AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

CANTER DP 9605

Department of Economics, University of Canterbury
Christchurch, New Zealand

Discussion Paper No. 9605

June 1996

POSTERIOR SIMULATION AND MODEL CHOICE IN LONGITUDINAL GENERALIZED LINEAR MODELS

**Siddhartha Chib, Edward Greenberg,
Rainer Winkelmann**

LIBRARY
WITHDRAWN
AUG 14 1996
GIANNINI FOUNDATION OF
AGRICULTURAL ECONOMICS

Posterior Simulation and Model Choice in Longitudinal Generalized Linear Models

Siddhartha Chib*
Washington University, St. Louis, MO 63130, USA.

Edward Greenberg
Washington University, St. Louis, MO 63130, USA.

Rainer Winkelmann
University of Canterbury, Christchurch, New Zealand.

June 30, 1996

Abstract

This paper presents an approach to posterior simulation and model comparison for generalized linear models with multiple random effects. Alternative MCMC approaches for posterior simulation and alternative parameterizations are considered and compared in the context of panel data and multiple random effects. A straightforward approach for the calculation of Bayes factors from the MCMC output is developed. This approach relies on the computation of the marginal likelihood of each contending model. Estimation of modal estimates based on Monte Carlo versions of the E-M algorithm is also discussed. The methods are illustrated with several real data applications involving count data and the Poisson link function.

Keywords: Bayes factor; Count data; Gibbs sampling; Importance sampling; Marginal likelihood; Metropolis-Hastings algorithm; Markov chain Monte Carlo; Poisson regression.

1 Introduction

This paper is concerned with the problems of fitting and comparing longitudinal generalized linear models via simulation-based methods. The models of interest contain random effects within a non-linear model form and are not easily analyzed. This has led to growing interest in Markov chain Monte Carlo (MCMC) simulation methods and data augmentation methods as organizing computational tools to fit generalized linear models [Albert (1992), Bennett et al. (1996), Gamerman (1994), Wakefield et al. (1994) and Zeger and Karim (1991)]. Gelfand, Sahu, and Carlin (1996), however, have pointed out that some identification problems arise when the joint posterior distribution of the parameters and random effects is simulated. These problems can severely compromise the performance of MCMC methods. Our work is related to this literature but advances it in three important directions: first, we propose a simple parameterization of the model that is related to that in Gelfand, Sahu and Carlin (1996); second, we provide and (systematically) compare several alternative simulation methods for the random effects and isolate those that work well; and third, we develop an approach for model selection based on the computation of Bayes factors from the MCMC output.

Our interest in longitudinal GLM's arose from the desire to fit longitudinal Poisson regression models with random effects to a data set consisting of the number of absences from work for a sample of 704 male German workers. A second application that motivated our work is a data set on patents for a longitudinal sample of 680 firms in the United States. The purpose of the analysis is to explain the number of counts after allowing for regression effects (such as R&D spending) and firm-specific coefficients. Another data set of interest involves the effects of treating epilepsy with the drug progabide.

The estimation of these models raises several interesting problems. The new parameterization of the model proves to be important for the efficient simulation of the posterior

distribution. For contrast, we show that the MCMC output from the standard parameterization displays much higher serial correlation. Another general question concerns the simulation of the random effects, especially in the context of numerous clusters with multiple cluster-specific random effects. Existing approaches for simulating these random effects (for example those based on the accept-reject method) are too slow, whereas those based on the Metropolis-Hastings (M-H) algorithm (with Gaussian proposal densities) tend not to mix well. We report on some simple modifications of the M-H algorithm, requiring a multivariate-t proposal density as one component in a mixture proposal, that mitigate both problems. Other proposal densities are also discussed and compared.

The problem of model comparison is clearly important but it has not received much attention in the literature. Carlin and Chib (1995) and Green (1995) have developed model indicator-MCMC approaches for model comparison, but the use of these methods for longitudinal GLM's seems quite difficult. Lewis and Raftery (1994) have discussed another approach and applied it binary logistic models with a single random effect. Their approach relies on the Laplace method and consequently has an asymptotic justification that proves unreliable for small cluster sizes. Chib (1995) has developed a more flexible and accurate approach that forms the basis of the method in this paper. This approach leads to an estimate of the marginal likelihood of each fitted model and requires an estimate at a single point in the parameter space of the likelihood function, the prior density and the posterior density. The approach is quite straightforward and represents, we believe, an important advance in GLM model selection.

We also consider the use of MCMC methods for computing the maximum likelihood (ML) estimate. It is shown that the Monte Carlo EM (MCEM) algorithm of Wei and Tanner (1991) can be usefully applied for this purpose. We think that it important and interesting that MCMC methods can also be used to deliver the ML estimate. The latter is a useful summary of the likelihood function that can be used as a starting point for the full

Bayesian MCMC simulation. It turns out that the ML estimate (along with the posterior mean) is an ingredient in the marginal likelihood computation.

The rest of the paper is organized as follows. In Section 2 we discuss the simulation of the posterior distribution with the Metropolis-Hastings algorithm [Tierney (1994), Chib and Greenberg (1995)]. We consider several different implementations, each defined by a particular choice of proposal density in the M-H step. In Section 3 we show how the marginal likelihood may be computed from the MCMC output. This section also takes up the calculation of the maximum likelihood estimates and the computation of the likelihood function. In Section 4 we consider applications of the techniques to the epilepsy data, the patent data, and absentee data. The article ends with some concluding remarks in the final section.

2 MCMC sampling methods

2.1 The model

Let $y = \{y_{it}\}$ be data on subjects $i = 1, \ldots, n$ across time periods $t = 1, \ldots, T_i$. The generalized linear model of interest assumes that

$$y_{it}|\beta, b_i \sim p(y_{it}|\beta, b_i)$$

where $p(\cdot)$ is a member of the regular exponential family with conditional mean

$$\mu_{it} = E(y_{it}|\beta, b_i) = h(x'_{it}\beta + w'_{it}b_i)$$

and

$$b_i \sim \mathcal{N}_q(\eta, D).$$

Here the parameters $\beta \in \Re^k$ and $\eta \in \Re^q$ are the fixed effects, $b_i \in \Re^q$ are the random effects, h is the link function, x_{it} and w_{it} are vectors of covariates containing *no* variables in common, and $\mathcal{N}_q$ is the q-variate normal distribution. The form of h depends on the nature of the observations. For Poisson count data, for example, $\mu_{it} = \exp(x'_{it}\beta + w'_{it}b_i)$.

We complete the model by assuming that (β, η, D) follow the prior distributions

$$\beta \sim N(\beta_0, B_0^{-1}), \quad \eta \sim N(\eta_0, M_0^{-1}), \quad D^{-1} \sim \text{Wish}(\nu_0, R_0),$$

where $(\beta_0, B_0, \eta_0, M_0, \nu_0, R_0)$ are known hyperparameters and Wish $(\cdot, \cdot)$ is the Wishart distribution with ν_0 degrees of freedom and scale matrix R_0 [Press (1982)].

The likelihood function of this model is rather difficult to calculate although it may be expressed formally as follows. Let $y_i = (y_{i1}, \ldots, y_{iT_i})$ denote the observations on the ith cluster. Then (under conditional independence)

$$f(y_i|\beta, b_i) = \prod_{t=1}^{T_i} p(y_{it}|\beta, b_i).$$

The joint density of (y_i, b_i) is $f(y_i, b_i|\beta, \eta, D) = f(y_i|\beta, b_i)\, \phi(b_i|\eta, D)$, where ϕ is the density of the normal distribution. The likelihood function of the parameters given $y = (y_1, \ldots, y_n)$ may therefore be written as

$$\begin{aligned} L(y|\beta, \eta, D) &= \prod_{i=1}^{n} \int f(y_i, b_i|\beta, \eta, D)\, db_i \\ &\equiv \prod_{i=1}^{n} L_i(y_i|\beta, \eta, D), \end{aligned} \tag{1}$$

which is the product of the n likelihood contributions $L_i(y_i|\beta, \eta, D)$.

