Identifiability, Improper Priors and Gibbs Sampling for Generalized Linear Models

Alan E. Gelfand and Sujit K. Sahu

September 8, 1998

Authors' Note

Alan E. Gelfand is a Professor in the Department of Statistics at the University of Connecticut, Storrs, CT 06269. Sujit K. Sahu is a Lecturer at the School of Mathematics, University of Wales, Cardiff, CF2 4YH, UK. The research of the first author was supported in part by NSF grant DMS 9301316 while the second author was supported in part by an EPSRC grant from UK. The authors thank Brad Carlin, Kate Cowles, Gareth Roberts and an anonymous referee for valuable comments.

Identifiability, Improper Priors and Gibbs Sampling for Generalized Linear Models

Abstract

Markov chain Monte Carlo algorithms are widely used in the fitting of generalized linear models (GLM). Such model fitting is somewhat of an art form requiring suitable trickery and tuning to obtain results one can have confidence in. A wide range of practical issues arise. The focus here is on parameter identifiability and posterior propriety. In particular, we clarify that non-identifiability arises for usual GLM's and discuss its implications for simulation based model fitting. Since often, some part of the prior specification is vague we consider whether the resulting posterior is proper, providing rather general and easy to check results for GLM's. We also show that if a Gibbs sampler is run with an improper posterior, it may be possible to use the output to obtain meaningful inference for certain model unknowns.

KEY WORDS AND PHRASES: Convergence; Embedded Posterior; Estimability; Integrability; Non-full rank models.

1 Introduction

Currently, simulation-based methods offer the best and often the only feasible approach to fitting complex hierarchical models. As such, these methods are enjoying widespread usage. Those who have worked with Markov chain Monte Carlo (MCMC) methods will have come to recognize a wide range of practical and theoretical issues which arise. Indeed, such model fitting is somewhat of an art form requiring suitable trickery and tuning to obtain results one can have confidence in. Elaboration of these various concerns and available remedies has been the subject of numerous articles in recent years. Accessible discussion is provided for example, in the book edited by Gilks *et al.* (1996) with further references therein.

We focus here on parameter identifiability and posterior propriety. In particular, the models we address are generalized linear models, in particular those introducing random effects referred to as generalized linear mixed models (GLMM's), which we argue are generally not identifiable. The remark of Lindley (1971, p.46), "In passing it might be noted that unidentifiability causes no real difficulty in the Bayesian approach" recognizes that a Bayesian analysis, in theory, is always possible by assigning proper priors for the model unknowns. However, the formal notion of Bayesian identifiability from Dawid (1979) and its equivalence to identifiability in the likelihood has implications for MCMC-based model fitting. We shall see below that for non-identifiable models we seek a middle ground with regard to the prior specification. Too precise a prior distribution will drive the inference mechanism; a prior too close to improper will yield an ill behaved posterior surface.

With GLM's, often at least some of the parametric components of the mean are, for convenience, given flat improper prior specifications raising the question of whether the resulting posterior is proper. We use a result from Ghosh *et al.* (1998) to obtain a rather general answer. Even with an improper posterior on the full dimensional posterior, it may be possible that there is an embedded model with a reduced dimensional parameter, for which we are able to sensibly define a unique proper posterior. The difficulty is apparent; one can not uniquely marginalize an infinite measure to a finite one. A simple example demonstrates and motivates one embedding construction in which we can provide such a determination using a formal limiting argument. Under this construction, if we run a Gibbs sampler with an improper posterior for which all the full conditional distributions are proper, we can use the output to obtain meaningful inference for the embedded model. Surprisingly, extending results from Roberts and Sahu (1997), for certain embeddings, an associated convergence result under MCMC model fitting encourages the specification of a full dimensional improper posterior.

A logistic regression with random effects provides a direct illustration of the foregoing points. It is not identifiable. It will yield an improper posterior under a flat prior for all of the effects. With such a prior, determination of a unique embedded proper posterior can

be implemented. Finally, fitting the model to a dataset from Agresti (1990), we observe, unexpectedly, that as we make the prior increasingly vague, the convergence behavior of the Gibbs sampler improves.

Thus, the format of the paper is as follows. In Section 2 we discuss the identifiability issue. Section 3 takes up the propriety problem. The embedding notion is developed in Section 4. Section 5 extends the convergence results for Gaussian Gibbs samplers from Roberts and Sahu (1997) to offer a possibly surprising finding under an improper posterior. The logistic regression extends this to a non-Gaussian case. We conclude with a brief summary.

2 Bayesian Identifiability

Here we consider notions related to model identifiability which are used in the sequel. In hierarchical models, stage-wise specification often introduces random effects yielding an overall parametric model of high dimension. Typically, for at least some of the parameters, there is a sense that the data provide little information, i.e., that these parameters are *weakly identified* and hence, that the model is weakly identified. The informality of this notion contrasts with Dawid's (1979) formal definition of Bayesian identifiability which we now recall.

In particular, suppose that the Bayesian model is denoted by likelihood $L(\boldsymbol{\theta}; \mathbf{y})$ and prior $f(\boldsymbol{\theta})$ and we partition $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. If

$$f(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1, \mathbf{y}) = f(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1) \tag{1}$$

we say that $\boldsymbol{\theta}_2$ is not identifiable. That is, if observing data $\mathbf{y}$ does not increase our prior knowledge about $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$, then $\boldsymbol{\theta}_2$ is not identified by the data. Non-identifiability occurs in the most rudimentary hierarchical specification, $f(\mathbf{y} \mid \boldsymbol{\theta}_1)\, f(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2)\, f(\boldsymbol{\theta}_2)$. Note that non-identifiability does not assert that there is no Bayesian learning. It does not imply that

$$f(\boldsymbol{\theta}_2 \mid \mathbf{y}) = f(\boldsymbol{\theta}_2). \tag{2}$$

In addition, since

$$f(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1, \mathbf{y}) \propto L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \mathbf{y})\, f(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)\, f(\boldsymbol{\theta}_1), \tag{3}$$

$\boldsymbol{\theta}_2$ is not identifiable if and only if $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \mathbf{y})$ is free of $\boldsymbol{\theta}_2$. Hence, Dawid's formal definition of Bayesian non-identifiability is equivalent to lack of identifiability in the likelihood. This equivalence is useful for arguing, as we do below, that any over-parameterized first stage generalized linear model, as for instance in customary ANOVA model specifications, will yield Bayesian non-identifiability. This equivalence also implies that identifiability does not depend upon the nature of the prior specification. Poirier (1996) investigates the above concerns using the terminology that the data $\mathbf{y}$ are marginally uninformative for $\boldsymbol{\theta}_2$ if and only if (2) holds while the data $\mathbf{y}$ are conditionally uninformative for $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$ if and only if (1) holds.

