, e.g., Rao (1973, p.29) we can write $(X^T X + V_{\boldsymbol{\sigma}^2})^{-1}$ in (5) as

$$\begin{pmatrix} W_0^{-1} & - \quad W_0^{-1} A_{\boldsymbol{\sigma}^2}^T \\ - \quad A_{\boldsymbol{\sigma}^2} W_0^{-1} & \Omega_{\boldsymbol{\sigma}^2}^{-1} + A_{\boldsymbol{\sigma}^2} W_0^{-1} A_{\boldsymbol{\sigma}^2}^T \end{pmatrix} \tag{7}$$

where $W_0 = \Omega_{\sigma_0^2} - X_0^T X_0 \Delta_{(0)} A_{\boldsymbol{\sigma}^2}$, $A_0 = \Omega_{\sigma_0^2}^{-1} X_0^T X_0 \Delta_{(0)}$, and $A_{\boldsymbol{\sigma}^2} = \Omega_{\boldsymbol{\sigma}^2}^{-1} \Delta_{(0)}^T X_0^T X_0$. Here, $\Delta_{(0)} = (\Delta_1, \Delta_2, \ldots, \Delta_b)$ and $\Omega_{\boldsymbol{\sigma}^2} = \Delta_{(0)}^T X_0^T X_0 \Delta_{(0)} + V_{(0)}$, with $V_{(0)}$ block diagonal having blocks $V_{\sigma_i^2}$, $i = 1, \ldots, b$. Note that, as $\sigma_0^2 \to \infty$, $A_0 \to \Delta_{(0)}$.

Now suppose we are at convergence so that

$$f(\boldsymbol{\beta}^{(t)} \mid \mathbf{Y}) = N\left((X^T X + V_{\boldsymbol{\sigma}^2})^{-1} X^T \mathbf{Y},\ (X^T X + V_{\boldsymbol{\sigma}^2})^{-1}\right).$$

We implement a Gibbs sampler, updating the $\boldsymbol{\beta}$'s in any order, but updating $\boldsymbol{\beta}_0$ last. (By relabeling the $\boldsymbol{\beta}_i$'s we can, without loss of generality assume the order is $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_b$.) Then

$$cov\left(\boldsymbol{\beta}_0^{(t+1)} + \sum_{i=1}^{b} \Delta_i \boldsymbol{\beta}_i^{(t+1)},\ \boldsymbol{\beta}_0^{(t)} + \sum_{i=1}^{b} \Delta_i \boldsymbol{\beta}_i^{(t)} | \mathbf{Y}\right) = cov\left(\boldsymbol{\beta}_0^{(t+1)} + \Delta_{(0)}\boldsymbol{\beta}_{(0)}^{(t+1)},\ \boldsymbol{\beta}_0^{(t)} + \Delta_{(0)}\boldsymbol{\beta}_{(0)}^{(t)} | \mathbf{Y}\right) \tag{8}$$

where $\boldsymbol{\beta}_{(0)}^T = (\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_r^T)$. But then, using (6), (8) becomes

$$cov\left(-A_0\boldsymbol{\beta}_{(0)}^{(t+1)} + \Delta_{(0)}\boldsymbol{\beta}_{(0)}^{(t+1)},\ \boldsymbol{\beta}_0^{(t)} + \Delta_{(0)}\boldsymbol{\beta}_{(0)}^{(t)} \mid \mathbf{Y}\right) = (\Delta_{(0)} - A_0) cov\left(\boldsymbol{\beta}_{(0)}^{(t+1)},\ \boldsymbol{\beta}_0^{(t)} + \Delta_{(0)}\boldsymbol{\beta}_{(0)}^{(t)} \mid \mathbf{Y}\right). \tag{9}$$

Hence, as $\sigma_0^2 \to \infty$, equation (9) $\to 0$. In fact, more detailed calculation shows that, at whatever stage we update $\boldsymbol{\beta}_0$, the covariance calculation in (8) introduces a $(\Delta_{(0)} - A_0)$ term so that again, as $\sigma_0^2 \to \infty$, equation (8) approaches 0. Hence we have the following result.

Theorem 1. For the model in (2) under the prior in (4), if we implement a Gibbs sampler which updates the blocks, $\boldsymbol{\beta}_i$, in any order, then provided (5) exists, after convergence $cov(\boldsymbol{\eta}^{(t+1)}, \boldsymbol{\eta}^{(t)} \mid \mathbf{Y}) \to 0$ as $\sigma_0^2 \to \infty$.

In other words, once the Gibbs sampler has converged, if σ_0^2 is large, successive iterates of $\boldsymbol{\eta}$, hence of any estimable function, hence of any function of an estimable function, will be approximately uncorrelated.

Remark 1. Note the distinct role played by σ_0^2 above. If we define $A_i = \Omega_{\sigma_i^2}^{-1} \Delta_i^T X_0^T X_0 \Delta_{(i)}$ where $\Delta_{(i)} = (I_{r_0 \times r_0}, \Delta_1, ..., \Delta_{i-1}, \Delta_{i+1}, ..., \Delta_b)$, as $\sigma_i^2 \to 0$, $A_i \to (\Delta_i^T X_0^T X_0 \Delta_i)^{-1} \Delta_i^T X_0^T X_0 \Delta_{(i)} \neq \Delta_{(i)}$. Hence, in (8), if, for instance, we update $\boldsymbol{\beta}_i$ last and factor out $\Delta_{(i)} - A_i$ analogously to (9), we do *not* obtain covariance tending to 0 as $\sigma_i^2 \to \infty$. However, in the special case where, for some $i \neq 0$, $\Delta_i = I$ then as $\sigma_i^2 \to \infty$, equation (8) *does* tend to 0.

Remark 2. Suppose, for example, we update $\boldsymbol{\beta}_0$ first. Then $cov(\boldsymbol{\beta}_0^{(t+1)}, \boldsymbol{\beta}_0^{(t)} | \mathbf{Y}) = A_0 A_{\boldsymbol{\sigma}^2} W_0^{-1}$.

From (7), $cov(\boldsymbol{\beta}_0^{(t)}, \boldsymbol{\beta}_0^{(t)} \mid \mathbf{Y}) = cov(\boldsymbol{\beta}_0^{(t+1)}, \boldsymbol{\beta}_0^{(t+1)} \mid \mathbf{Y}) = W_0^{-1}$. To study the correlation between, say, $\beta_{0\ell}^{(t+1)}$ and $\beta_{0\ell}^{(t)}$ we need to investigate $(W_0^{-1})_{\ell\ell}$ and $(A_0 A_{\boldsymbol{\sigma}^2} W_0^{-1})_{\ell\ell}$. This is more easily done using an alternative form of (7), again obtained from standard identities,

$$\begin{pmatrix} \Omega_{\sigma_0^2}^{-1} + A_0 V_{(0)}^{-1} A_0^T & -A_0 V_{(0)}^{-1} \\ -V_{(0)}^{-1} A_0^T & V_{(0)}^{-1} \end{pmatrix},$$

provided V_i^{-1} exists for $i = 1, 2, \ldots, b$.

Hence, as $\sigma_0^2 \to \infty$, $W_0^{-1} = \Omega_{\sigma_0^2}^{-1} + A_0 V_{(0)}^{-1} A_0^T \to (X_0^T X_0)^{-1} + \Delta_{(0)} V_{(0)}^{-1} \Delta_{(0)}^T = (X_0^T X_0)^{-1} + \sum_i \Delta_i V_{\sigma_i^2}^{-1} \Delta_i^T$. Also, $A_0 A_{\boldsymbol{\sigma}^2} W_0^{-1} = A_0 V_{(0)}^{-1} A_0^T \to \Delta_{(0)} V_{(0)}^{-1} \Delta_{(0)}^T = \sum_i \Delta_i V_{\sigma_i^2}^{-1} \Delta_i^T$. We see that if, in addition, any $\sigma_i^2 \to \infty$ then $corr(\beta_{0\ell}^{(t+1)}, \beta_{0\ell}^{(t)} \mid \mathbf{Y}) \to 1$. More detailed calculation shows that this result holds regardless of updating order. We can also show that if $\sigma_0^2 \to \infty$ and $\sigma_i^2 - \infty$, $cov(\beta_{i\ell}^{(t+1)}, \beta_{i\ell}^{(t)} \mid \mathbf{Y}) \to 1$.