Remark: The parameterization above may be contrasted with that of Laird and Ware (1982) in which w_{it} is a subset of x_{it} and $E(b_i) = 0$. We do not recommend this parameterization with MCMC methods that rely on the simulation of the random effects. To see this, suppose for simplicity that the only overlap between x_{it} and w_{it} is x_{itk} and define $A_{itk} = \mu_{it} - x_{itk}(\beta_k + b_{ik})$ so that $\mu_{it} = (\beta_k + b_{ik})x_{itk} + A_{itk}$. But the first term is observationally equivalent to $b_{ik}x_{itk}$, implying that β_k is not likelihood identified [O'Hagan (1995)]. Identification is, therefore, achieved entirely through the *prior* distribution of b_i. As a result, if the variance D is large, an MCMC algorithm that simulates both β and b_i does not mix well. Transferring the "common" effect of x_k to η_k removes the nonidentified parameter

β_k. This parameterization is related to hierarchical centering introduced by Gelfand, Sahu, and Carlin (1995). It is easily shown by a change of variable that after integrating out the b_i the two parameterizations lead to identical likelihoods—our parameterization is thus completely general.

2.2 Sampling the random effects

As mentioned in the foregoing discussion, an operational MCMC scheme for simulating the posterior distribution requires data augmentation (Tanner and Wong (1987). The MCMC algorithm is then based on the blocks $b = (b_1, b_2, \ldots, b_n)$, β, η and D, and the associated full conditional distributions

$$[b|y, \beta, D]\ ;\ \ [\beta|y, \eta, b]\ ;\ \ [\eta|b, D]\ ;\ \ [D^{-1}|\eta, b]\ . \tag{2}$$

Starting with an (arbitrary) point in the parameter space, these distributions are sampled recursively, where the most recent values of the conditioning variables are used in the simulation. To implement this procedure we require methods for sampling each of the full conditional distributions. We show how this can done assuming for concreteness that the link function is exponential and the $\{y_{it}\}$ are count data distributed according to a Poisson distribution.

The main computational problem arises in the sampling of the random effects from the distribution $(b|y, \beta, \eta, D)$. A little algebra shows that this amounts to the simulation of b_i one at a time from the n (unnormalizable) *target densities* $\pi(b_i|y_i, \beta, \eta, D)$,

$$\begin{aligned}\pi(b_i|y_i, \beta, \eta, D\,) &\propto f(y_i, b_i|\beta, \eta, D) \\ &= \phi(b_i|\eta, D) \prod_{t=1}^{T_i} \exp[-\exp(x_{it}'\beta + w_{it}'b_i)]\ [\exp(x_{it}'\beta + w_{it}'b_i]^{y_{it}},\end{aligned}$$

where $\phi(b_i|\eta, D)$ is the density of the normal distribution with mean η and covariance D. We now discuss several methods for sampling this density. All these methods rely on the Metropolis-Hastings algorithm [Tierney (1994) and Chib and Greenberg (1995)].

Alternative methods do not appear to be practical. For example, the accept-reject approach (discussed by Zeger and Karim (1991) in a similar context) is generally difficult to apply with numerous clusters and random effects.

Before proceeding, we include a brief description of the M-H algorithm to fix notation. For a given target density $f(\psi)$, the M-H algorithm is defined by (1) a proposal density $q(\psi, \psi^\dagger)$ that is used to supply a proposal value $\psi^\dagger$ given the current value ψ and (2) a probability of move defined as

$$\alpha(\psi, \psi^\dagger) = \min\left\{\frac{f(\psi^\dagger)q(\psi^\dagger, \psi)}{f(\psi)q(\psi, \psi^\dagger)}, 1\right\}. \tag{3}$$

The proposal value $\psi^\dagger$ is accepted with probability $\alpha(\psi, \psi^\dagger)$; if rejected, the next sampled value is taken to be ψ. The different methods discussed next are defined by the choice of q.

Method 1: Random walk proposal

For this method let $q_1(b_i, b_i^\dagger) = \phi(b_i^\dagger | b_i, \tau_1 D)$, $i \leq n$, where τ_1 is a scalar that is adjusted in trial runs to obtain suitable candidates. With this choice, proposal values are obtained with little effort, but the sample can display considerable serial correlation.

Method 2: Tailored proposal

In this case, the proposal density is tailored to the target density around its modal value [Gammerman (1994)]. Let $\hat{b}_i$ denote the mode of $\ln f(y_i, b_i | \beta, \eta, D)$ and $V_{b_i} = (-H_{b_i})^{-1}$ the curvature around the mode. By direct computation, it can be seen that the gradient vector and Hessian matrix are given by

$$g_{b_i} = -D^{-1}(b_i - \eta) + \sum_{t=1}^{T_i} \left(y_{it} - \exp(x_{it}'\beta + w_{it}'b_i)\right) w_{it} \tag{4}$$

and

$$H_{b_i} = -D^{-1} - \sum_{t=1}^{T_i} \left(\exp(x_{it}'\beta + w_{it}'b_i)\right) w_{it}w_{it}', \tag{5}$$

respectively. These form the basis of a Newton-Raphson scheme to deliver the relevant quantities. We now define the (tailored) proposal density as $q_2 = \mathrm{MV}t(b_i | \hat{b}_i, \tau_2 V_{b_i}, \nu)$, where

τ_2 is another scaling factor and MVt is the multivariate-t distribution with ν degrees of freedom. In this case, the probability of move reduces to

$$\alpha(b_i, b_i^\dagger) = \min\left\{\frac{w(b_i^\dagger)}{w(b_i)}, 1\right\}, \; w(s) \equiv \frac{f(y_i, s|\beta, \eta, D)}{\text{MV}t(s|\hat{b}_i, \tau_2 V_{b_i}, \nu)}.$$

The adoption of the multivariate-t distribution is important. Gaussian proposals that have been recommended in the literature lead to much lower acceptance rates due to the fact that the weight $f(y_i, b_i|\beta, \eta, D)/\phi(b_i|\hat{b}_i, \tau_2 V_{b_i})$ in α above is unbounded as a result of the exponential decline of the Gaussian tails. Therefore, the probability of move is effectively zero in those places where the Gaussian density is thin. This leads to stickiness and poor mixing. Our suggestion (which is generally applicable) provides a simple and effective remedy.

Method 3: Mixture proposal—tailored proposal

In this method the proposal values are drawn from a mixture of proposal densities q_1 and q_2. To moderate the set up computations, q_2 is selected less frequently than q_1 (say every fifth iteration). It is important that the selection of the components be not based on the output of the chain so as to preserve the Markov property of the simulation.