For the Bayesian specification $L(\boldsymbol{\theta}; \mathbf{y})\, f(\boldsymbol{\theta})$ suppose that there exists a one-to-one transformation from $\boldsymbol{\theta}$ to $(\boldsymbol{\delta}, \boldsymbol{\rho})$ such that, say $\boldsymbol{\rho}$ is not identifiable in the sense that (1) holds when we replace $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ by $\boldsymbol{\delta}$ and $\boldsymbol{\rho}$ respectively. Then the model is not Bayesianly identifiable. Section 3 clarifies how to construct $\boldsymbol{\delta}$ and $\boldsymbol{\rho}$ generally; obviously the reparameterization is not unique.

Continuing this logic, models which are not Bayesianly identifiable are contained within the class of *weakly identified* models. In particular, we could say that the parameter $\boldsymbol{\theta}_2$ is weakly identifiable if there exists $\boldsymbol{\theta}_1$ such that $f(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1, \mathbf{y}) \approx f(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)$ where, assuming the latter is proper, the approximation means that the two distributions are close under some suitable metric. This impression concurs with the aforementioned vagueness attached to the terminology and suggests an area for possibly fruitful research. Alternatively, one might prefer to think of $\boldsymbol{\theta}_2$ as weakly identified if $f(\boldsymbol{\theta}_2 \mid \mathbf{y}) \approx f(\boldsymbol{\theta}_2)$, though this may be difficult to formalize when $f(\boldsymbol{\theta}_2)$ is improper.

To illustrate the identifiability problem, consider the Gaussian linear model,

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4}$$

where $X_{n\times p}, \boldsymbol{\beta}_{p\times 1}$ with $r(X) = r < p$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$, with, for the moment, σ^2 known. If $\gamma = \ell^T\boldsymbol{\beta}$ and $\ell \in m(X^T)$, the manifold spanned by X, γ is said to be estimable. Viewed from a different perspective, we can say that $\mathbf{y}$ informs about $\boldsymbol{\eta} = X\boldsymbol{\beta}$ and hence, about any linear combination of the η's, $\mathbf{c}^T\boldsymbol{\eta} = \mathbf{c}^T X\boldsymbol{\beta} = \boldsymbol{\ell}^T\boldsymbol{\beta}$, i.e., about any estimable function. In fact, with regard to identifiability, we need not confine ourselves to linear transformations (since unbiased estimation is not the issue) whence, more generally $\mathbf{y}$ informs about any function $h(\boldsymbol{\eta})$.

We obtain similar identifiability problems as we move from the Gaussian linear model (4) to a generalized linear model. Here one assumes $E(\mathbf{y}) = \boldsymbol{\mu}$ and the linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta}$ is related to $\boldsymbol{\mu}$ using a link function g so that $g(\boldsymbol{\mu}) = \boldsymbol{\eta} = X\boldsymbol{\beta}$. As in the Gaussian linear model suppose further that the rank of X is $r < p$. Here also, we can say that $\mathbf{y}$ informs about $\boldsymbol{\mu}$, hence about $\boldsymbol{\eta} = X\boldsymbol{\beta}$. Consequently, $h(\boldsymbol{\eta})$ is identifiable in the likelihood for any function h. For linear functions of the parameters, we can say that the likelihood identifies estimable functions, e.g., contrasts and means but not, e.g., model main effects and interactions. We shall return to this model with more details in Section 3.2.

A practical consequence of a non-identifiable model when implementing simulation based model fitting using the Gibbs sampler (Gelfand and Smith, 1990), is the possibility of drifting. If the prior on $\boldsymbol{\rho}$ is not informative, trajectories of the Markov chain for components of $\boldsymbol{\rho}$ will tend to exhibit drift to very extreme values; there is nothing in the structure to *center* them. Such extreme parameter values can make convergence very difficult to assess and can lead to unstable computations and hence inaccurate estimates. Of course, priors which are too informative, i.e., too precise, will limit Bayesian learning from the data. With regard to prior specification in nonidentifiable models, we seek the ground between too informative and insufficiently informative. Often considerable simulation work is needed to clarify this middle ground.

3 Propriety of the Posterior

Obviously, a proper prior specification insures a proper posterior but with GLM's, improper priors (prior distributions) are often used for the parametric components of the mean to reflect *ignorance* or for mathematical convenience, allowing the possibility that the resultant posterior (posterior distribution) is improper. When, as in (1), the data are conditionally uninformative for $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$, Ghosh *et al.* (1998) provide a simple generic result which is apparent from (3).

Lemma 1 *Suppose that* $\mathbf{y}$ *has a non-identifiable density, i.e.,* $f(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = f(\mathbf{y}|\boldsymbol{\theta}_1)$ *and we adopt the improper prior* $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. *Then the posterior* $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})$ *is proper if and only if* $f(\boldsymbol{\theta}_1|\mathbf{y})$ *and* $f(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ *are both proper.*

Perhaps surprisingly, we can use this result to provide a general answer to and an easily checked condition for the integrability question for GLM's as we show below.