The model in (2) is more flexible than might first appear. For instance, consider the general two stage multilevel linear model

$$\begin{aligned} \mathbf{Y}_i &= X_i \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \\ \boldsymbol{\eta}_i &= Z_i \boldsymbol{\gamma} + \boldsymbol{v}_i, \; i = 1, \ldots, I \end{aligned} \qquad \mathbf{(10)}$$

where $\mathbf{Y}_i$ is $n_i \times 1$, X_i is $n_i \times m_0$ with full column rank, $\boldsymbol{\eta}_i$ is $m_0 \times 1$ and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma_e^2 I_{m_i \times n_i})$. Also Z_i is $m_0 \times m_1$, γ is $m_1 \times 1$, $\boldsymbol{v}_i \sim N(0, \sigma_v^2 H_v)$ and finally $\boldsymbol{\gamma} \sim N(0, \sigma_\gamma^2 H_\gamma)$.

Next, let $\mathbf{Y}^T = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_I^T)$, let X_0, $\Sigma n_i \times I m_0$, be block diagonal with i^{th} block X_i and let $\boldsymbol{\beta}_0^T = (\boldsymbol{v}_1^T, \ldots, \boldsymbol{v}_I^T)$. Finally, let Δ_1 be $I m_0 \times m_1$, such that $\Delta_1^T = (Z_1^T, \ldots, Z_I^T)$ and let $\boldsymbol{\beta}_1 = \boldsymbol{\gamma}$. Then

(10) can be written as

$$\mathbf{Y} = (X_0 \;\; X_0\Delta_1)\begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{\epsilon}\,.$$

Here $V_{\sigma_0^2} = \frac{1}{\sigma_v^2}V$ where V, $Im_0 \times Im_0$, is block diagonal with blocks H_v^{-1} and $V_{\sigma_1^2} = \frac{1}{\sigma_\gamma^2}H_\gamma^{-1}$. Hence, Theorem 1 applies when σ_v^2 is large relative to σ_e^2.

The extension to a general three stage model is apparent but we give brief details to provide structural clarification. Now, let

$$\begin{aligned} \mathbf{Y}_{ij} &= X_{ij}\boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{ij}, \\ \boldsymbol{\eta}_{ij} &= Z_{ij}\boldsymbol{\gamma}_i + \boldsymbol{v}_{ij}, \text{ and} \\ \boldsymbol{\gamma}_i &= W_i\boldsymbol{\delta} + \boldsymbol{\mu}_i. \end{aligned} \tag{11}$$

The *extended form* for (11) is $\mathbf{Y}_{ij} = X_{ij}Z_{ij}W_i\boldsymbol{\delta} + X_{ij}Z_{ij}\boldsymbol{\mu}_i + X_{ij}\boldsymbol{v}_{ij} + \boldsymbol{\epsilon}_{ij}$. In (11), $\mathbf{Y}_{ij}$ is $n_{ij} \times 1$, $i = 1, ..., I$, $j = 1, ..., J_i$, X_{ij} is $n_{ij} \times m_0$, $\boldsymbol{\eta}_{ij}$ is $m_0 \times 1$ and $\boldsymbol{\epsilon}_{ij} \sim N(\mathbf{0}, \sigma_e^2 I_{n_{ij}\times n_{ij}})$. Now, Z_{ij} is $m_0 \times m_1$, $\boldsymbol{\gamma}_i$ is $m_1 \times 1$ and $\boldsymbol{v}_{ij} \sim N(\mathbf{0}, \sigma_v^2 H_v)$. Lastly, W_i is $m_1 \times m_2$, $\boldsymbol{\mu}_i$, $m_1 \times 1$, $\sim N(\mathbf{0}, \sigma_\mu^2 H_\mu)$ and $\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_\delta^2 H_\delta)$.

Again, concatenating the $\mathbf{Y}_{ij}$'s into a column vector $\mathbf{Y}$, let X_0 be block diagonal with blocks X_{ij} and concatenate the $\boldsymbol{v}_{ij}$ into a column vector $\boldsymbol{\beta}_0$. Next let

$$\Delta_1^T = \begin{pmatrix} Z_{11}^T & \cdots & Z_{1J_1}^T & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & Z_{21}^T & \cdots & Z_{2J_2}^T & \cdots & 0 & \cdots & 0 \\ & \vdots & & & \vdots & & \ddots & & \vdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & Z_{I1}^T & \cdots & Z_{IJ_I}^T \end{pmatrix}.$$

Let $\boldsymbol{\beta}_1$ concatenate the $\boldsymbol{\mu}_i$ into a column vector, let $\Psi_1^T = (W_1^T, ..., W_I^T)$, and let $\boldsymbol{\beta}_2 = \boldsymbol{\delta}$. Then

(11) can be written as

$$\mathbf{Y} = (X_0 \;\; X_0\Delta_1 \;\; X_0\Delta_1\Psi_1) \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon} \, .$$

Again, Theorem 1 applies when σ_v^2 is large relative to σ_e^2.

We remark that, for models such as (10) and (11), Gelfand, Carlin and Sahu (1995, 1996) argued that a "hierarchically centered" parametrization would typically provide a better behaved Gibbs sampler. For instance, in (10), $(\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_I, \boldsymbol{\gamma})$ would be preferred to $(\boldsymbol{v}_1, ..., \boldsymbol{v}_I, \boldsymbol{\gamma})$, i.e., $(\boldsymbol{\beta}_0 + \Delta_1\boldsymbol{\beta}_1, \boldsymbol{\beta}_1)$ to $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$. Similarly in (11), $(\boldsymbol{\eta}_{11}, ..., \boldsymbol{\eta}_{IJ}, \boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_I, \boldsymbol{\delta})$ would be preferred to $(\boldsymbol{v}_{11}, ..., \boldsymbol{v}_{IJ}, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_I, \boldsymbol{\delta})$, i.e., $(\boldsymbol{\beta}_0 + \Delta_1\boldsymbol{\beta}_1 + \Delta_1\Psi_1\boldsymbol{\beta}_2,\ \boldsymbol{\beta}_1 + \Psi_1\boldsymbol{\beta}_2,\ \boldsymbol{\beta}_2)$ to $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. Theorem 1 only applies to the vector of highest order hierarchically centered parameters.

To conclude this subsection we note that (2) will include certain models of the form

$$\mathbf{Y}_i = X_i\boldsymbol{\alpha} + Z_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \; = 1. \ldots, I \, , \tag{12}$$

the so-called Laird-Ware (1982) models. We provide two illustrations.

In the first case, suppose that X_i, $n_i \times p$, is nested with Z_i, $n_i \times (p+q)$, i.e., $Z_i = (X_i \;\; U_i)$. Then if X_0 is block diagonal with blocks Z_i, if $\boldsymbol{\beta}_0$ collects the $\boldsymbol{\beta}_i$ into a vector, if Δ_1 is $I(p+q) \times p$ of the form

$$\Delta_1^T = (I_{p\times p} \; 0_{p\times q} \; I_{p\times p} \; 0_{p\times q} \cdots I_{p\times p} \; 0_{p\times q}),$$

and if $\boldsymbol{\beta}_1 = \boldsymbol{\alpha}$ then (12) becomes $\mathbf{Y} = (X_0 \;\; X_0\Delta_1)\binom{\boldsymbol{\beta}_0}{\boldsymbol{\beta}_1} + \boldsymbol{\epsilon}$.

Alternatively, suppose $X_i = \mathbf{1}_{n_i \times 1} \cdot \mathbf{X}_i^T$ where $\mathbf{X}_i$ is $p \times 1$ and suppose the first column of Z_i consists of 1's, i.e., Z_i is $n_i \times (1+q)$ such that $Z_i = (\mathbf{1} \;\; \widetilde{Z}_i)$. Then again let X_0 be block diagonal with blocks Z_i and let $\boldsymbol{\beta}_0$ collect the $\boldsymbol{\beta}_i$ into a vector. Now, if Δ_1 is $I(1+q) \times p$ of the form

$$\Delta_1^T = (\mathbf{X}_1\ 0_{pxq}\ \mathbf{X}_2\ 0_{pxq} \cdots \mathbf{X}_I\ 0_{pxq})$$

and if $\boldsymbol{\beta}_1 = \boldsymbol{\alpha}$ then again (12) becomes $\mathbf{Y} = (X_0\ \ X_0\Delta_1)\binom{\boldsymbol{\beta}_0}{\boldsymbol{\beta}_1} + \epsilon$.