Method 4: Acceptance-rejection with tailored proposal

In this case, the proposal value is obtained by an acceptance-rejection procedure applied to the pseudo-dominating function $c_i\text{MV}t(b_i|\hat{b}_i, \tau_2 V_{b_i}, \nu)$, where c_i is a positive number (its choice is discussed below). Note that we have again utilized the MVt distribution rather than the multivariate normal. Let $b_i^\dagger$ be a value generated from $\text{MV}t(b_i|\hat{b}_i, \tau_2 V_{b_i}, \nu)$ that satisfies the condition

$$u \leq f(y_i, b_i^\dagger|\beta, \eta, D)/c_i\text{MV}t(b_i^\dagger|\hat{b}_i, \tau_2 V_{b_i}, \nu),$$

where $u \sim \text{Unif}(0, 1)$. Let $C_1 = I(f(y_i, b_i|\beta, \eta, D) \leq c_i\text{MV}t(b_i|\hat{b}_i, \tau_2 V_{b_i}, \nu))$ be an indicator of whether the proposal density dominates the target at the current value b_i, and let

$C_2 = I(f(y_i, b_i^\dagger|\beta, D) \leq c_i \text{MV}t(b_i^\dagger|\hat{b}_i, \tau_2 V_{b_i}, \nu))$ be an indicator of domination at the proposal value $b_i^\dagger$. Then the probability of move [see Chib and Greenberg (1995, pg. 332)] is defined as

(a) $\alpha(b_i, b_i^\dagger) = 1$ if $C_1 = 1$;

(b) $\alpha(b_i, b_i^\dagger) = c_i \text{MV}t(b_i|\hat{b}_i, \tau_2 V_{b_i}, \nu)/f(y_i, b_i|\beta, D)$ if $C_1 = 0$ and $C_2 = 1$;

(c) $\alpha(b_i, b_i^\dagger) = \min\left\{f(y_i, b_i^\dagger|\beta, D)\text{MV}t(b_i|\hat{b}_i, \tau_2 V_{b_i}, \nu)/[f(y_i, b_i|\beta, D)\text{MV}t(b_i^\dagger|\hat{b}_i, \tau_2 V_{b_i}, \nu)], 1\right\}$ if $C_1 = 0$ and $C_2 = 0$.

Remark: We have developed a simple and automatic process for determining c_i for use in this algorithm (the value of ν is fixed at 15 in the examples). The recommendation is that

$$c_i = \frac{.6 \times f(y_i, \hat{b}_i|\beta, \eta, D)}{\text{MV}t(\hat{b}_i|\eta, D, \nu)},$$

which can be explained in the following way. The term $f(y_i, \hat{b}_i|\beta, \eta, D)/\text{MV}t(\hat{b}_i|\eta, D, \nu)$ forces the ordinates of the pseudo-dominating density and the (unnormalized) target density to agree at the mode $\hat{b}_i$. The factor .6 (other values might be tried) decreases the ordinates of the pseudo-dominating density at all values of b_i to improve the probability of generating values away from the mode and thereby attain greater mixing.

2.3 Sampling β, η, and D

Given the random effects, the remaining simulations are actually quite straightforward with both η and D being simulated from standard distributions. For β, the sampling requires the use of a M-H algorithm with an easily constructed (tailored) proposal density. In this case, the target density is proportional to

$$\phi(\beta|\beta_0, B_0^{-1}) \prod_{i=1}^{n} \prod_{t=1}^{T_i} \exp[-\exp(x_{it}'\beta + w_{it}'b_i)]\,[\exp(x_{it}'\beta + w_{it}'b_i)]^{y_{it}}.$$

It turns out that the mode $\hat{\beta}$ and curvature $V_\beta = [-H_\beta]^{-1}$ of the logarithm of this function at the mode are readily obtained, usually through a few Newton-Raphson steps. The required

gradient vector and Hessian matrix are given by

$$g_\beta = -B_0(\beta - \beta_0) + \sum_{i=1}^{n}\sum_{t=1}^{T_i}[y_{it} - \exp(x_{it}'\beta + w_{it}'b_i)]x_{it}$$

and

$$H_\beta = -B_0 - \sum_{i=1}^{n}\sum_{t=1}^{T_i}[\exp(x_{it}'\beta + w_{it}'b_i)]x_{it}x_{it}',$$

respectively. A tailored MVt density can now be constructed. We suggest that the proposal be obtained by the method of reflection. The general idea is to reflect the current value around the modal value before adding a Gaussian increment with variance $\tau_\beta V_\beta$. It is easy to check that the resulting proposal density is given by $q(\beta, \beta^\dagger) = \phi(\hat{\beta} - (\beta - \hat{\beta}), \tau_\beta V_\beta)$. This density is symmetric. Chib and Greenberg (1995) have documented the importance of reflection in other problems. We do not think that it is necessary to use a mixture proposal density in this case because the computational burden of finding the tailored density is minimal.

To complete one cycle of the MCMC simulation one now samples η from

$$\pi(\eta|b, D) = \text{MV}t(\eta|\hat{\eta}, M_1^{-1}, \nu), \tag{6}$$

where $\hat{\eta} = M_1^{-1}(M_0\eta_0 + \sum_{i=1}^{n} D^{-1}b_i)$ and $M_1 = (M_0 + nD^{-1})$, and D^{-1} from

$$\pi(D^{-1}|b) = f_W(D^{-1}|n + v_0, [R_0^{-1} + \sum_{i=1}^{n}(b_i - \eta)(b_i - \eta)']^{-1}),$$

where $f_W(\cdot|a, A)$ denotes a Wishart density with a degrees of freedom and scale matrix A. This completes the derivation and simulation of the full conditional densities required in the MCMC sampling.

3 Marginal likelihood by MCMC

From a practical viewpoint, the problem of model choice is one of the most important in fitting generalized linear models. We now show how this problem can be tackled given one

of the posterior simulator techniques discussed in the previous section. We focus on one of the central quantities in Bayesian model choice - the marginal likelihood of a model and show how it may be computed from the MCMC output. The marginal likelihood of given model $\mathcal{M}$ is the integral of the likelihood with respect to the prior density of the parameters, i.e.,

$$m(y|\mathcal{M}) = \int L(y|\mathcal{M}, \beta, \eta)\, \pi(\beta, \eta, D|\mathcal{M})\, d\beta\, dD, \tag{7}$$

where $\pi(\beta, \eta, D|\mathcal{M})$ is the model specific prior density (Jeffreys (1961) and Kass and Raftery (1995)). On the basis of the marginal likelihood one may compute the Bayes factor in favor of model $\mathcal{M}_k$ (and against model $\mathcal{M}_l$) as

$$B_{k,l} = \frac{m(y|\mathcal{M}_k)}{m(y|\mathcal{M}_l)}. \tag{8}$$

Chib (1995) discusses an alternative representation of the marginal likelihood

$$m(y|\mathcal{M}) = \frac{L(y|\mathcal{M}, \theta^*)\pi(\theta^*|\mathcal{M})}{\pi(\theta^*|\mathcal{M}, y)}, \tag{9}$$

leading to the estimate

$$\ln \hat{m}(y|\mathcal{M}) = \ln L(y|\mathcal{M}, \theta^*) + \ln \pi(\theta^*|\mathcal{M}) - \ln \hat{\pi}(\theta^*|\mathcal{M}, y). \tag{10}$$

where $\theta^* = (\beta^*, \eta^*, D^*)$ is some point in the parameter space, $\hat{\pi}(\theta^*|\mathcal{M}, y)$ is an estimate of the posterior ordinate at θ^* and all the functions on the right hand side are normalized. To gain some insight into the relevance of this approach, we note that the choice of point θ^* is arbitrary since the expression above is an identity in θ. Still, it has been suggested in Chib (1995) that the identity be evaluated at a high density point, such as the posterior mean of θ or the maximum likelihood estimate (whose computation is discussed below).

We now consider the calculation of each term in (10) from the MCMC output.

3.1 Likelihood function

We begin with the computation of the likelihood function at the point θ^*. It should be noted that this estimate is required at only a single point, which minimizes the computational

burden. Consider now the contribution of y_i to the likelihood at the point θ^*,

$$L_i(y_i|\theta^*) = \int f(y_i|b_i,\beta^*)\,\phi(b_i|\eta^*,D^*)\,db_i, \tag{11}$$

where the normalizing constants for both of the functions that appear under the integral are known. If b_i is of low dimension it is possible to compute this integral numerically by the method of quadrature. The likelihood contribution can also be computed by the Laplace approximation [see Tierney and Kadane (1986)] if the cluster size T_i is large. Then,

$$\ln \hat{L}_i(y_i|\theta^*) = \ln\{f(y_i|\hat{b}_i,\beta^*)\phi(\hat{b}_i|\eta^*,D^*)\} + 0.5q\ln(2\pi) + 0.5\ln|-H_{\hat{b}_i}^{-1}|,$$

where $\hat{b}_i$ denotes the mode of $\ln\{f(y_i|b_i,\beta^*)\phi(b_i|\eta^*,D^*)\}$, $H_{\hat{b}_i}$ the Hessian at the mode, and q is the dimension of b_i. These quantities are obtained by the methods discussed earlier in connection with the simulation of b_i.