3.1 Propriety for Gaussian Models

Consider the one-way ANOVA model, $y_i = \mu + \alpha_i + \epsilon_i,\ i = 1, \ldots, n$ with $\epsilon_i \overset{iid}{\sim} N(0,1)$. Let $\boldsymbol{\beta}$ be the collection of parameters $(\boldsymbol{\alpha}, \mu)$ and define $\eta_i = \mu + \alpha_i$. Then the likelihood, denoted by $L(\boldsymbol{\beta}; \mathbf{y})$, is proportional to $\exp\{-\sum(y_i - \mu - \alpha_i)^2/2\}$. Let $f(\boldsymbol{\beta})$ be the prior on $\boldsymbol{\beta}$. Under a flat choice of $f(\boldsymbol{\beta})$, the posterior is clearly improper (integrate over $\boldsymbol{\alpha}$ first). However, a single proper prior on any component of $\boldsymbol{\beta}$ yields a proper posterior. For instance, suppose that $f(\boldsymbol{\beta}) = f(\mu)$ where $f(\mu)$ is proper. Reparameterize $\boldsymbol{\beta}$ to $(\boldsymbol{\eta}, \mu)$. Then the posterior distribution is proportional to,

$$\exp\left\{-\sum(y_i - \eta_i)^2/2\right\} f(\mu), \tag{5}$$

whence integrability with respect to $(\boldsymbol{\eta}, \mu)$, hence with respect to $\boldsymbol{\beta}$, follows. In contrast, suppose that we place a proper prior on say, $\boldsymbol{\eta}$. The posterior is still improper as may be

seen from (5). (As an aside, note that a conditionally exchangeable prior on the η_i, given μ, e.g., $f(\boldsymbol{\eta} \mid \mu) = N(\mu \mathbf{1}, \sigma_\alpha^2 I)$ with $f(\mu) = 1$ is equivalent to an unconditional exchangeable prior on the α_i, e.g., $f(\boldsymbol{\alpha}) = N(\mathbf{0}, \sigma_\alpha^2 I)$ and yields a proper posterior.)

The concept of estimability helps to explain what is going on in this simple model and more generally for the Gaussian linear model in (4). Without loss of generality we can write $X = (X_1^T, X_2^T)^T$ with X_1 being $r \times p$ with full row rank. Hence there exists A, $(n-r) \times r$, such that $X_2 = AX_1$ and also R, $(p-r) \times p$, such that the matrix $T = (X_1^T, R^T)^T$ is of full rank. Consider the transformation

$$\begin{pmatrix} \boldsymbol{\delta} \\ \boldsymbol{\rho} \end{pmatrix} = T\boldsymbol{\beta} = \begin{pmatrix} X_1\boldsymbol{\beta} \\ R\boldsymbol{\beta} \end{pmatrix}. \tag{6}$$

Hence, $X\boldsymbol{\beta} = (I_{r\times r}, A^T)^T \boldsymbol{\delta} = V\boldsymbol{\delta}$, say. The transformation from $\boldsymbol{\beta}$ to $(\boldsymbol{\delta}, \boldsymbol{\rho})$ is linear and one-to-one and the likelihood can be written as $L(X\boldsymbol{\beta}; \mathbf{y}) = L(V\boldsymbol{\delta}; \mathbf{y})$. Suppose that our prior information is expressed through a proper prior on $\boldsymbol{\gamma} = \Omega\boldsymbol{\beta}$. If $\boldsymbol{\gamma}$ is estimable, there exists a W such that $\boldsymbol{\gamma} = WX\boldsymbol{\beta} = WV\boldsymbol{\delta}$. The posterior for $\boldsymbol{\beta}$ is thus improper since $\int L(X\boldsymbol{\beta}; \mathbf{y})\, f(\Omega\boldsymbol{\beta})\, d\boldsymbol{\beta} < \infty$ if and only if $\int L(V\boldsymbol{\delta}; \mathbf{y})\, f(WV\boldsymbol{\delta})\, d\boldsymbol{\delta}\, d\boldsymbol{\rho} < \infty$ and the latter condition fails. We thus have the following result.

Theorem 1 *Consider the Gaussian linear model* $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ *with* $X_{n\times p}, \boldsymbol{\beta}_{p\times 1}$ *with* $r(X) = r < p$ *and* $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$. *Suppose that the prior* $f(\boldsymbol{\beta}, \sigma^2)$ *takes the form* $f(\boldsymbol{\beta})\, f(\sigma^2)$ *where* $f(\boldsymbol{\beta}) = f(\boldsymbol{\gamma})$, *a proper prior on* $\boldsymbol{\gamma} = \Omega\boldsymbol{\beta}$ *with* $\Omega\boldsymbol{\beta}$ *estimable. Then* $f(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})$ *is improper.*

A noteworthy class of priors on estimable linear functions are the pairwise difference priors discussed in, e.g., Besag *et al.* (1995) whence the posterior for $\boldsymbol{\beta}$ will be improper as discussed in Roberts *et al.* (1995). In this context, typically in practice, a non-estimable constraint is imposed on the sampling. This is a special (degenerate) case of the more general situation where a proper prior specification on some non-estimable parameters will be needed to enable a proper posterior. For instance, if σ^2 is known in (4) and the prior information on $\boldsymbol{\beta}$ is of the form of a proper prior on $\boldsymbol{\rho}$ we do obtain a proper posterior since,

with $r(X_1) = r \leq n$, $L(V\boldsymbol{\delta};\, \mathbf{y})$ is integrable with respect to $\boldsymbol{\delta}$. More generally, a prior on $\boldsymbol{\beta}$ induces a prior on $(\boldsymbol{\delta}, \boldsymbol{\rho})$. If this prior is integrable with respect to $\boldsymbol{\rho}$ and the marginal prior for $\boldsymbol{\delta}$ combined with $L(V\boldsymbol{\delta};\mathbf{y})$ yields a proper posterior, then by Lemma 1, $f(\boldsymbol{\delta}, \boldsymbol{\rho}|\mathbf{y}, \sigma^2)$ is proper. Note that the marginal prior for $\boldsymbol{\delta}$ need not be proper.