3.2 An exact sampling result for estimable parameters

Returning to the general setting of equations (2) and (4), we extend an exact sampling result which appears in Gelfand and Sahu (1999, Section 5). Recalling (5), let $Q_{\boldsymbol{\sigma}^2} \equiv X^TX + V_{\boldsymbol{\sigma}^2}$. Also, let $X^TX = L - U$ where L is the lower triangular part of X^TX including all of the diagonal elements and U is obtained by subtraction. Then $Q_{\boldsymbol{\sigma}^2} = L_{\boldsymbol{\sigma}^2} - U$ where $L_{\boldsymbol{\sigma}^2} = L + V_{\boldsymbol{\sigma}^2}$. Updating in the order $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_b)$, Roberts and Sahu (1997) show that the Gibbs sampler transition kernel is given by

$$\boldsymbol{\beta}^{(t+1)} \mid \boldsymbol{\beta}^{(t)},\ \mathbf{Y} \sim N\left(B_{\boldsymbol{\sigma}^2}\boldsymbol{\beta}^{(t)} + \mathbf{b}_{\boldsymbol{\sigma}^2},\ Q_{\boldsymbol{\sigma}^2}^{-1} - B_{\boldsymbol{\sigma}^2}Q_{\boldsymbol{\sigma}^2}^{-1}B_{\boldsymbol{\sigma}^2}^T\right) \tag{13}$$

where $B_{\boldsymbol{\sigma}^2} = L_{\boldsymbol{\sigma}^2}^{-1}U$ and $\mathbf{b}_{\boldsymbol{\sigma}^2} = (I - B_{\boldsymbol{\sigma}^2})Q_{\boldsymbol{\sigma}^2}^{-1}X^T\mathbf{Y}$. They also show that the rate of convergence of the Gibbs sampler is given by the maximum modulus eigenvalue of $B_{\boldsymbol{\sigma}^2}$.

If all $\sigma_i^2 \to \infty$, the posterior distribution of $\boldsymbol{\beta}$ approaches an improper distribution. Since the full conditional distributions are proper, by direct calculation, the above transition density still remains valid in the limit if we replace $Q_{\boldsymbol{\sigma}^2}^{-1}$ by a generalized inverse of X^TX. However, following Gelfand and Sahu, $\eta = X\boldsymbol{\beta} = X_0(\boldsymbol{\beta}_0 + \sum_i \Delta_i\boldsymbol{\beta}_i)$ has a unique proper posterior distribution even as $\min_i \sigma_i^2 \to \infty$. Theorem 2 describes what happens to the Gibbs sampler asymptotically.

Theorem 2. Suppose that a Gibbs sampler with the target density $f(\boldsymbol{\beta} \mid Y)$ in (5) is run with a customary sequential updating scheme. Suppose further that L as defined above is such that L^{-1}

is a generalized inverse of $Q = X^T X$, i.e.,

$$QL^{-1}Q = Q. \tag{14}$$

Then the Gibbs sampler on the full parameter vector $\boldsymbol{\beta}$ becomes divergent as $\min_i \sigma_i^2 \to \infty$. In this limiting case, the iterates $\boldsymbol{\eta}^{(t)}$ are an exact sample from the unique density $f(\boldsymbol{\eta} \mid Y)$.

Note that, because X is not of full column rank, in the limit (5) becomes improper so the first conclusion follows. Also note that the second conclusion implies that, in the limiting case, the Gibbs sampler produces identically distributed draws from the posterior for $\boldsymbol{\eta}$. The improper prior specification for $\boldsymbol{\beta}$ results in a Gibbs sampler which yields exact samples for the proper posterior of any estimable function and hence, any function of an estimable function.

We now prove the second conclusion. Straightforwardly, in the limit, the unique proper posterior for $\boldsymbol{\eta}$ is

$$f(\boldsymbol{\eta} \mid Y) = N\left(XQ^{-}X^{T}Y,\ XQ^{-}X^{T}\right) \tag{15}$$

for an arbitrary generalized inverse Q^{-}.

Next, note that L^{-1} always exists due to the propriety of the full conditional distributions. Let $B = L^{-1}U$. It is apparent that B is idempotent if and only if (14) holds. In fact, (14) holds also if and only if $XB = 0$. Since $B_{\boldsymbol{\sigma}^2} \to B$ as $\min_i \sigma^2 \to \infty$, $XB_{\boldsymbol{\sigma}^2} \to 0$. From (13), for any $\boldsymbol{\sigma}^2$ we have

$$\boldsymbol{\eta}^{(t+1)} \mid \boldsymbol{\beta}^{(t)}, \mathbf{Y} \sim N\left(XB_{\boldsymbol{\sigma}^2}\boldsymbol{\beta}^{(t)} + X\mathbf{b}_{\boldsymbol{\sigma}^2},\ X(Q_{\boldsymbol{\sigma}^2}^{-1} - B_{\boldsymbol{\sigma}^2}Q_{\boldsymbol{\sigma}^2}^{-1}B_{\boldsymbol{\sigma}^2}^{T})X^{T}\right). \tag{16}$$

Letting $\min_i \sigma^2 - \infty$ in (16) with $\lim_{\sigma^2 \to \infty} Q_{\boldsymbol{\sigma}^2}^{-1} = L^{-1}$, we obtain (15) with $Q^{-} = L^{-1}$. That is, for each t, the distribution of $\boldsymbol{\eta}^{(t)}$ is the posterior for $\boldsymbol{\eta}$.

To summarize our two results, Theorem 1 states that at convergence, weak association between

$\eta^{(t+1)}$ and $\eta^{(t)}$ arises as σ_0^2 grows large. Theorem 2 states that if *all* the σ_i^2 grow large, for each t, $\eta^{(t)}$ is approximately a sample from $f(\eta \mid \mathbf{Y})$.

3.3 The non-Gaussian first stage case

Suppose the first stage specification for the data is not Gaussian but a usual one parameter exponential family so that a generalized linear multilevel response model arises. If the joint posterior for β is approximately normal, we would expect Theorems 1 and 2 to still roughly hold.

The logic is as follows. Suppose that the likelihood is approximately proportional to $\exp\{-(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(X^TM^{-1}X)^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/2\}$ where $\widehat{\boldsymbol{\beta}}$ is the MLE and M is a diagonal matrix with M_{ii} equal to the square of the derivative of the link function evaluated at the estimated mean of Y_i multiplied by the variance function evaluated at the mean of Y_i (see, e.g., Agresti 1990, pp.448-449). Then with the prior in (4) we have that $\boldsymbol{\beta} \mid \widehat{\boldsymbol{\beta}}$ is approximately distributed as

$$N\left((X^TM^{-1}X + V_{\boldsymbol{\sigma}^2})^{-1}X^TM^{-1}X\widehat{\boldsymbol{\beta}}\,,\,(X^TM^{-1}X + V_{\boldsymbol{\sigma}^2})^{-1}\right), \tag{17}$$

analogous to (5). Since $\mathbf{Y}$ is treated as fixed, if $\widetilde{X}_0 = M^{-\frac{1}{2}}X_0$, then $X^TM^{-1}X$ is identical to (3) with $\widetilde{X}_0$ replacing X_0. Thus the calculations regarding $\boldsymbol{\eta}$ defined in Subsection 3.1 apply approximately here.

With regard to Theorem 2, if the target posterior is approximately normal, the Gaussian approximation approach (Sahu and Roberts, 1999) anticipates a similar continuity with regard to exact posterior sampling of $\boldsymbol{\eta}$. Indeed, in Subsection 3.2 we need only replace X^TX with $X^TM^{-1}X$.

4 Computational Findings with Normal Data

There seems little benefit in routine numerical illustration of Theorems 1 and 2. Of greater practical interest is whether these results continue to hold when variance components are unknown, particularly when the prior for the component is imprecise but with a large mean. In such cases, analytical calculation becomes intractable. As a first illustration of this, consider the usual balanced one-way ANOVA model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \;\; i = 1, \ldots, k, \; j = 1, \ldots, m, \tag{18}$$

where $\epsilon_{ij} \overset{iid}{\sim} N(0, 1)$. Similarly to Subsection 3.1, (18) is of the form in (2) with X_0 being block diagonal having blocks equal to $m \times 1$ column vectors of 1's, Δ_1 being a $k \times 1$ column vector of 1's, $\beta_0 = (\alpha_1, \ldots, \alpha_k)^T$, and $\beta_1 = \mu$.