The accuracy of the Laplace method depends crucially on T_i, the size of the ith cluster. To see how the asymptotic approximation can fail for small T_i, consider Poisson count data generated from the following model in which there are $n = 200$ clusters, two random effects ($q = 2$), and two fixed effect parameters and $T_i = 5$. Let

$$\beta = 0.5, \quad \eta = (-.5,-.8)', \quad \text{and } D = \begin{pmatrix} .3 & -.1 \\ -.1 & .2 \end{pmatrix},$$

and assume that $x_{it} \sim N(0,1)$, $w_{it1} = 1$, and $w_{it2} \sim N(0,1)$. The very accurate estimate of the log likelihood function based on quadrature is -1215.30, while the Laplace approximation is -1435.78, which is clearly in error.

An alternative method that is more reliable for small cluster sizes is importance sampling [see Geweke (1989)]. If $g(b_i)$ denotes an importance sampling function, the importance sampling estimate of $L_i(y_i|\theta^*)$ is

$$\hat{L}_i(y_i|\theta^*) = M^{-1}\sum_{j=1}^{M} \frac{f(y_i|b_i^{(j)},\beta^*)\,\phi(b_i^{(j)}|\eta^*,D^*)}{g(b_i^{(j)})},$$

where $b_i^{(j)}$ $(j = 1, \ldots, M)$ are i.i.d. draws from $g(b_i)$. A convenient choice for the latter is a multivariate-t distribution with location $\hat{b}_i$, scale matrix $(-H_{b_i})^{-1}$ and ν degrees of freedom. The log-likelihood function is obtained by adding the $\ln \hat{L}_i(y_i|\theta^*)$.

For the simulated data set described above, we let $M = 2000$ and specify 10 degrees of freedom for the multivariate-t importance function (the result are not sensitive to these choices). The importance sampling estimate of the likelihood is -1215.32, which agrees with the quadrature estimate up to the first decimal place. Thus, in this example with small cluster sizes, the importance sampling estimate of the likelihood is far more accurate than that based on the Laplace approximation.

3.2 Estimation of $\pi(\theta^*|y)$

We now develop a methodology for estimating the posterior density at θ^*. This approach is adapted from Chib (1995) where more details may be found. First, write the denominator of (9) as

$$\ln \pi(\theta^*|y) = \ln \pi(D^{-1*}|y) + \ln \pi(\eta^*|y, D^{-1*}) + \ln \pi(\beta^*|y, \eta^*, D^{-1*}), \tag{12}$$

and note that

$$\pi(D^{-1*}|y) = \int \pi(D^{-1*}|b, \eta)\pi(b, \eta|y)\, db, \tag{13}$$

$$\pi(\eta^*|y, D^*) = \int \pi(\eta^*|b, D^*)\pi(b|y, D^*)\, db, \text{ and} \tag{14}$$

$$\pi(\beta^*|y, \eta^*, D^*) = \int \pi(\beta^*|y, b, D^*)\pi(b|y, \eta^*, D^*)\, db. \tag{15}$$

The second step is concerned with the estimation of each of these ordinates from the MCMC output. A little reflection shows that to estimate (13) one simply requires output from the initial MCMC run consisting of the distributions

$$[\beta|y, b], \quad [b|y, \beta, \eta, D], \quad [\eta|b, D], \quad [D^{-1}|\eta, b]$$

The draws $\{b, \eta\}$ from this run are distributed according to $\pi(b, \eta|y)$. Therefore, an estimate of $\pi(D^{-1*}|y)$ is given by averaging the Wishart density $\pi(D^{-1*}|b, \eta)$ in (13) over these

simulated draws. Next, a reduced MCMC simulation consisting of the distributions

$$[\beta|y,b], \quad [b|y,\beta,\eta,D^*], \quad [\eta|b,D^*]\,,$$

where D is set equal to D^*, produces draws of $\{b\}$ that are distributed according to $\pi(b|y,D^*)$. These draws can be used to average the Gaussian full conditional density $\pi(\eta^*|b,D^*)$ in (14) at the point η^*. Finally, a reduced Gibbs run consisting of

$$[\beta|y,b], \quad [b|y,\beta,\eta^*,D^*]$$

leads to draws of β from the density $\pi(\beta|y,\eta^*,D^*)$. Kernel smoothing an be applied to these draws to estimate the density at the point β^*.

Given these estimates, the marginal likelihood is estimated as

$$\ln \hat{m}(y) = \ln L(y|\theta^*) + \ln \pi(\theta^*) - \left(\ln \hat{\pi}(D^{-1*}|b,\eta) + \ln \hat{\pi}(\eta|y,D^*) + \ln \hat{\pi}(\beta^*|y,\eta^*,D^*)\right).$$

The numerical standard error of this estimate may be derived.

3.3 Computation of modal estimates

We now turn to the question of finding the modal estimate, which, along with the posterior mean, may serve as θ^* for the marginal likelihood calculation. We are interested in the ML estimate because it provides (i) an approximate summary of the posterior density; (ii) an input into the AIC or BIC model information functions, and (iii) a starting point for the full MCMC iterations.

The E-M algorithm [Dempster, Laird, and Rubin (1987)] requires the recursive implementation of two steps: the expectation or E-step and the maximization or M-step. In the E-step, given the current guess of the maximizer $\theta^{(j)} = (\beta^{(j)}, \eta^{(j)}, D^{(j)})$, one computes

$$\begin{aligned} Q(\theta^{(j)},\theta) &= \int \ln\{f(y,b|\beta)\}\,\pi(b|y,\theta^{(j)})\,db \\ &= \int \{\sum_{i=1}^{n} [\ln \Pr(y_i|\theta,b_i) + \ln \phi(b_i|\eta,D)]\}\,\pi(b|y,\theta^{(j)})\,db, \end{aligned} \tag{16}$$

which is the expectation of the log of the complete data density with respect to the conditional density of b_i given the data and the current guess of the maximizer $\theta^{(j)}$. Although the Q function cannot be calculated in closed form, it can be estimated by Monte Carlo as suggested by Wei and Tanner (1990). Let $\{b^{(1)},\ldots,b^{(K)}\}$, where $b^{(j)} \sim [b|y,\theta^{(j)}]$, be a sample obtained by one of the methods discussed in Section 2. Wei and Tanner (1990) recommend that K depend on j - a small value of K is used at the start of the iterations and increased as the maximizer is approached. Then

$$\hat{Q}(\theta^{(j)},\theta) = K^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n}\left\{\ln \Pr(y_i|\beta,b_i^{(k)}) + \ln \phi(b_i^{(k)}|D)\right\} \tag{17}$$

is an ergodic average that, under regularity conditions, converges to Q as $K \to \infty$. (The Q function may also be estimated from a (synthetically) independent sample constructed by using every lth draw of the sequence $\{b^{(1)},\ldots,b^{(K)}\}$.) In the M-step, the $\hat{Q}$ function is maximized to obtain a revised guess of the maximizer $\theta^{(j+1)}$, i.e.,

$$\theta^{(j+1)} = \arg\max_{\theta} \hat{Q}(\theta^{(j)},\theta).$$

This maximization is accomplished in a sequence of two conditional maximization steps:

- Given the current value of D, $\hat{Q}(\theta^{(j)},\theta)$ is maximized over β and η to produce $\beta^{(j+1)}$ and $\eta^{(j+1)}$. The latter is seen to be $\eta^{(j+1)} = (nK)^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n} b_i^{(k)}$ (the sample mean of all the draws), whereas $\beta^{(j+1)}$ is obtained by the Newton-Raphson method applied to the function $K^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n}\ln \Pr(y_i|\beta,b_i^{(k)})$. The gradient and Hessian for the N-R algorithm, similar to those of Section 3, are given by

$$K^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(y_{it} - \exp(x_{it}'\beta + w_{it}'b_i^{(k)})\right)x_{it}$$

and

$$-K^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(\exp(x_{it}'\beta + w_{it}'b_i^{(k)})\right)x_{it}x_{it}',$$

respectively.