If σ^2 is unknown and we adopt a proper prior $f(\sigma^2)$ then $f(\boldsymbol{\delta}, \boldsymbol{\rho}, \sigma^2|\mathbf{y})$ is proper if

$$\int L(V\boldsymbol{\delta}, \sigma^2; \mathbf{y})\, f(\boldsymbol{\delta})\, f(\boldsymbol{\rho}|\boldsymbol{\delta})\, f(\sigma^2)\, d\boldsymbol{\rho}\, d\boldsymbol{\delta}\, d\sigma^2 < \infty.$$

Again if $f(\boldsymbol{\rho}|\boldsymbol{\delta})$ is proper and $f(\boldsymbol{\delta})$ is bounded then a sufficient condition for integrability is

$$\int \left(\frac{1}{\sigma^2}\right)^{n/2} f(\sigma^2) d\sigma^2 < \infty.$$

This holds for typical inverse gamma choice for $f(\sigma^2)$. As a slightly more general illustration, extend (4) to the customary mixed effects form

$$\mathbf{y}|\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \sigma_e^2 \sim N\left(X^{(1)}\boldsymbol{\beta}^{(1)} + X^{(2)}\boldsymbol{\beta}^{(2)}, \sigma_e^2 I\right),\ \boldsymbol{\beta}^{(2)}|\boldsymbol{\mu}, D \sim N\left(Z\boldsymbol{\mu}, D\right)$$

where D is a diagonal matrix whose first q_1 entries are σ_1^2, the next q_2 entries are σ_2^2, etc. Assume flat prior on $\boldsymbol{\beta}^{(1)}$ and on $\boldsymbol{\mu}$ with a proper prior on σ_e^2 and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)$. Then $f(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\mu}, \sigma_e^2, \boldsymbol{\sigma}^2|\mathbf{y})$ is proper if σ^2's are a priori independent with inverse gamma priors. Hobert and Casella (1996) provide necessary and sufficient conditions for propriety under improper limits of these inverse gamma priors.

Note that setting unidentified parameters to zero, as is often done in classical analysis of linear models to obtain unique solutions to the normal equations (see e.g., Searle, 1971 p209) amounts to putting a proper point mass prior on the unidentified parameters $\boldsymbol{\rho}$, which makes the posterior distribution proper. However, if inference regarding estimable parameters is of interest, we argue in Section 5 that, when fitting such a model using the Gibbs sampler, often the less informative the choice for $\boldsymbol{\rho}$, the more rapid the convergence.

3.2 Propriety for the Generalized Linear Model

We now turn to the integrability for the Bayesian generalized linear model with non-full rank design matrix and known or intrinsically specified dispersion parameter. Suppose then we observe conditionally independent y_i given θ_i such that

$$f(y_i \mid \theta_i) = c(y_i; \phi) \exp[\upsilon_i \phi^{-1}\{y_i\theta_i - b(\theta_i)\}], \quad i = 1, \ldots, n. \tag{7}$$

That is, y_i comes from a one parameter exponential family with given sample size υ_i and dispersion parameter ϕ whence $E(y_i \mid \theta_i) = b'(\theta_i) \equiv \mu_i$ and $\text{var}(y_i \mid \theta_i) = \upsilon_i \phi^{-1} b''(\theta_i) = \upsilon_i \phi^{-1} b''^{-1}(\mu_i) \equiv \upsilon_i \phi^{-1} V(\mu_i)$. We further assume that the θ_i are explained through a linear model by the relation $\mathbf{x}_i^T \boldsymbol{\beta} = \eta_i = g(\mu_i)$, where $\boldsymbol{\beta}$ is a $p \times 1$ unknown coefficient vector, g is the link function of the model and $\mathbf{x}_i$ is the design vector for the ith observation. Then $\theta_i = h(\eta_i) \equiv b'^{-1}(g^{-1}(\eta_i))$ so that the likelihood function of $\boldsymbol{\beta}$ is proportional to

$$\exp\left[\sum_i \left\{\upsilon_i \phi^{-1} \left(h(\mathbf{x}_i^T \boldsymbol{\beta}) y_i - b(h(\mathbf{x}_i^T \boldsymbol{\beta}))\right)\right\}\right]. \tag{8}$$

Assembling the $\mathbf{x}_i$'s into an $n \times p$ design matrix X and the η_i into a vector $\boldsymbol{\eta}$, on the transformed scale we have the linear model $\boldsymbol{\eta} = X\boldsymbol{\beta}$. Following the same argument as for the Gaussian model in Section 3.1, if $r(X) = r < p$, we can write (8) as

$$\exp\left[\sum_i \upsilon_i \phi^{-1}\left\{h(V_i \boldsymbol{\delta}) y_i - b(h(V_i \boldsymbol{\delta}))\right\}\right] \tag{9}$$

where V_i denotes the ith row of V. Again, with a proper prior solely on 'estimable' parameters, an improper posterior results. If, as before, we assume our prior information on $\boldsymbol{\beta}$ is captured through a proper prior on $\boldsymbol{\rho}$, then propriety of $f(\boldsymbol{\beta}|\mathbf{y})$ hinges upon the integrability of (9) with respect to $\boldsymbol{\delta}$, i.e., upon $f(\boldsymbol{\delta} \mid \mathbf{y})$ being proper. In general, applying Lemma 1, if the prior on $\boldsymbol{\beta}$ induces a proper prior on $\boldsymbol{\rho}$ given $\boldsymbol{\delta}$ we only need establish that $f(\boldsymbol{\delta} \mid \mathbf{y})$ is proper.