Turning to the prior, we assume $\alpha_i \overset{iid}{\sim} N(0, \sigma_\alpha^2)$, and $\mu \sim N(0, \sigma_\mu^2)$ independently of the α_i. In the notation of (4), this means $V_0 = I_k$ and $V_1 = 1$. We use the BUGS language (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml) as our computational engine. This package makes the necessary programming essentially trivial, but does require a reparametrization to $\tau_\mu \equiv 1/\sigma_\mu^2$ and $\tau_\alpha \equiv 1/\sigma_\alpha^2$, with gamma priors for each. We use the notation $\tau \sim G(a, b)$ to denote a gamma distribution with mean a/b, and $IG(a, b)$ to denote an inverse Gamma distribution with mean $b/(a-1)$. Markov's inequality is useful in suggesting priors to encourage σ^2 large or small. That is, $P(\sigma^2 < c) = P(\tau^2 > c^{-1}) < ca/b$. So if a/b is small, e.g. $a/b = 10^{-2}$ and $c = 10$, then $P(\sigma^2 > 10) > .9$. Also, if $a > 1$, $P(\sigma^2 > c) < b/[(a-1)c]$. So if $a = 2, b = .1$ and $c = 1$, then $P(\sigma^2 < 1) > .9$.

We thus consider four illustrative specifications for the pair (τ_α, τ_μ), where in each case τ_α and τ_μ are a priori independent:

i	observations Y_{ij}, $j = 1, \ldots, 10$									
1	1.447	−0.331	2.673	0.288	−0.955	−0.247	1.795	−0.481	−0.841	0.867
2	−1.095	0.100	−0.022	0.520	0.305	1.650	0.153	0.808	0.886	0.854
3	1.594	1.975	1.722	−0.215	1.102	1.427	3.355	2.339	0.684	0.450
4	−1.775	−1.678	−2.638	−1.217	0.915	−0.399	−1.011	−2.307	0.636	−1.016
5	0.263	−0.246	0.962	0.041	0.656	1.319	0.427	2.441	−0.259	1.985

Table 1: Illustrative generated dataset, oneway ANOVA model.

(i) $\tau_\mu \sim G(2,2)$, $\tau_\alpha \sim G(1000,1)$

(ii) $\tau_\mu \sim G(2,2)$, $\tau_\alpha \sim G(2,100)$

(iii) $\tau_\mu \sim G(2,100)$, $\tau_\alpha \sim G(2,100)$

(iv) $\tau_\mu \sim G(\gamma,\gamma)$, $\tau_\alpha \sim G(\gamma,\gamma)$, for $\gamma = .001$

Case (ii) roughly meets the conditions of Theorem 1, while Case (i) is very far from these conditions ($P(\sigma_\alpha^2 < .1) > .99$). Case (iii) roughly meets the requirements for Remark 2. Case (iv) is a typical "default" specification in BUGS, yielding a prior which is quite vague and nearly improper.

In our investigation, we take $k = 5$, $m = 10$, and generate an illustrative dataset from the model (18) with $\mu = 0$ and $\tau_\alpha = 1$. The resulting data are shown in Table 1, and arise from a sampled α vector of (0.391, 0.320, 1.265, −0.918, 0.622). Initializing the mean parameters to 0 and the precision parameters to 1, we used BUGS to produce a single chain of 10,000 samples from the joint posterior distribution, following a burn-in period of 1000 iterations (more than sufficient for the chain to be in its post-convergence steady state). Table 2 gives the resulting lag 1 sample autocorrelations for μ, α_1, and $\eta_1 \equiv \mu + \alpha_1$ under each of the four prior specifications listed above. As expected, the post-convergence η_1 chain is essentially uncorrelated in Case (ii), but similarly small η_1 correlations are seen in all four cases. In Case (iii), the correlations in the μ and α_1 chains are very near 1, in concert with Remark 2. Finally, the BUGS default prior leads to correlations

case:	(i)	(ii)	(iii)	(iv)
prior for τ_μ:	$G(2,2)$	$G(2,2)$	$G(2,100)$	$G(\gamma,\gamma)$
prior for τ_α:	$G(1000,1)$	$G(2,100)$	$G(2,100)$	$G(\gamma,\gamma)$
μ	−.00769	.977	.996	.871
α_1	.0129	.891	.982	.494
η_1	−.0169	−.00504	−.00437	.00237

Table 2: Post-convergence lag 1 sample autocorrelations, oneway ANOVA model, with priors for τ_μ and τ_α as indicated (in Case (iv), $\gamma = .001$).

rather intermediate to those in the preceding cases.

5 A Poisson Regression Example

In this section we investigate whether the implications of our theorems still hold when we depart from the normal errors setting. In particular, we consider a spatial Poisson model that features the identifiability and overparametrization issues present in model (2). Let Y_i denote the number of disease events in region i. We assume $Y_i \overset{ind}{\sim} Poisson(E_i \exp(\eta_i))$, where E_i is a known *expected* number of events, and thus η_i is the log-relative risk of disease in region i, modeled linearly as

$$\eta_i = \mu + \theta_i + \phi_i,\ i = 1, \ldots, n\,. \tag{19}$$

Here μ is an overall intercept, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$ are vectors of region-specific random effects capturing regional *heterogeneity* and *clustering*, respectively (see the prior specification below). Clearly the mean structure in (19) can be written in the general form used in (2) by letting $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\beta}_0 = \boldsymbol{\theta}$, $\boldsymbol{\beta}_1 = \boldsymbol{\phi}$, $\boldsymbol{\beta}_2 = \mu$, and subsequently setting $X_0 = I_n$, $\Delta_1 = I_n$, and $\Delta_2 = \mathbf{1}_n$.

Turning to the prior specification, all three model components are given Gaussian specifications, namely $\mu \sim N(0, 1/\tau_\mu)$, $\theta_i \overset{iid}{\sim} N(0, 1/\tau_h)$, and $\phi_i \sim CAR(\tau_c)$. This lattermost notation refers to a *conditionally autoregressive* specification in which $\phi_i \,|\, \phi_{j \neq i} \sim N\left(\frac{1}{m_i}\sum_{j\; adj\; i} \phi_j \,,\, \frac{1}{m_i \tau_c}\right)$, where m_i is the number of regions adjacent to region i, and the sum in the prior mean is taken over these regions. Besag (1974) showed that this formulation is equivalent to an (improper) joint multivariate normal distribution for ϕ. The CAR prior is translation invariant, so a sum-to-zero constraint $\sum_{i=1}^{n} \phi_i = 0$ is typically imposed. A fully Bayesian model specification is completed by specifying fixed values or prior distributions for each of τ_μ, τ_h, and τ_c. Appropriate choices in this regard (seeking a "fair" prior balance between heterogeneity and clustering) are discussed in Bernardinelli et al. (1995), Best et al. (1999), and Carlin and Pérez (2000).

To illustrate this model, we return to the oft-analyzed Scottish lip cancer data of Clayton and Kaldor (1987). This dataset provides observed and expected cases of lip cancer in the 56 districts of Scotland for 1975-1980. Eberly and Carlin (2000) investigate convergence and Bayesian learning for this dataset and model, using fixed values for τ_μ, τ_h, and τ_c. We investigate Theorems 1 and 2 using several mutually independent prior specifications for these three parameters, namely

(i) $\tau_\mu \sim G(2,2)$, $\tau_h \sim G(1000,1)$, $\tau_c \sim G(1000,1)$

(ii) $\tau_\mu \sim G(2,2)$, $\tau_h \sim G(2,2)$, $\tau_c \sim G(2,100)$

(iii) $\tau_\mu \sim G(2,2)$, $\tau_h \sim G(2,100)$, $\tau_c \sim G(2,2)$

(iv) $\tau_\mu \sim G(2,100)$, $\tau_h \sim G(2,100)$, $\tau_c \sim G(2,100)$

(v) $\tau_\mu \sim G(10,10^5)$, $\tau_h \sim G(.001,.001)$, $\tau_c \sim G(.1,.1)$

(vi) $\tau_\mu \sim G(10,10^5)$, $\tau_h \sim G(3.2761,1.81)$, $\tau_c \sim G(1,1)$

Here Case (i) fails to meet the conditions of Theorems 1 or 2. Cases (ii) and (iii) roughly satisfy the conditions of Theorem 1, with $\beta_0 = \phi$ in (ii) and $\beta_0 = \theta$ in (iii) so we can compare these two possible choices. Case (iv) meets the conditions of Theorem 2. Case (v) is the "fair" specification

case:	(i)	(ii)	(iii)	(iv)	(v)	(vi)
prior for τ_μ:	$G(2,2)$	$G(2,2)$	$G(2,2)$	$G(2,100)$	$G(10,10^5)$	$G(10,10^5)$
prior for τ_h:	$G(1000,1)$	$G(2,2)$	$G(2,100)$	$G(2,100)$	$G(.001,.001)$	$G(3.2761,1.81)$
prior for τ_c:	$G(1000,1)$	$G(2,100)$	$G(2,2)$	$G(2,100)$	$G(.1,.1)$	$G(1,1)$
μ	.956	.996	.998	.999	.999	.998
θ_{28}	−.00392	.744	.810	.956	.198	.441
ϕ_{28}	.994	.891	.950	.975	.956	.954
η_{28}	.134	−.0266	−.0133	−.0177	−.00379	−.0712
d_1	.696	.719	.837	.951	.179	.500
d_2	.00499	.724	.770	.950	.254	.413
d_3	.322	−.0411	−.0127	−.00687	.0330	−.0571

Table 3: Post-convergence lag 1 sample autocorrelations, Scottish lip cancer data model, with priors for τ_μ, τ_h, and τ_c as indicated.

recommended by Best et al. (1999), while Case (vi) is an alternative such specification proposed by Carlin and Pérez (2000). Note that neither of these two papers uses a prior for τ_μ; the above specifications for τ_μ in these two cases essentially fix $\tau_\mu = .0001$.