- Given $\beta^{(j+1)}$ and $\eta^{(j+1)}$, the random effects $\{b_i\}$ are drawn from $\pi(b|y, \eta^{(j+1)}, D^{(j)})$, and the update of D is obtained from the revised $\hat{Q}$ function

$$D^{(j+1)} = (nK)^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n} \left(b_i^{(k)} - \eta^{(j+1)}\right) \left(b_i^{(k)} - \eta^{(j+1)}\right)',$$

which is found by equating to zero the derivative of $\hat{Q}$ with respect to D.

The calculation of $\hat{Q}$ and the maximization over θ are terminated when the change in successive parameter values is sufficiently small. The value θ^* at the end of these iterations is the maximum likelihood estimate. Standard errors of the estimate θ^* can be obtained from Louis (1982), where it is shown that the observed information matrix (the negative of $\frac{\partial^2 l}{\partial\theta\partial\theta'}$) is given by

$$-E\left[\frac{\partial^2 \ln\{f(y,b|\theta)}{\partial\theta\partial\theta'}\right] - \mathrm{Var}\left[\frac{\partial \ln\{f(y,b|\theta)}{\partial\theta}\right];$$

the expectation and variance are taken with respect to $[b|y,\theta^*]$. Although direct evaluation is not feasible, each of these terms can be estimated by using the M-H step to produce a sample $\{b^{(1)},\ldots,b^{(J)}\}$, where $b^{(j)} \sim [b|y,\theta^*]$. The observed information matrix is estimated as

$$-J^{-1} \sum_{k=1}^{J} \frac{\partial^2 \ln\{f(y,b^{(k)}|\theta^*)}{\partial\theta\partial\theta'} - J^{-1} \sum_{k=1}^{J} \left(\frac{\partial \ln\{f(y,b^{(k)}|\theta^*)}{\partial\theta} - m\right)\left(\frac{\partial \ln\{f(y,b^{(k)}|\theta^*)}{\partial\theta} - m\right)' \tag{18}$$

where $m = J^{-1}\sum_{k=1}^{J} \frac{\partial \ln\{f(y,b^{(k)}|\hat{\theta})}{\partial\theta}$. The derivatives of $\ln f(y,b^{(k)}|\theta^*)$ can be computed analytically or by numerical differentiation via a packaged routine. The relevant standard errors are given by the square root of the diagonal elements of the inverse of the estimated information matrix.

The constants K and J can be chosen pragmatically, motivated by the speed of the computing environment and the accuracy desired. We allow K to increase gradually as a function of the iterations and, near the mode, usually set K to be about 1000, which appears to be satisfactory for most problems. A value of J about 5000 has been found to be adequate.

A variant of the MCEM algorithm can also be used to find θ^*. This variant, called the simulated EM (SEM) algorithm, is based on the algorithm of Celeux and Diebolt (1985). Unlike the MCEM algorithm, the SEM generates a Markov chain sample, and the mean of the simulated sample can be used as an estimate of the modal value. In this algorithm the evaluation of Q is replaced by a maximization of the complete data log-likelihood. Specifically, given the current sample of the random effects $b^{(k)}$, the next item in the parameter sequence is obtained by maximizing

$$\sum_{i=1}^{n} \left\{ \ln \Pr(y_i|\beta, b_i^{(k)}) + \ln \phi(b_i^{(k)}|\eta, D) \right\}$$

over the parameter space. The random effects are then simulated as described above, and the process is iterated. From the MLE and $\ln \hat{L}(y|\theta^*)$ (through the approach in Section 2.1), it is possible to compute the AIC and BIC information functions: Subtracting the penalty $2n^{-1}\dim(\theta)$ from $\ln \hat{f}(y|\theta^*)$ produces the AIC, and subtracting the penalty $n^{-1}\dim(\theta)\log(n)$ produces the BIC.

4 Examples

We next present three applications of the methods developed above to count data. The first is to data on treatment for epilepsy, the second to the patent data, and the third to workplace absences.

4.1 Epilepsy data

Diggle, Liang, and Zeger (1995) consider the data on four successive two-week seizure counts (y_{ij}) for each of 59 epileptics ($i = 1, \ldots, 59; j = 0, \ldots, 4$), some of whom are treated with progabide (observation 49 is eliminated from the computations because of the "unusual pre- and post-randomization seizure counts"). The covariates are

$$x_{i1} = \begin{cases} 1 & \text{if treatment group} \\ 0 & \text{if control} \end{cases} ; \; x_{i2} = \begin{cases} 1 & \text{if visit 1, 2, 3 or 4} \\ 0 & \text{if baseline} \end{cases} ;$$

Obs	y_{i1}	y_{i2}	y_{i3}	y_{i4}	Treat	Base	Obs	y_{i1}	y_{i2}	y_{i3}	y_{i4}	Treat	Base
1	5	3	3	3	0	11	31	0	4	3	0	1	19
2	3	5	3	3	0	11	32	3	6	1	3	1	10
3	2	4	0	5	0	6	33	2	6	7	4	1	19
4	4	4	1	4	0	8	34	4	3	1	3	1	24
5	7	18	9	21	0	66	35	22	17	19	16	1	31
6	5	2	8	7	0	27	36	5	4	7	4	1	14
7	6	4	0	2	0	12	37	2	4	0	4	1	11
8	40	20	23	12	0	52	38	3	7	7	7	1	67
9	5	6	6	5	0	23	39	4	18	2	5	1	41
10	14	13	6	0	0	10	40	2	1	1	0	1	7
11	26	12	6	22	0	52	41	0	2	4	0	1	22
12	12	6	8	5	0	33	42	5	4	0	3	1	13
13	4	4	6	2	0	18	43	11	14	25	15	1	46
14	7	9	12	14	0	42	44	10	5	3	8	1	36
15	16	24	10	9	0	87	45	19	7	6	7	1	38
16	11	0	0	5	0	50	46	1	1	2	4	1	7
17	0	0	3	3	0	18	47	6	10	8	8	1	36
18	37	29	28	29	0	111	48	2	1	0	0	1	11
19	3	5	2	5	0	18	49	102	65	72	63	1	151
20	3	0	6	7	0	20	50	4	3	2	4	1	22
21	3	4	3	4	0	12	51	8	6	5	7	1	42
22	3	4	3	4	0	9	52	1	3	1	5	1	32
23	2	3	3	5	0	17	53	18	11	28	13	1	56
24	8	12	2	8	0	28	54	6	3	4	0	1	24
25	18	24	76	25	0	55	55	3	5	4	3	1	16
26	2	1	2	1	0	9	56	1	23	19	8	1	22
27	3	1	4	2	0	10	57	2	3	0	1	1	25
28	13	15	13	12	0	47	58	0	0	0	0	1	13
29	11	14	9	8	1	76	59	1	4	3	2	1	12
30	8	7	9	4	1	38							

Table 1: Epilepsy data

and t_{ij} (the offset term) which equals 8 if $j = 0$ and 2 if $j = 1, 2, 3$, or 4. The complete data set appears in Table 1. Following Diggle, Liang, and Zeger, we model the counts by a Poisson link with mean

$$\log E(y_{ij}|\beta, b_i) = \log t_{ij} + \beta_1 + \beta_2 x_{ij1} + \beta_3 x_{ij2} + \beta_4 x_{ij1} x_{ij2} + b_{i1} + b_{i2} x_{ij1}.$$

The intercept and x_{i1} (time) variables are thus treated as random effects.