When is $f(\boldsymbol{\delta} \mid \mathbf{y})$ proper? Assuming $f(\boldsymbol{\delta})$ is bounded (though not necessarily proper), if $r(X_1) = n$, then for the canonical link we can set $\theta_i = V_i \boldsymbol{\delta}$ so that integrability of (9) is the

same as integrability of $\exp\left[\sum_{i=1}^{n} \upsilon_i \phi^{-1} \{\theta_i y_i - b(\theta_i)\}\right]$ with respect to $\boldsymbol{\theta}$, i.e., the same as integrability of the conjugate prior associated with (7) at each y_i and $\upsilon_i \phi^{-1}$. But, Diaconis and Ylvisaker (1979) show that if each $\upsilon_i \phi^{-1} > 0$ and y_i belongs to the interior of the domain of μ_i, these conjugate priors are proper. For link functions other than the canonical link, integrability depends upon the behavior of the Jacobian from $\boldsymbol{\delta}$ to $\boldsymbol{\theta}$. If it is bounded, we may again appeal to the result of Diaconis and Ylvisaker. Under the canonical link, the likelihood is bounded for most GLM's of interest (Barndorff-Nielsen, 1977). Hence, if $r(X_1) = r < n$ we only need have integrability with respect to r of the θ_i's rather than all n of them. We may be able to select r θ_i's so that the associated $\upsilon_i \phi^{-1}$ and y_i satisfy the Diaconis and Ylvisaker conditions in which case again $f(\boldsymbol{\beta}|\mathbf{y})$ will be proper. We collect the foregoing discussion into the following theorem.

Theorem 2 *Consider the GLM in (8) with $\upsilon_i > 0, i = 1, \ldots, n$ and a canonical link. Collect the $\mathbf{x}_i$s into a matrix X with rank say, $r \leq n$. Let $(\boldsymbol{\delta}, \rho)$ be as given in (6) and assume that prior on $\boldsymbol{\beta}$ is such that $f(\boldsymbol{\rho}|\boldsymbol{\delta})$ is proper and $f(\boldsymbol{\delta})$ is bounded. Then a sufficient condition for a proper $f(\boldsymbol{\beta}|\mathbf{y})$ is that the likelihood is bounded and that at least r of the y_i's belong to the interior of their respective domains.*

We conclude with a summary of other propriety results in the literature for GLM's. Ibrahim and Laud (1991) investigate the propriety of fixed effects models for one parameter exponential families. They assume that X is of full column rank and obtain a sufficient condition for posterior integrability under either Jeffreys' or a flat prior for $\boldsymbol{\beta}$. The condition requires checking integrability of univariate integrals. Dey, Gelfand and Peng (1997) extend the class of GLM's to over-dispersed GLM's and, in the process, extend the integrability condition of Ibrahim and Laud to one requiring the checking of two-dimensional integrals. Dey *et al.* also note that when the likelihood is log concave and the prior is bounded, integrability follows immediately. Using bounding hyper-planes, upon exponentiation one can bound the posterior by an integrable exponential function. Wedderburn (1976) shows that log concave likelihoods arise under canonical links and, in some cases, under other

link functions as well. Recently, Natarajan and McCulloch (1995) establish necessary and sufficient conditions for posterior propriety in the class of binomial GLMs using a polyhedral cone argument.

4 Proper Embedded Posteriors

If the posterior is not integrable, can we sensibly argue that an embedded lower-dimensional parameter has a unique proper posterior? As a simple example under the one-way ANOVA model in Section 3.1, we have $y_i - y_j$ normal with mean $\alpha_i - \alpha_j$ and variance 2. A flat prior on $\boldsymbol{\beta}$ yields an improper posterior. It also induces a flat prior on $\alpha_i - \alpha_j$. Thus, the *embedded* Bayesian model with "data" $y_i - y_j$ and a flat prior on the "parameter" $\alpha_i - \alpha_j$ results in a proper posterior. More generally, a flat prior on $\boldsymbol{\beta}$ induces a flat prior on $\boldsymbol{\eta} = X\boldsymbol{\beta}$. The posterior for $\boldsymbol{\beta}$ is improper but the embedded Bayesian model with data $\mathbf{y}$ and a flat prior on $\boldsymbol{\eta}$ results in the proper posterior $f(\boldsymbol{\eta} \mid \mathbf{y}) = N(\mathbf{y}, I)$ and indeed for any function of $\boldsymbol{\eta}$, e.g., if $\boldsymbol{\gamma} = A\boldsymbol{\eta}$, $f(\boldsymbol{\gamma} \mid \mathbf{y}) = N(A\mathbf{y}, AA^T)$. Apparently, within Bayesian models yielding improper posteriors, it is possible to extract embedded models with parameters having proper posteriors.

It is natural to ask whether these posteriors are uniquely determined. In general, the answer must be no since there is no unique marginalization of an infinite measure to a finite one. As a simple illustration (suggested by a referee), suppose $f(x, y) = 1$ on the set $\{0, 1\} \times \{1, 2, \dots\}$. Though the discrete joint density is improper, intuitively, the marginal density for X appears to be proper with mass $1/2$ at 0 and at 1. Consider the one-to-one transformation from (X, Y) to (U, V) defined as follows: if $X = 0, Y = k$ then $U = 0, V = k$; if $X = 1, Y = 2k$ then $U = 1, V = k$; if $X = 1, Y = 2k - 1$ then $U = 2, V = k$. Thus operating formally, $f(u, v) = 1$ on the set $\{0, 1, 2\} \times \{1, 2, \dots\}$. The marginal distribution for U appears to place mass $1/3$ on 0, 1 and 2. Inverting back to (X, Y) the marginal distribution for X now appears to put mass $1/3$ at 0 and $2/3$ at 1. By obvious extension we

can create a one-to-one transformation which, upon inversion, will yield $Pr(X = 0)$ to be any rational number in (0, 1).

Thus starting with $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ such that $f_{\mathbf{Z}}(\mathbf{z}) \geq 0$ and $\int_{\mathbb{Z}} f_{\mathbf{Z}}(\mathbf{z})d\mathbf{z} = \infty$ we seek simple conditions which yield unique proper marginal density $f_{\mathbf{X}}(\mathbf{x})$ for $\mathbf{X}$. A bit more generally, suppose that we consider a one-to-one transformation of $\mathbf{z}$ to $\mathbf{w}$, $\mathbf{w} = t(\mathbf{z})$ which maps $\mathbb{Z}$ onto $\mathbb{W}$. Operating formally, define

$$f_{\mathbf{W}}(\mathbf{w}) = f_{\mathbf{Z}}(t^{-1}(\mathbf{w})) \mid J_{\mathbf{z}\to\mathbf{w}} \mid$$

where J denotes the Jacobian of the transformation. Then, $\int_{\mathbb{W}} f_{\mathbf{W}}(\mathbf{w})d\mathbf{w} = \infty$.