Initializing all the parameters to 0, we again used BUGS to produce a single chain of 10,000 samples from the joint posterior distribution, following a burn-in of 1000 iterations (a period which again appears more than adequate in all cases). Table 3 is pertinent to Theorem 1, showing the lag 1 sample autocorrelations for four model parameters, μ, η_{28}, θ_{28}, ϕ_{28}, and three parameter contrasts, $d_1 \equiv \phi_{28} - \phi_1$, $d_2 \equiv \theta_{28} - \theta_1$, and $d_3 \equiv \eta_{28} - \eta_1$. (The 56 counties are arranged in increasing order of crude disease rate, so county 28 was selected as an "average" county.) Note that, of these seven quantities, only η_{28} and d_3 are estimable. The results are similar to those in Table 2 above. Autocorrelations are higher for η_{28} and d_3 in Case (i), but low in Cases (ii), (iii), and (iv), in concert with Theorem 1. The two "fair" specifications given in Cases (v) and (vi) also seem to produce acceptable autocorrelations for these two estimable parameters, and slightly lower autocorrelations for θ_{28}, d_1, and d_2 than in Cases (ii), (iii), and (iv).

We illustrate Theorem 2 by comparing trace plots and kernel density estimates (KDEs) for

η_1, η_{28}, and η_{56} using iterates 1–1000 to those using iterates 10,001–11,000. We use Cases (i) and (iv), and initialize all the chains to "bad" starting values far from the true posterior ($\mu^{(0)} = \phi_i^{(0)} = \theta_i^{(0)} = -3$, and $\tau_\mu^{(0)} = \tau_h^{(0)} = \tau_c^{(0)} = 1$) so that any resulting slow convergence will be apparent in the plots.

Figures 1(a) and (b) compare the results for Case (i). The burn-in period is clearly visible in the former, and the KDEs pairs look rather different. Figures 2(a) and (b) consider Case (iv). Now convergence is essentially immediate, and the sample trace and KDE pairs look very similar, suggesting that the $\eta_i^{(t)}$ are roughly draws from their true posterior for every t.

6 A Binary Response Three-Level Example

We turn to an illustration of our results using a three-stage multilevel generalized linear model where the response variable is binary. In particular, the response concerns the health status of root apexes of oak trees from the Mesola forest in the Veneto region of northeastern Italy. Full description of the dataset along with the questions of interest and a thorough data analysis, using multilevel models, is provided in Trevisani (1999).

Here we consider a portion of the data consisting of trees classified as nondeclining. We are naturally led to a multilevel structure. That is, for the nondeclining class, 5 trees were randomly selected. The area below the crown of each tree was partitioned into 6 sectors. Within each sector, 15 roots were randomly drawn. Finally, within each root 15 apexes were examined, starting at the distal part, for presence of ectomycorrhization. Ectomycorrhiza is a symbiosis occurring at the fine root apexes of the trees with some species of fungi, which improves uptake of water and nutrients and as a result, resistance to stress. Also recorded is the vitality of the apex as a binary response (1 = healthy, 0 = not). The primary objective of the study is to examine the relationship between vitality and ectomycorrhization.

Almost surely, the responses at the apexes are not independent. Correlation is introduced through sector level and root-within-sector level random effects. Covariate information at the apex level is a (centered) indication of ectomycorrhization. At the root level a centered and scaled root length is recorded as well as a categorical measure of extent of mycorrhiza (the number of apexes) having 4 categories: 0, 1–7, 8–14, and 15. Dummy variables are introduced for the last three categories; order is ignored. Finally, a sector-level classification, to reflect root distribution of ectomycorrhization, is introduced.

Hence, the model becomes:

$$\log \frac{\mathbf{p}_{ij}}{\mathbf{1} - \mathbf{p}_{ij}} = X_{ij}\boldsymbol{\eta}_{ij}$$

where $\mathbf{p}_{ij}$ is 15×1 with entries p_{ijk} denoting the probability of vitality status = 1 for the k^{th} apex in the j^{th} root in the i^{th} sector. X_{ij} is 15×2 with the first column consisting of 1's and the second of $\tilde{X}_{ijk}$, the apex level ectomycorrhization indicator. $\boldsymbol{\eta}_{ij}$ is 2×1 with

$$\eta_{ij1} = \varphi_1 c_{ij1} + \varphi_2 c_{ij2} + \varphi_3 c_{ij3} + \varphi_4 \tilde{\ell}_{ij} + \gamma_{i1} + v_{ij1},$$

$$\eta_{ij2} = \gamma_{i2} + v_{ij2}.$$

Here $\tilde{\ell}_{ij}$ is the standardized root length, the c_{ij}'s are the root level dummies and v_{ij1} and v_{ij2} are root-within-sector random effects. Finally,

$$\gamma_{i1} = \delta_1 + \delta_2 s_i + \mu_{i1},$$

$$\gamma_{i2} = \delta_3 + \delta_4 s_i + \mu_{i2},$$

where s_i is a dichotomous measure of sector level mycorrhizal distribution and μ_{i1} and μ_{i2} are sector level random effects.

prior case:	(i)	(ii)
δ_2	0.5140	0.9960
δ_4	0.5220	0.9880
v_{111}	0.0115	0.2830
v_{112}	0.0144	0.2220
μ_{11}	−0.0010	0.9870
μ_{12}	−0.0059	0.9790
d_1	0.1830	−0.0082
d_2	0.0490	−0.0042

Table 4: Post-convergence lag 1 sample autocorrelations, three-level forest data model.

Paralleling Subsection 3.1, but omitting details, we may write the mean vector on the logit scale as $X_0\boldsymbol{\beta}_0 + X_0\Delta_1\boldsymbol{\beta}_1 + X_0\Delta_2\boldsymbol{\beta}_2 + X_0\Delta_3\boldsymbol{\beta}_3$, where $\boldsymbol{\beta}_0$ is the set of v_{ij1}'s and v_{ij2}'s, $\boldsymbol{\beta}_1$ is the set of μ_{i1}'s and μ_{i2}'s, $\boldsymbol{\beta}_2^T = (\delta_1, \delta_2, \delta_3, \delta_4)$ and $\boldsymbol{\beta}_3^T = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)$. We model $\boldsymbol{v}_{ij} = \binom{v_{ij1}}{v_{ij2}} \sim N\left(\binom{0}{0}, \tau_v^{-1}I\right)$ so that $\sigma_0^2 = 1/\tau_v$ and V_0 is block diagonal with $I_{2\times2}$ as the blocks. Similarly, we model $\boldsymbol{\mu}_i = \binom{\mu_{i1}}{\mu_{i2}} \sim N\left(\binom{0}{0}, \tau_\mu^{-1}I\right)$ so that $\sigma_1^2 = 1/\tau_\mu$ and V_1 is again block diagonal with $I_{2\times2}$ as blocks. Under a binary regression, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ require proper priors to provide a proper posterior. For illustration, we use multivariate normals with mean $\mathbf{0}$ and diagonal covariance matrices which are a multiple of the diagonal part of the respective asymptotic covariance matrix resulting from fitting a standard logistic regression, ignoring all random effects.

Note that, with $i = 1, \ldots, 30$, $j = 1, \ldots, 15$, and $k = 1, \ldots, 15$, the response vector $\mathbf{Y}$ is 6750×1. Sharply discerning the qualitative conclusions of Theorems 1 and 2 using this large dataset with the foregoing complex model will be difficult. Nevertheless, we investigate using the following fixed values for the precisions τ_v and τ_μ: (i) $\tau_v = \tau_\mu = 1000$, and (ii) $\tau_v = \tau_\mu = 0.01$. Case (i) is far from the conditions of Theorems 1 and 2, while Case (ii) supports both. We keep the variability for $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ unchanged in both cases.