We specify the following vague priors on β, η and D:

$$\beta \sim N_2(0, 10^{-2} \times I),\ \eta \sim N_2(0, 10^{-2} \times I),\ D^{-1} \sim W(4, I)$$

and experiment with the four alternative proposal generating densities for b discussed in

Section 2. Any tuning constants in these methods (such as τ_1 and τ_2) are obtained by short, preliminary runs, by focusing on the acceptance rates and the serial correlations of the output. The values of these adjustable constants are included in our tabular output. The final MCMC iterations are then run for 10,000 cycles beyond a burn-in of 1000 iterations.

Table 2 contains a set of results for this data in (β, η) parameterization. The table contains the posterior mean (mean), the posterior standard deviations (s.d.), the acceptance rates in both the b_i and β steps, and the autocorrelation at lag 20 of the generated sample for each of the alternative methods. Because there are a large number of acceptance rates in the case of b_i, we report the minimum and maximum rates achieved in the sampling. We have found that this diagnostic is a useful summary of the performance of the M-H simulations given that the acceptance rate for each random effect cannot be sensibly monitored in real time.

From these results we conclude that all four methods for simulating b yield similar posterior means and standard deviations. These, in turn, are close to the maximum likelihood estimators reported in Diggle, Liang, and Zeger (1995) and to those obtained from the MCEM algorithm developed above. The posterior point estimates of D_{ij} also agree with the maximum likelihood estimates. The results indicate an important time $\times$ treatment interaction effect and substantial heterogeneity in the intercepts.

We now examine the effect of parameterization and apply each of the four methods anew after setting $\eta = 0$ and letting w_{it} be a subset of x_{it}. The prior on β in these runs is now $N_4(0, 10^{-2} \times I)$)For brevity we focus on method 4 and simulate 10,000 draws from the posterior distribution using $\tau_\beta = 1.5$ and $\tau_2 = 1.5$. We summarize the results obtained in Figure 1 for $(\beta_1, \beta_4, D_{11}, D_{22})$. The figure contains Q-Q and autocorrelation plots for output from the recommended (β, η) parameterization [second column] and from the *no* η parameterization [third column]. From these figures we can conclude that the Q-Q plots are linear, and that the chain displays generally less serial correlation in the (β, η)

	Method 1	Method 2	Method 3	Method 4
M-H const				
τ_β	1.5	1.5	1.5	1.5
τ_1	.7	n.a.	.7	n.a.
τ_2	1.5	1.5	1.5	1.5
Param				
Const	1.093 (.128)	1.076 (.134)	1.080 (.143)	1.066 (.134)
Treat	- .051 (.170)	-.023 (.180)	- .029 (.204)	- .002 (.185)
Time	.017 (.101)	.016 (.115)	.021 (.108)	.013 (.114)
Interact	- .370 (.133)	- .363 (.166)	- .373 (.147)	- .360 (.159)
D_{11}	.474 (.099)	.478 (.100)	.481 (.100)	.476 (.100)
D_{21}	.017 (.056)	.015 (.058)	.013 (.058)	.014 (.057)
D_{22}	.241 (.062)	.245 (.065)	.244 (.063)	.246 (.064)
Acf(20)				
Const	.429	.435	.395	.368
Treat	.872	.779	.804	.721
Time	.421	.276	.362	.195
Interact	.686	.471	.580	.321
D_{11}	.042	.010	.024	.024
D_{21}	.096	.017	.018	.003
D_{22}	.124	.005	.045	.012
Accpt rate				
β	.392	.401	.401	.399
b_i min	.084	.587	.187	.895
b_i max	.429	.610	.466	.911

Table 2: Epilepsy data: M-H tuning constants, posterior moments and performance summaries in the (β, η) parameterization. Results are based on $G = 10,000$ samples beyond an initial transient stage of a 1000 cycles.

parameterization.

In terms of the methods, the best results overall are obtained when the random effects are simulated by the accept-reject method with a pseudo-dominating density (Method 4) in the (β, η) formulation. It is interesting to note that even the random-walk chain for simulating the random effects (Method 1), yields point estimates that are similar to the others, although its autocorrelations are quite large. This suggests that exploratory work can be done with this rather fast approach, and final results can be computed with one of the slower, but more satisfactory, methods.

We also consider the question of model choice for these data and compute the log marginal likelihoods for the model discussed above and for an alternative model in which

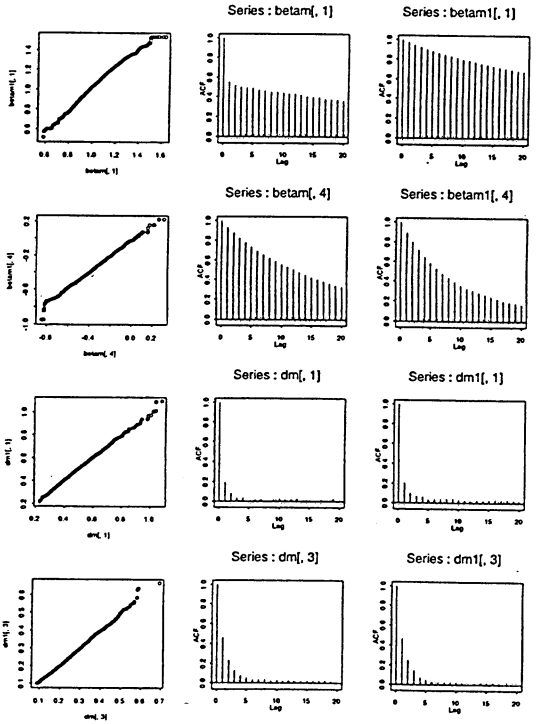

Figure 1: Diggle data. Q-Q and acf plots under alternative parameterizations: output from (β, η) is in second column.

the intercept is the only random effect. The marginal likelihoods are computed from the (β, η) parameterization. Method 4 is used to simulate the random effects. Each of the reduced MCMC iterations are run for 10,000 iterations and the marginal likelihood identity is evaluated at the maximum likelihood estimate. We obtain $\ln m(y) = -915.404$ for the former model and -969.824 for the latter. This is very strong evidence in favor of including the second random effect.

4.2 Patent data

We next illustrate the methods using data on patents. This data has previously been analyzed by Hausman, Hall, and Griliches (1984) and Blundell, Griffith and Van Reenen (1995) by classical means. The data set contains information on the research and development (R&D) expenditures of 642 firms and the number of patents received over the time period 1975–1979. Letting y_{it} denote the number of patents received by firm i in year t, the model of interest specifies that

$$\log E(y_{ij}|\beta, b_i) = \beta_1 + \beta_2 x_{ij1} + \beta_3 x_{ij2} + \beta_4 x_{ij3} + b_{i1} + b_{i2} x_{ij1},$$

where x_{ij1} is the logarithm of R&D spending ($\log R_0$) and x_{ij2} to x_{ij4} are lagged values of the logarithm of R&D spending ($\log R_{-1}, \log R_{-2}, \log R_{-3}$). The intercept and $\log R_0$ are thus treated as random effects. The model also contains time dummies for 1976–1979 but these are suppressed here and in the output for notational and visual convenience. The data set contains additional variables - a dummy variable for whether or not a firm is in a group of scientifically based industries and the inflation adjusted book value of the firm in 1971 - but these cannot be not included as covariates in the model because they exhibit no within variation and hence are indistinguishable from the random intercept.

The MCMC design and the priors for this model correspond to those discussed above. Once again we investigate the efficacy of the four methods for simulating the random effects and of the alternative parameterizations. The first set of results (based on 10,000 simulations

after dropping the first 2,000) appear in Table 3.