Letting $\mathbf{w} = (\mathbf{u}, \mathbf{v})$, suppose that $\mathbb{W}$ arises as a product space for $\mathbf{u}$ and $\mathbf{v}$, i.e., $\mathbb{W} = \mathbb{U}\times\mathbb{V}$. Suppose further that for almost every $(\mathbf{u}, \mathbf{v})$ we have $f_{\mathbf{W}}(\mathbf{u}, \mathbf{v}) = f_1(\mathbf{u})\, f_2(\mathbf{v})$. Consider two increasing sequences of sets $\{A_n\}$ and $\{B_n\}$ such that $\{A_n\} \uparrow \mathbb{U}$ and $\{B_n\} \uparrow \mathbb{V}$. Defining $C_n = t^{-1}(A_n \times B_n) = \{\mathbf{z} : t(\mathbf{z}) \in A_n \times B_n\}$, assume that, for each n, $\int_{C_n} f_{\mathbf{z}}(\mathbf{z})dz < \infty$. Then

$$\int_{C_n} f_{\mathbf{Z}}(\mathbf{z})d\mathbf{z} = \int_{A_n\times B_n} f_{\mathbf{W}}(\mathbf{w})d\mathbf{w} = \int_{A_n} f_1(\mathbf{u})d\mathbf{u} \int_{B_n} f_2(\mathbf{v})d\mathbf{v}.$$

Hence, there exists a sequence $\{a_n,\ n = 1, 2, ...\}$ such that for each n, $f_1(\mathbf{u})/a_n$ is a proper density on A_n. Assume that $\int_{\mathbb{U}} f_1(\mathbf{u})d\mathbf{u} = a < \infty$. Then by the monotone convergence theorem $a_n \to a$. Hence, $f_1(\mathbf{u})/a$ is the *unique* proper density on $\mathbb{U}$ which arises under such limiting operations and we call $f_1(\mathbf{u})/a$ the embedded proper density for $\mathbf{u}$ on $\mathbb{U}$ within the improper joint density $f_{\mathbf{Z}}(\mathbf{z})$ on $\mathbb{Z}$. In practice, we only need check the integrability of $f_1(\mathbf{u})$ to assert that $\mathbf{u}$ has a unique, embedded proper density. Note the need for the condition that $\mathbb{W}$ arise as a product space. That is, to insure uniqueness it is not sufficient that $\int_{\mathbb{U}} f_1(\mathbf{u})d\mathbf{u} < \infty$. In the above discrete example, the inversion from (U, V) back to (X, Y) is not on $(\mathbb{X} \times \mathbb{Y})$; the set of Y's depends upon the value of X.

Working with Bayesian GLM's as in (8), our interest is in the case where $\mathbf{w} = (\boldsymbol{\delta}, \boldsymbol{\rho})$ is a linear transformation given by (6) of $\mathbf{z} = \boldsymbol{\beta}$ whence $\mathbb{Z} = \mathbb{W} = \mathbb{R}^p$ and $\mathbb{W}$ is immediately a product space for $\boldsymbol{\delta}$ and $\boldsymbol{\rho}$. Suppose that $f_{\mathbf{Z}}(\mathbf{z}) = L(X\boldsymbol{\beta}; \mathbf{y})\, f(\boldsymbol{\beta})$ is not integrable. Then

$$f_{\mathbf{W}}(\mathbf{w}) = L(Xt^{-1}(\boldsymbol{\delta}, \boldsymbol{\rho}); \mathbf{y})\, f(t^{-1}(\boldsymbol{\delta}, \boldsymbol{\rho})) \mid J_{\boldsymbol{\beta}\to(\boldsymbol{\delta},\boldsymbol{\rho})} \mid \tag{10}$$

which is also not integrable. In the case where $\boldsymbol{\rho}$ is non-identifiable given $\boldsymbol{\delta}$, (10) yields

$$f_{\mathbf{W}}(\mathbf{w}) \propto L(\boldsymbol{\delta};\mathbf{y})\, f(t^{-1}(\boldsymbol{\delta},\boldsymbol{\rho})).$$

Suppose that we can factor $f(t^{-1}(\boldsymbol{\delta},\boldsymbol{\rho})) = f_1(\boldsymbol{\delta})\, f_2(\boldsymbol{\rho})$. By the above argument, *regardless* of $f_2(\boldsymbol{\rho})$, if $L(\boldsymbol{\delta};\mathbf{y})\, f_1(\boldsymbol{\delta})$ is integrable over the support of $\boldsymbol{\delta}$ then

$$f(\boldsymbol{\delta}|\mathbf{y}) = L(\boldsymbol{\delta};\mathbf{y})\, f_1(\boldsymbol{\delta}) / \int L(\boldsymbol{\delta};\mathbf{y})\, f_1(\boldsymbol{\delta})\, d\boldsymbol{\delta} \tag{11}$$

is the unique proper posterior density for $\boldsymbol{\delta}$ embedded within the improper posterior for $\boldsymbol{\beta}$ and we can refer to $L(\boldsymbol{\delta};\mathbf{y})\, f_1(\boldsymbol{\delta})$ as an *embedded* proper Bayesian model for $\boldsymbol{\delta}$. Illustrative instances where the required factorization for $f(t^{-1}(\boldsymbol{\delta},\boldsymbol{\rho}))$ arises include a flat prior on $\boldsymbol{\beta}$ as well as any prior on $\boldsymbol{\beta}$ expressed through a proper prior on $\Omega\boldsymbol{\beta}$ where $\Omega\boldsymbol{\beta}$ is estimable.