Table 4 is pertinent to Theorem 1, showing the lag 1 autocorrelations for eight parameters of interest. Only $d_1 = \log\left(\frac{p_{111}}{1-p_{111}}\right)$ and $d_2 = \log\left(\frac{p_{111}}{1-p_{111}} \cdot \frac{1-p_{211}}{p_{211}}\right)$ are estimable. The observable patterns in Table 4 from Case (i) to (ii) include a small but decreasing autocorrelation for the well-identified parameters (particularly d_2), and an increase for the level 1 random effects, v_{111} and v_{112}, the level 2 random effects, μ_{11} and μ_{12}, and the fixed coefficients, δ_2 and δ_4. The generally low autocorrelation for the level 1 random effects is most likely due to the large sample size.

To illustrate Theorem 2, Figures 3 and 5 show trace and KDE plots of the first 1000 iterations for the two estimable parameters; similarly Figures 4 and 6 for the post-convergence iterations 4001–5000. Comments analogous to those in the previous section can be made. In Cases (i) and (ii), adding randomness to τ_v and τ_μ through hyperpriors provides patterns that are qualitatively similar to those in Table 4 and Figures 3–6, but, not surprisingly, a bit more obscured, and thus are not presented.

References

[1] Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley.

[2] Bernardinelli, L., Clayton, D.G. and Montomoli, C. (1995). Bayesian estimates of disease maps: how important are priors? *Statistics in Medicine*, **14**, 2411–2431.

[3] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc., Ser. B*, **36**, 192–236.

[4] Best, N.G., Arnold, R.A., Thomas, A., Waller, L.A. and Conlon, E.M. (1999). Bayesian models for spatially correlated disease and exposure data (with discussion). In *Bayesian Statistics 6*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford: Oxford University Press, pp. 131–156.

[5] Browne, W.J. and Draper, D. (1999). A comparison of Bayesian and likelihood methods for fitting multilevel models. Technical report, Institute of Education, University of London.

[6] Carlin, B.P. and Pérez, M.-E. (2000). Robust Bayesian analysis in medical and epidemiological settings. To appear in *Robust Bayesian Analysis*, eds. D.R. Insua and F. Ruggeri. New York: Springer-Verlag.

[7] Clayton, D.G. and Bernardinelli, L. (1992). Bayesian methods for mapping disease risk. In *Geographic and Environmental Epidemology: Methods for Small-Area Studies*, eds. P. Elliot, J. Cuzick, D. English and R. Stern. Oxford: Oxford University Press.

[8] Clayton, D.G. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

[9] Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. B*, **41**, 1–31.

[10] Eberly, L.E. and Carlin, B.P. (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. To appear *Statistics in Medicine.*

[11] Gelfand, A.E. and Sahu, S.K. (1999). Identifiability, improper priors and Gibbs sampling for generalized linear models. *J. Amer. Statist. Assoc.*, **94**, 247–253.

[12] Gelfand, A.E., Sahu, S.K. and Carlin, B.P. (1995). Efficient parametrizations for normal linear mixed models. *Biometrika*, **82**, 479–488.

[13] Gelfand, A.E., Sahu, S.K. and Carlin, B.P. (1996). Efficient parametrizations 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.

[14] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.

[15] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1995). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

[16] Goldstein, M. (1995). *Multilevel Statistical Models (2nd Edition)*. London: Edward Arnold.

[17] Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **78**, 963–974.

[18] Lindley, D.V. (1971). The estimation of many parameters (with discussion). In *Foundations of Statistical Inference*, eds. V.P. Godambe and D.A. Sprott. Toronto: Holt, Rinehart and Winston, pp. 435–452.

[19] Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. B*, **34**, 1–41.

[20] Poirier, D.J. (1998). Revising beliefs in nonidentified models. *Econometric Theory*, **14**, 483–509.

[21] Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.

[22] Roberts, G.O. and Sahu, S.K. (1997). Updating schemes, correlation structure, blocking and parametrization for the Gibbs sampler. *J. Roy. Statist. Soc. B*, **59**, 291–317.

[23] Sahu, S.K. and Roberts, G.O. (1999). On the convergence of the EM algorithm and the Gibbs sampler. *Statistics and Computing*, **9**, 55–64.

[24] Searle, S.R. (1971). *Linear Models*. New York: Wiley.

[25] Spiegelhalter, D.J., Thomas A., Best, N.G. and Gilks, W.R. (1995). *BUGS: Bayesian inference using Gibbs sampling, Version 0.50.* Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

[26] Trevisani, M. (1999). Bayesian analysis of multilevel models. Unpublished Ph.D. thesis, Department of Statistical Sciences, University of Padova, Italy.