	Method 1	Method 2	Method 3	Method 4
M-H const				
τ_β	.7	1.0	.7	1.0
τ_1	.7	n.a.	1.0	n.a.
τ_2	1.0	2.5	1.5	2.0
Param				
constant	.776 (.075)	.772 (.077)	.747 (.076)	.733 (.076)
$\log R_0$	.694 (.030)	.697 (.040)	.621 (.035)	.572 (.036)
$\log R_{-1}$	-.043 (.031)	-.055 (.033)	.005 (.032)	.046 (.033)
$\log R_{-2}$	.128 (.036)	.130 (.036)	.138 (.038)	.144 (.037)
$\log R_{-3}$	.092 (.030)	.089 (.030)	.113 (.030)	.129 (.031)
D_{11}	2.588 (.259)	2.668 (.256)	2.594 (.248)	2.547 (.252)
D_{21}	-.578 (.072)	-.618 (.079)	-.597 (.076)	-.585 (.076)
D_{22}	.215 (.027)	.293 (.035)	.287 (.034)	.282 (.032)
Acf(20)				
Constant	.153	.026	.048	.031
$\log R_0$	.480	.322	.221	.171
$\log R_{-1}$	.186	.263	.045	.155
$\log R_{-2}$	.034	.034	-.009	.007
$\log R_{-3}$	.182	.083	.050	.042
D_{11}	.515	.117	.204	.011
D_{21}	.550	.182	.253	.019
D_{22}	.630	.290	.385	.032
Accpt rate				
β	.377	.222	.387	.233
b_i min	.015	.259	.121	.818
b_i max	.590	.291	.482	.925

Table 3: Patent data: M-H tuning constants, posterior moments and performance summaries in the (β, η) parameterization. Results are based on $G = 10,000$ samples beyond an initial transient stage of a 1000 cycles.

We find that the results are broadly consistent across methods. The magnitudes of the posterior means and standard deviations of D lead us to conclude that there is considerable variation across firms and that firms with large intercepts have a smaller effect from current R&D expenditures. Furthermore, the posterior moments of the fixed effects reveal that the effect of the first lag in log R&D is close to zero, while those from the remaining lagged values of log R&D are positive but smaller than that of current R&D.

It is also interesting to mention that this data clearly illustrates the advantages of using a MVT tailored proposal as opposed to the Gaussian tailored proposal in the generation of

the random effects. The latter proposal was found to give minimum acceptance rates of 0 and poor mixing in some cases.

Next we report on the results from the no η parameterization by fitting the above model using Method 3 (setting $\tau_\beta = .7$, $\tau_1 = 1$ and $\tau_2 = 1.5$). For simplicity we consider the parameters β_1 and β_2 (the coefficients of our two random effects, the intercept and log R_0) and compare the marginal posterior distributions of these parameters from the alternative parameterizations. We also examine the autocorrelation plots of the sampled values. The results appear in Figure 2 where the first column corresponds to the recommended parameterization. It can be seen that the marginal posterior distributions for β_1 are different but those of β_2 are roughly identical. It appears that the distribution of the intercept in the no η parameterization has not converged even after 12,000 iterations due to the high serial correlation. For each parameter, the autocorrelation patterns are much better behaved in the (β, η) parameterization. This is the kind of improvement we expected given the pattern of heterogeneity in the data. A more extensive experiment with the other methods gave similar results.

Finally, we note that method 3 (which appears to inherit the strengths of method 2 without the drawbacks of method 1) gives results that are comparable to the more sophisticated method 4. This is potentially very useful because in the context of large data sets, method 3 can deliver an order of magnitude reduction in computing time.

4.3 Absence data

Our final illustration is with a data set on the number of absences from work for a random sample of 704 full-time workers in Germany covering the period 1986–1989. The data are drawn from the German Socio-Economic Panel [see Wagner, Burkhauser, and Behringer (1993)]. This is an interesting data set because, as noted by Brown and Sessions (1996), days lost due to absences can exceed those lost as a result of unemployment.

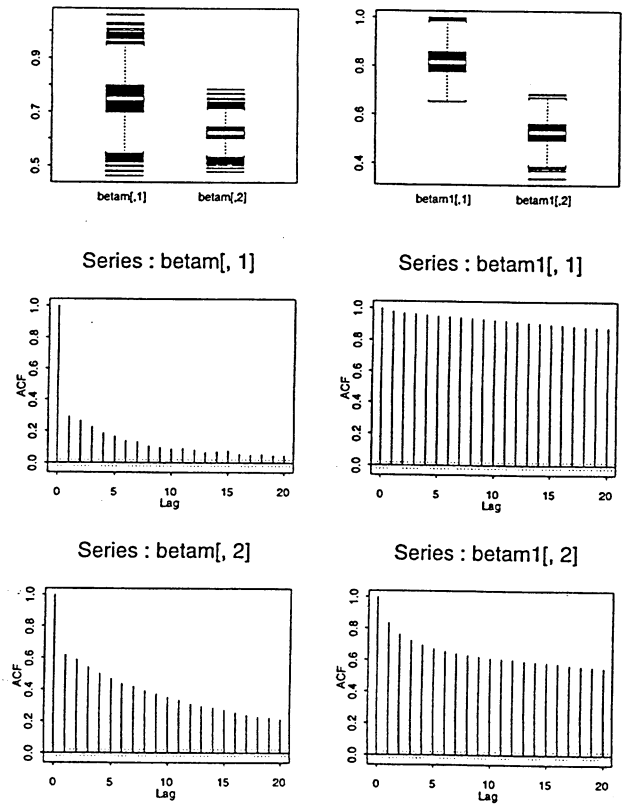

Figure 2: Patents data. Posterior densities and acf's for β_1, β_2 under alternative parameterizations: output from (β, η) is in first column.

The response variable y_{it} is the count of the number of days a worker has been absent from work during the calendar year t. In the survey, this question is asked in year $t+1$ retrospectively for year t. Some summary statistics of this data are as follows. The average number of absent days in the sample is 4.6, with a standard deviation 8.3. An important feature of the data is the high proportion of zeros: 59 percent of all observations are zero, and 23 percent of workers report no absent day in any of the four years. Both the high variability of the dependent variable and the excess of zeros suggest that the standard Poisson regression model without random effects is likely to be inappropriate. We therefore fit and compare alternative Poisson models with multiple random effects.

Potential covariates to explain the response variable include: years of job tenure in the current job; job satisfaction (an ordinal response coded $0, 1, \ldots, 10$, where 0 stands for "completely dissatisfied" and 1 stands for "completely satisfied"); the lagged number of absent days; the size of the employees' firm (1 if it is a large firm with 200 or more employees, 0 otherwise); the marital status of the worker (1 if married); the presence of children at home (1 if children are present); and the nature of the work contract (1 if limited time contract). These covariates fall into two categories. The first consists of the variables job tenure and job satisfaction that have within-variation for most workers. The second consists of the remaining variables with no within-variation for most workers. This distinction is important since it affects identification. In the presence of a random intercept, any variable with a random coefficient must have within variation for each individual in order to identify b_i. We ensure this by including only individuals for which the (4×3) matrix formed by the constant, job tenure and job satisfaction has full column rank.

We specify four models for this data. The first three models include the same set of covariates: tenure, satisfaction, and lagged absent days with a different assignment of the random effects in each case. The fourth model has job tenure and job satisfaction as the random effects and a different set of covariates. To summarize, the random effects in the

four models are specified as

- Model 1. Constant.
- Model 2. Constant, tenure.
- Model 3. Constant, satisfaction.
- Model 4. Job tenure, Job satisfaction.

For each model, we simulate the posterior density in η-parametrization, obtain the maximum likelihood estimator as a high density point using the MCEM algorithm, and then estimate the log marginal likelihood at the ML estimate.