5 Gibbs Sampling and Proper Embedded Posteriors

For a GLM with improper posterior on the parameter vector $\boldsymbol{\beta}$, if the Gibbs sampler is used to create a sequence of draws of $\boldsymbol{\beta}$, convergence for the entire sequence is not meaningful. However, if $\boldsymbol{\delta}$ is such that its posterior is proper, arising uniquely as in (11) can the associated sequence of $\boldsymbol{\delta}$'s be used to obtain inference about $f(\boldsymbol{\delta}|\mathbf{y})$? Such "marginalization over parameters" with suitable re-centering of the $\boldsymbol{\beta}$'s is routinely done, e.g., Besag *et al.* (1995, Section 4.1). Here we attempt a formal justification for this in the Gaussian case. For a proper Gaussian posterior, Roberts and Sahu (1997) formulate methodology for calculating the exact rate of convergence for a variety of Gibbs samplers. In the following discussion we use their methods to bring out a surprising result.

5.1 Theoretical Results

We return to the Gaussian linear model (4). We assume that $\mathbf{y} \sim N(X\boldsymbol{\beta}, I)$ and $r(X) = r < p$. In addition, suppose that a priori $\boldsymbol{\beta} \sim N(0, \tau^2 I)$. Let $\Sigma_{\tau^2}^{-1} \equiv Q_{\tau^2} = X^T X + \frac{1}{\tau^2} I$ and

$\boldsymbol{\mu}_{\tau^2} = \Sigma_{\tau^2} X^T \mathbf{y}$. Immediately we see that

$$\boldsymbol{\beta}|\mathbf{y} \sim N(\boldsymbol{\mu}_{\tau^2}, \Sigma_{\tau^2}). \tag{12}$$

Let $X^T X = L - U$ where L is the lower triangular part of $X^T X$ including all of the diagonal elements and U is obtained by subtraction. Let $Q_{\tau^2} = L_{\tau^2} - U$ where $L_{\tau^2} = L + \frac{1}{\tau^2} I$. Roberts and Sahu (1997) show that the Gibbs sampler transition kernel is given by,

$$\boldsymbol{\beta}^{(t+1)} \mid \boldsymbol{\beta}^{(t)}, \mathbf{y} \sim N(B_{\tau^2}\boldsymbol{\beta}^{(t)} + \mathbf{b}_{\tau^2}, \Sigma_{\tau^2} - B_{\tau^2}\Sigma_{\tau^2}B_{\tau^2}^T) \tag{13}$$

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

If $\tau^2 \to \infty$, 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 Σ_{τ^2} by a generalized inverse of $X^T X$. However, as above, $\boldsymbol{\delta} = X_1\boldsymbol{\beta}$ has a unique proper posterior distribution even as $\tau^2 \to \infty$. The following result describes what happens to the Gibbs sampler asymptotically.

Theorem 3 *Suppose that a Gibbs sampler with target density $f(\boldsymbol{\beta}|\mathbf{y})$ in (12) 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 $\tau^2 \to \infty$. In this limiting case, the iterates $\boldsymbol{\delta}^{(t)} = X_1\boldsymbol{\beta}^{(t)}$ are an exact sample from the unique density $f(\boldsymbol{\delta}|\mathbf{y})$.

Note that, because X is not of full column rank, in the limit (12) 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 $\boldsymbol{\delta}$. The

improper prior specification for $\boldsymbol{\beta}$ results in a Gibbs sampler which yields exact samples for the proper posterior of any estimable function. We now prove the second conclusion.

Proof: Applying the approach of Section 4 to the limiting case, it is straightforward to see that $f(\boldsymbol{\delta}|\mathbf{y})$ in (11) becomes

$$f(\boldsymbol{\delta}|\mathbf{y}) = N\left\{X_1(X^TX)^-X^T\mathbf{y},\ X_1(X^TX)^-X_1^T\right\} \tag{15}$$

for an arbitrary generalized inverse $(X^TX)^-$.

Next, note that L^{-1} always exist 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 $B_{\tau^2} \to B$ as $\tau^2 \to \infty$, both $XB_{\tau^2} \to 0$ and $X_1B_{\tau^2} \to 0$. From (13), for any τ^2 we have:

$$\boldsymbol{\delta}^{(t+1)} = X_1\boldsymbol{\beta}^{(t+1)} \mid X_1\boldsymbol{\beta}^{(t)}, \mathbf{y} \sim N\left\{X_1B_{\tau^2}\boldsymbol{\beta}^{(t)} + X_1\mathbf{b}_{\tau^2},\ X_1\left(\Sigma_{\tau^2} - B_{\tau^2}\Sigma_{\tau^2}B_{\tau^2}^T\right)X_1^T\right\}. \tag{16}$$

Letting $\tau^2 \to \infty$ in (16) with $\lim_{\tau^2\to\infty}\Sigma_{\tau^2} = L^{-1}$, we obtain (15) with $(X^TX)^- = L^{-1}$. That is, for each t, the distribution of $\boldsymbol{\delta}^{(t+1)}$ is the posterior for $\boldsymbol{\delta}$. □

When will L^{-1} be a generalized inverse of Q? An interesting practical result is that if X is of the form

$$X = (X_0\triangle_1,\ X_0\triangle_2, \ldots, X_0\triangle_{s-1},\ X_0), \tag{17}$$

when $s \geq 1$ and X_0 has full column rank, then (14) holds using an induction argument. (We can routinely check the $s = 2$ case. An X as in (17) arises, for instance in ANOVA specifications which are fully nested with main effects models which include an interaction involving all of the main effects. In either case X_0 is the portion of the design matrix associated with the highest order term in the model. The form (17) does not arise for additive models. For X as in (17), inversion of L provides an easy method to obtain a generalized inverse for X^TX.

Thus, in practice for handling Gaussian linear mixed effects models we can consider prior specifications which range from degeneracy (e.g., usual classical constraints) to flatness. Simulation study shows that the more vague the prior for $\boldsymbol{\beta}$, the more rapid the convergence for estimable functions suggesting some "continuity" in convergence behavior as we move from degeneracy to the flat specification covered by Theorem 3.