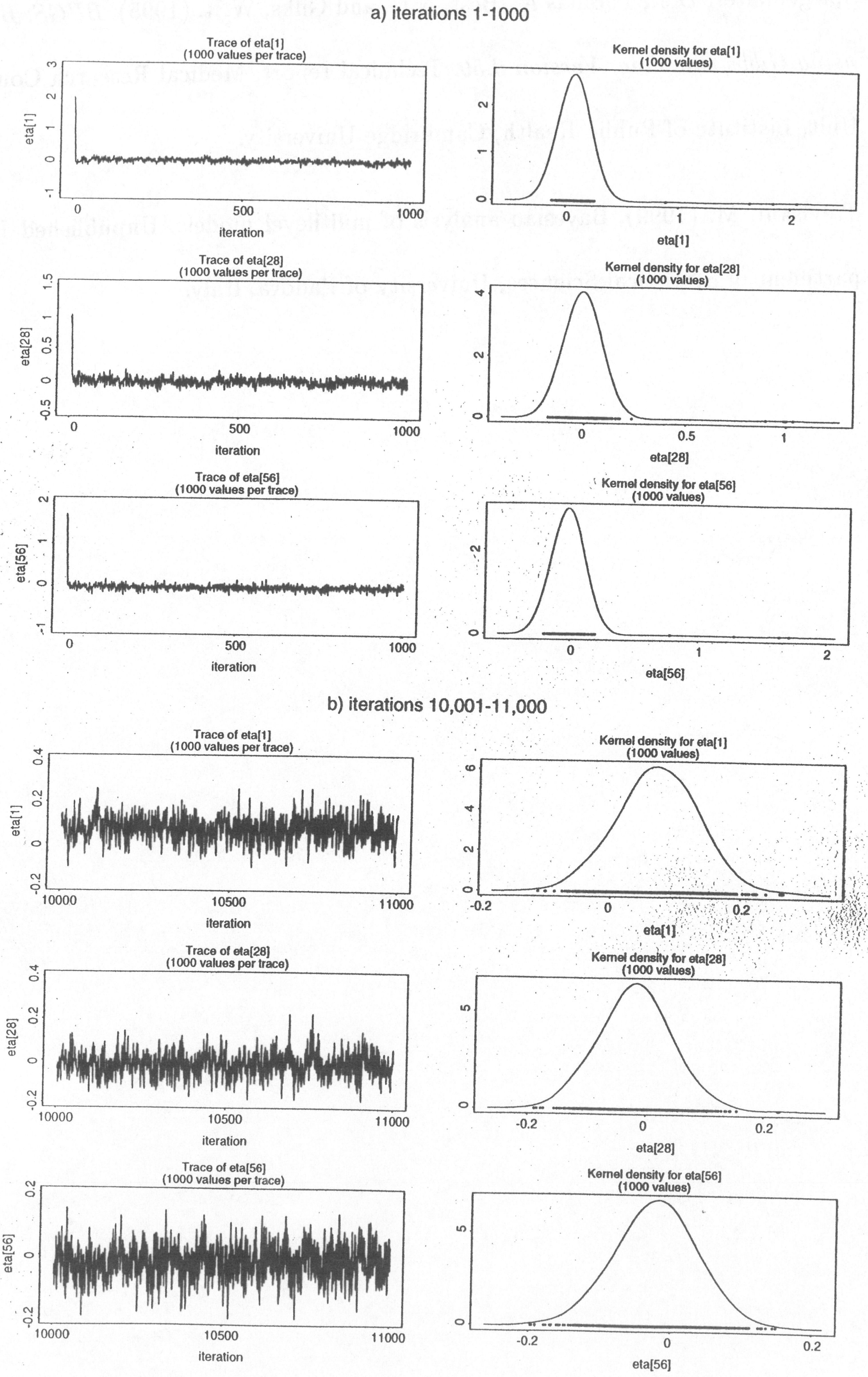
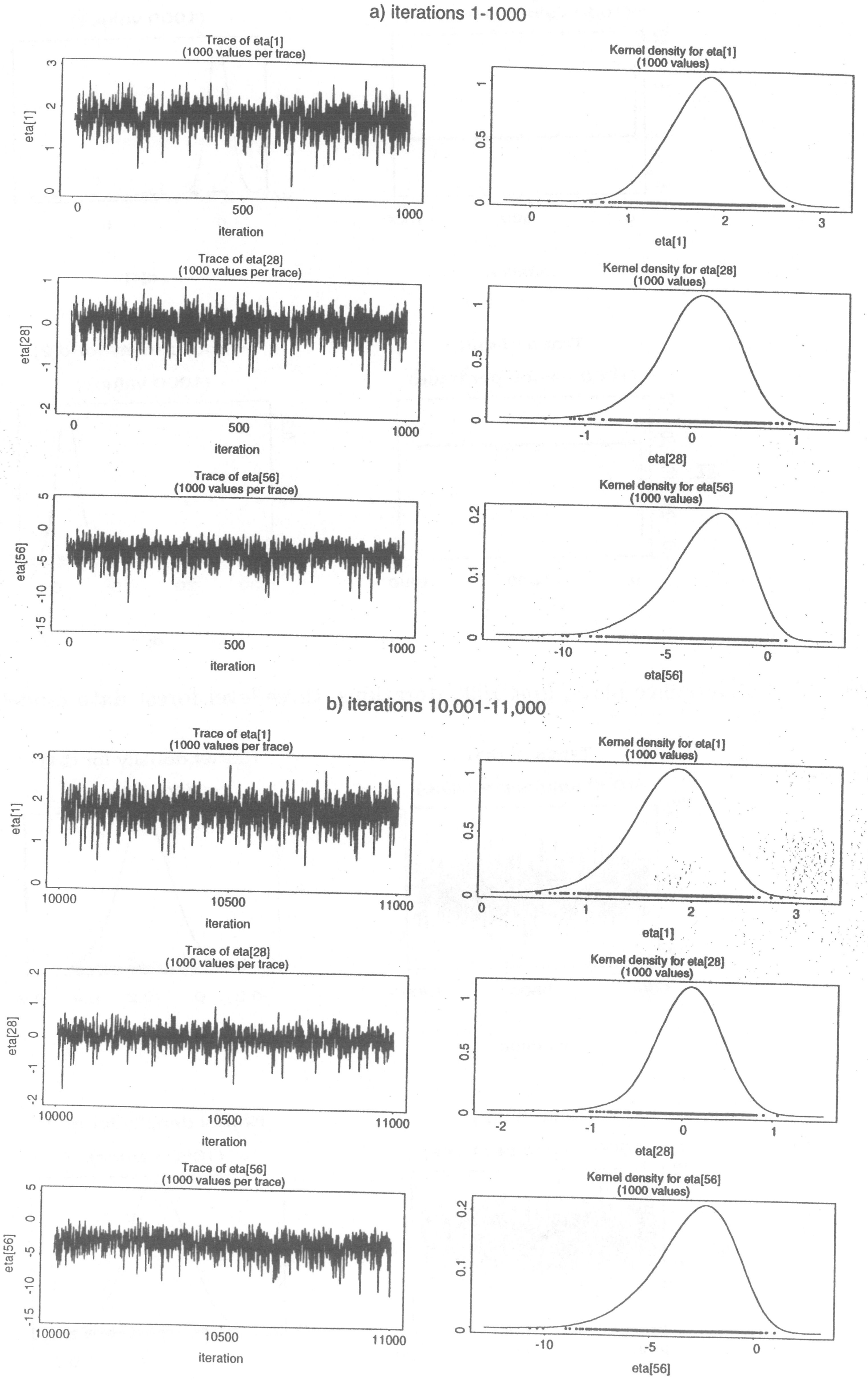


Figure 1: Convergence plots, first and last 1000 iterations, Scottish lip cancer data model, case (i) prior.



a) iterations 1-1000

b) iterations 10,001-11,000

Figure 2: Convergence plots, first and last 1000 iterations, Scottish lip cancer data model, case (iv) prior.

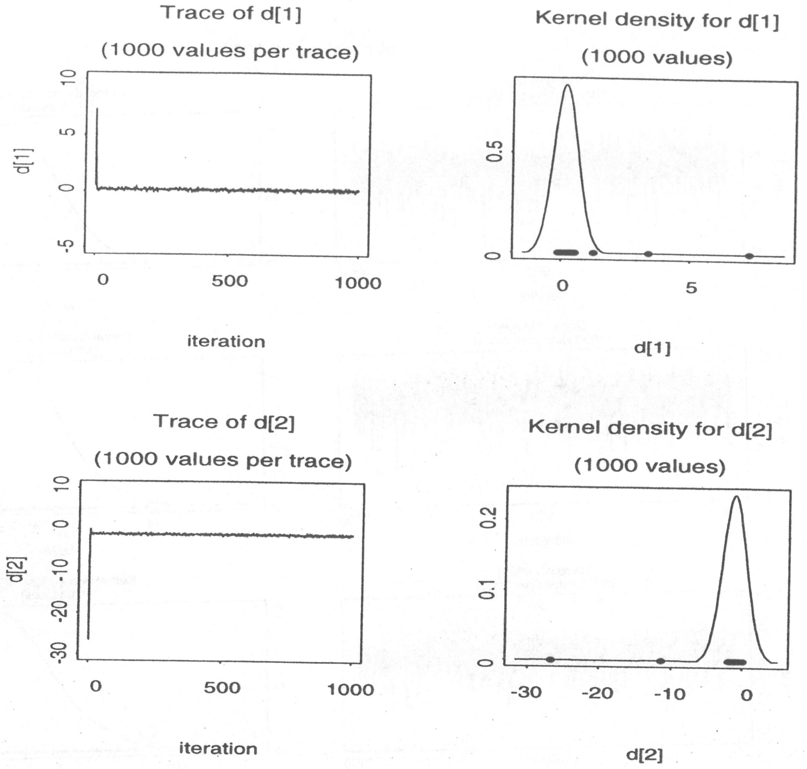


Figure 3: Convergence plots, first 1000 iterations, three-level forest data model, prior case (i).

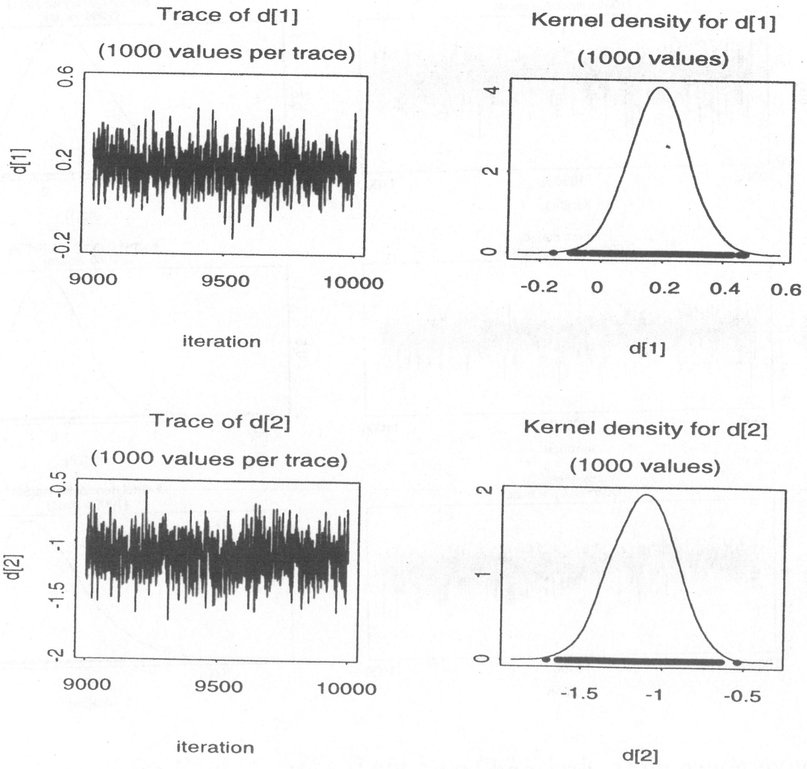


Figure 4: Convergence plots, last 1000 iterations, three-level forest data model, prior case (i).