The above argument can be straightforwardly extended to the case where the error distribution is $N(\mathbf{0}, S)$ and also when the prior for $\boldsymbol{\beta}$ is $N(\mathbf{0}, \tau^2\Lambda)$. Lastly, if the target posterior distribution is only approximately normal, as it might be for $\boldsymbol{\beta}$ in a GLM, the Gaussian approximation approach of Sahu and Roberts (1998) suggests similar continuity in convergence behavior. The following example provides empirical support.

5.2 An Example

We illustrate the foregoing discussion with the fitting of a logistic growth curve model using data from Agresti (1990, p.397). He fits the model

$$\text{logit}(p_{ijk}) = \mu + \alpha_i^D + \alpha_j^T + \beta_0 x_k + \beta_j x_k$$

for comparing a new drug versus a standard drug ($j = 1$ for standard,=2 for new), for the ith diagnosis group, ($i = 1$ for mild, =2 for severe) at the kth occasion, ($k = 1, 2, 3$). The occasions are assumed equally spaced, i.e., $x_1 = -1, x_2 = 0$ and $x_3 = 1$. Here the model is not identifiable; we have 8 parameters but the rank of the X matrix is 5. Agresti imposes three constraints $\alpha_2^D = \alpha_2^T = \beta_2 = 0$. We assume that $\mu, \alpha_1^D, \alpha_1^T, \beta_0$ and β_1 follow independent $N(0, \tau_1^2)$ while α_2^D, α_2^T and β_2 follow independent $N(0, \tau_2^2)$. We choose $\tau_1^2 = 1$ (which is quite large for this problem). By setting τ_2^2 very small (10^{-5} in this case) we get essentially Agresti's constraints.

By direct calculation, the X matrix satisfies condition (13). As we let τ_2^2 increase from 10^{-5} to 1, we observe convergence to the conclusions of Theorem 3. In particular, the Gibbs samplers are implemented using the `BUGS` (Spiegelhalter *et al.*, 1996) software package.

$\log_{10}(\tau_2^2)$	-5	-4	-2	-1.8	-1.5	-1.3	-1	0
Lag 1 Cor	0.6166	0.5958	0.4288	0.3791	0.2821	0.1741	0.0804	0.0117

Table 1: Prior specification for the unidentified parameters and associated lag 1 auto-correlations for $\beta_1 - \beta_2$ under the Gibbs sampler.

Table 1 gives the lag 1 auto-correlations of $\beta_1 - \beta_2$ using 2000 values from the Gibbs samplers for different values of τ_2^2. It is seen that worst mixing (slowest convergence) of the Markov chain happens when we impose Agresti's constraints. Mixing improves as we specify an increasingly flat prior for the unidentified parameters α_2^D, α_2^T and β_2.

6 Summary

We have touched upon several issues relevant to the analysis of generalized linear models which are over-parameterized. We have noted that, while non-identifiability arises in such models, it does not preclude Bayesian inference as long as suitable informative prior is specified. Selection of suitable proper priors determines posterior propriety and can be connected to the familiar notion of estimability. The strength of this prior information affects the extent of Bayesian learning from the data as well as how well behaved the posterior is and thus how successful simulation-based model fitting will be. Even with an improper posterior, in certain cases we can uniquely define a proper posterior for a lower dimensional parameter. Moreover, if a Gibbs sampler is run in the customary way on the improper posterior, iterates can be used to infer about the lower dimensional proper posterior. Surprisingly, in some cases as the posterior for the full model tends to impropriety, the convergence behavior of the Gibbs sampler, in the context of the lower dimensional posterior, improves.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis,* New York: John Wiley and Sons.

Barndorff-Nielsen, O. (1977) *Information and Exponential Families in Statistical Theory.* New York: John Wiley and Sons.

Besag, J., Green, E., Higdon, D. and Mengersen, K. (1995) Bayesian Computation and Stochastic Systems, (with discussion). *Statistical Science*, **10**, 3–66.

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

Dey, D. K., Gelfand, A. E. and Peng, F. (1997) Overdispersed generalized linear models. *J. Statist. Planning and Inference*, **64**, 93–107.

Diaconis, P. and Ylvisaker, D. (1979) Conjugate priors for exponential families. *Ann. Statist.,* **7**, 269-281.

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

Ghosh, M., Ghosh, A., Chen, M. H. and Agresti, A. (1998) Bayesian Estimation for Item Response Models Using Improper Priors. Technical Report, University of Florida, Gainesville.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. G. (1996) *Markov Chain Monte Carlo In Practice*, London: Chapman and Hall.

Hobert, J. and Casella, G. (1996) The Effect of Improper Priors on Gibbs Sampling in Hierarchical Mixed Models. *J. Amer. Statist. Assoc.* **91**, 1461–1473.

Ibrahim, J. G. and Laud, P. W. (1991) On Bayesian analysis of generalized linear models using Jeffreys' prior. *J. Amer. Statist. Assoc.,* **86**, 981–986.

Lindley, D. V. (1971) *Bayesian Statistics: A Review* (SIAM).

Natarajan, R. and McCulloch, C. E. (1995) A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika*, **82**, 639–643.

Poirier, D. J. (1996) Revising beliefs in non-identified models. Technical Report, Department of Economics, University of Toronto.

Roberts, G. O. and Sahu, S. K. (1997) Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *J. R. Statist. Soc.*, B **59**, 291–317.

Roberts, G. O., Sahu, S. K. and Gilks, W. R. (1995) Comment on "Bayesian Computation and Stochastic Systems". *Statistical Science*, **10**, 49–51.

Sahu, S. K. and Roberts, G. O. (1998) On Convergence of the EM Algorithm and the Gibbs Sampler. To appear, *Statistics in Computing.*

Searle, S. R. (1971) *Linear Models*, New York: John Wiley and Sons.

Speigelhalter, D. J., Thomas, A., Best, N. G. and Gilks, W. R. (1995) *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50.* MRC Biostatistics Unit, Cambridge, England.

Wedderburn, R. W. M. (1976) On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, **63**, 27–32.