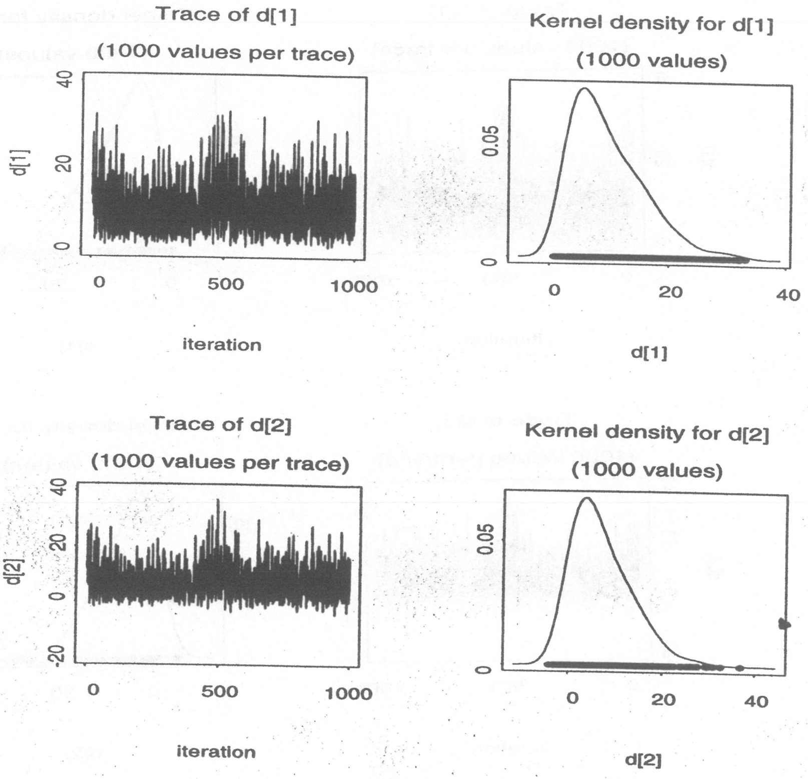


Figure 5: Convergence plots, first 1000 iterations, three-level forest data model, prior case (ii).

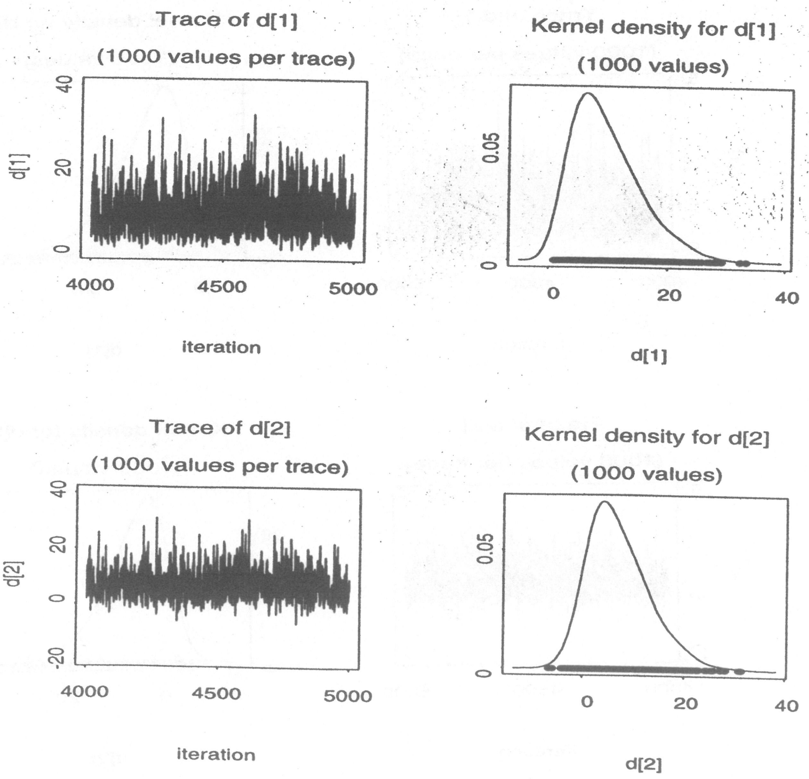


Figure 6: Convergence plots, last 1000 iterations, three-level forest data model, prior case (ii).

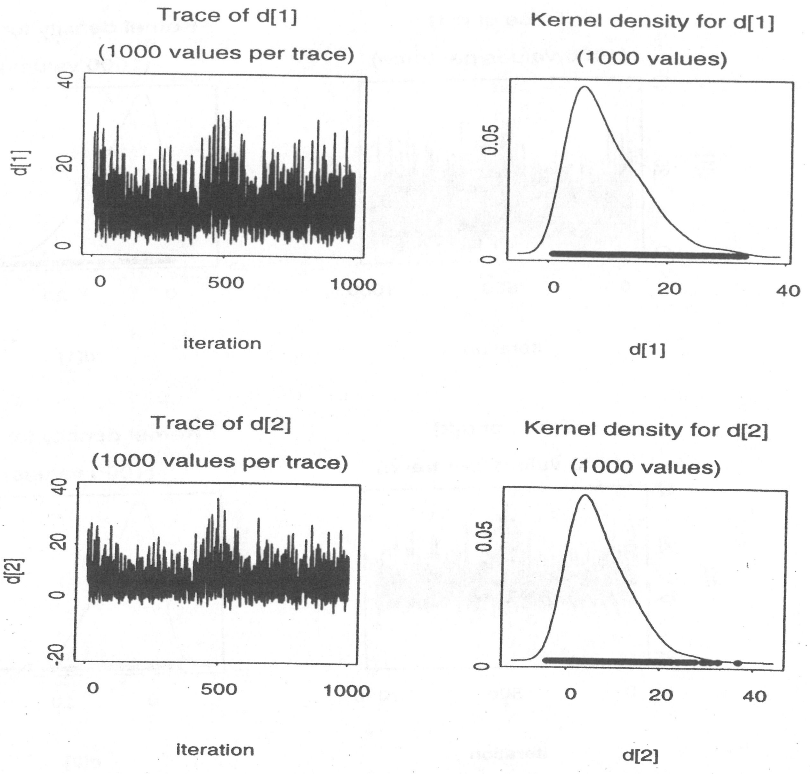


Figure 6: Convergence plots, first 1000 iterations, three-level forest data model, prior case (ii).

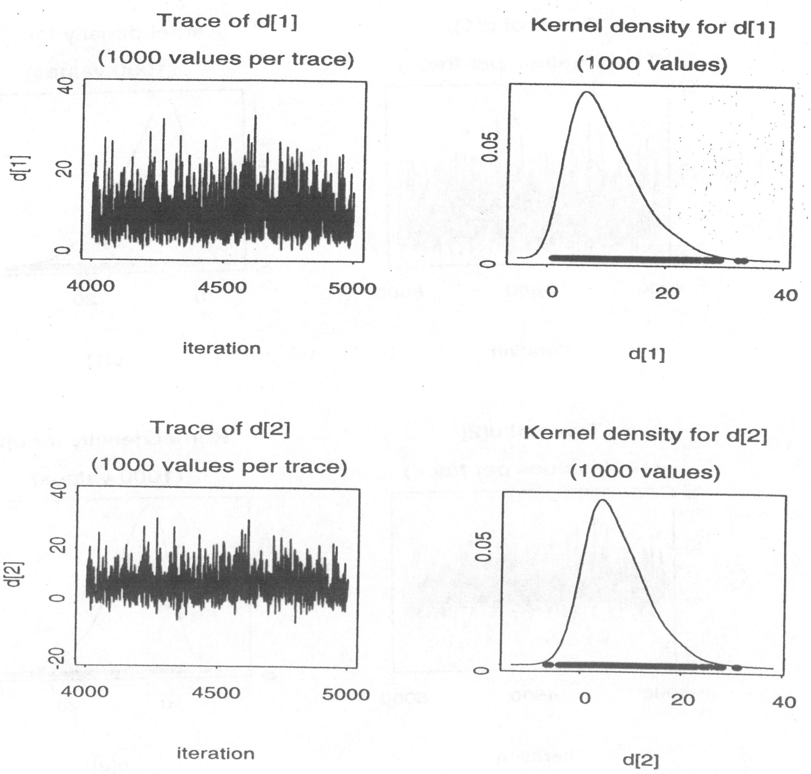


Figure 7: Convergence plots, last 1000 iterations, three-level forest data model, prior case (ii).