Variable	Model 1	Model 2	Model 3	Model 4
β:				
Constant				1.116 (.092)
Job tenure	0.035 (.002)	0.055 (.001)		
Job satisfaction	-0.044 (.004)		-0.037 (.003)	
Absent days$_{t-1}$	-0.024 (.001)	-0.030 (.001)	-0.053 (.002)	-0.040 (.001)
Firm size				0.145 (.061)
Married				-0.311 (.067)
Children				-0.157 (.042)
Limited contract				-0.102 (.122)
η:				
Constant	0.047 (.008)	-0.046 (.045)	-0.106 (.051)	
Job tenure			0.100 (.034)	0.004 (.024)
Job satisfaction		-0.007 (.056)		-0.076 (.035)
D:				
D_{11}	0.038 (.003)	0.851 (.084)	1.364 (.142)	0.252 (.039)
D_{12}		-1.039 (.101)	-0.811 (.079)	-0.301 (.046)
D_{22}		1.368 (.125)	0.654 (.056)	0.569 (.082)

Table 4: Absence data: Maximum likelihood estimates and standard errors from the Markov chain Expectation Maximization algorithm. The results are the final iterate values at convergence. The standard errors are computed using $M = 1000$ random effects draws after convergence.

The prior densities and the MCMC design is again similar to that used in the earlier examples. Based on our experience from those runs, we use method 3 to simulate the random effects. To achieve a balanced D matrix, the constant term is scaled by a factor

10. We start the MCEM algorithm with $K = 4$ draws for b, and later increase K to 1000. Again, the tuning constants are adjusted in order to produce acceptance rates between 0.3 and 0.5.

Table 4 displays the results from the MCEM estimation, while Table 5 shows the posterior means, the posterior standard deviations and the log marginal likelihoods from the MCMC simulation.

Variable	Model 1	Model 2	Model 3	Model 4
β:				
Constant				1.034 (.177)
Job tenure	0.035 (.005)	0.060 (.008)		
Job satisfaction	-0.045 (.007)		-0.021 (.010)	
Absent days$_{t-1}$	-0.024 (.001)	-0.029 (.001)	-0.053 (.002)	-0.040 (.001)
Firm size				0.161 (.136)
Married				-0.188 (.108)
Children				-0.141 (.064)
Limited contract				-0.156 (.145)
η:				
Constant	0.046 (.013)	-0.060 (.046)	-0.122 (.054)	
Job tenure			0.101 (.037)	0.004 (.023)
Job satisfaction		-0.004 (.056)		-0.085 (.036)
D:				
D_{11}	0.040 (.003)	0.879 (.076)	1.384 (.120)	0.254 (.022)
D_{12}		-1.065 (.093)	-0.822 (.072)	-0.302 (.029)
D_{22}		1.401 (.120)	0.665 (.053)	0.583 (.051)
Log marginal likelihood	-11606.84	-9663.58	-9648.52	-9693.72

Table 5: Absence data: Posterior moments and marginal likelihoods. Random effects are simulated using Method 3. Results are based on $G = 10,000$ samples beyond an initial transient stage of a 2000 cycles.

We note that both maximum likelihood estimator and estimated standard errors are very similar to the posterior means and standard deviations. The number of reported absent days increases with job tenure and decreases with job satisfaction. The preferred model is Model 3 with a marginal likelihood of -9648.5; this model with random individual specific intercept is better than Model 4: the included time invariant covariates are not able to explain the between individual variation in the absence intensity.

5 Conclusions

This paper has shown how MCMC methods make possible the analysis of rather complex variants of the generalized linear model. We have discussed several different M-H based approaches for simulating the (augmented) posterior distribution. One useful approach for sampling the random effects is based on a mixture proposal density. The first component of this mixture is a random walk chain and the second is a tailored MVT density. We have found that it is important to use a MVT proposal density instead of one based on the Gaussian distribution. Another method was shown to be even more effective (though computationally more demanding). This is the M-H accept-reject algorithm with a pseudo-dominating density. The use of this method in the context of our models is quite promising. The paper also documents the value of a new parameterization that is related to the idea of hierarchical centering.

In addition, we have considered the problems of ML estimation and model choice. It is interesting that access to a MCMC random effects simulator is sufficient to find the ML estimate and the associated standard errors, due to a Monte Carlo version of the EM algorithm. Interest in this algorithm can be expected to increase. Finally, we have developed a practical methodology for the computation of marginal likelihoods and Bayes factors without constraining assumptions about the size of the clusters and number of random effects. This advance is useful and important as well.

References

Albert, J (1992), "A Bayesian analysis of a Poisson random-effects model," *American Statistician*, 46, 246–253.

Brown, S. and J.G. Sessions (1996), "The economics of absence: Theory and evidence," *Journal of Economic Surveys*, 10, 23–53.

Bennett, J. E., A. Racine-Poon, and J. C. Wakefield (1995), "MCMC for nonlinear hierarchical models," in *Markov chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson, and D. J. Spiegelhalter), London: Chapman and Hall (in press).

Blundell, R., R. Griffith, and J. Van Reenan (1995), "Dynamic count data models of technological innovation," *Economic Journal*, 105, 333–344.

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

Carlin, B and S. Chib (1995), "Bayesian Model Choice via Markov Chain Monte Carlo," *Journal of the Royal Statistical Society*, Ser B, 57, 473-484.

Celeux, G. and J. Diebolt (1985), "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem," *Computational Statistics Quarterly*, 2, 73–82.

Chib, S. (1995), "Marginal likelihood from the Gibbs output," *Journal of the American Statistical Association*, 90, 1313–1321.

Chib, S. and E. Greenberg (1995), "Understanding the Metropolis-Hastings Algorithm," *American Statistician*, 49, 327–335.

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

Diggle, P., K-Y Liang, and S.L. Zeger (1995), *Analysis of Longitudinal Data*, Oxford, Oxford University Press.

Gammerman, D (1994), "Efficient sampling from the posterior distribution in generalized linear mixed models," Technical Report, Universidade federal do Rio de Janeiro.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (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.

Geweke, J. (1989), "Bayesian inference in econometric models using Monte Carlo integration," *Econometrica*, 57, 1317–1340.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computations and Bayesian model determination," *Biometrika*, 82, 711-732.

Hausman, J. A., B. H. Hall, and Z. Griliches (1984), "Econometric models for count data with an application to the Patents-R & D relationship," *Econometrica*, 52, 909–938.

Jeffreys, H (1961), Theory of Probability (3rd edition), New York: Oxford University Press.

Kass, R.E. and A.E. Raftery (1995), "Bayes factors," *Journal of the American Statistical Association* 90, 773-795.

Lewis, S. and A.E. Raftery (1994), "Estimating Bayes Factors via Posterior Simulation with the Laplace-Metropolis Estimator," Technical Report No. 279, Department of Statistics, University of Washington.

Louis, T. A. (1982), "Finding the observed information matrix using the EM algorithm," *Journal of the Royal Statistical Society* B, 44, 226–233.

O'Hagan, A. (1994), *Kendall's Advanced Theory of Statistics* ,Volume 2B, Bayesian Inference,, New York: Halsted Press.

Press (1989), *Bayesian Statistics: Principles, Models and Applications*, New York: John Wiley.

Tanner, M. A. and W. H. Wong (1987), "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 528–549.

Tierney, L. (1991), "Markov chains for exploring posterior distributions (with discussion)," *Annals of Statistics*, 22, 1701–1762.

Tierney, L. and J. Kadane (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82-86.

Wagner, G. G., R. V. Burkhauser, and F. Behringer (1993), "The English Language Public Use File of the German Socio-Economic Panel," *Journal of Human Resources*, 28, 429–433.

Wakefield, J. C., A. F. M. Smith, A. Racine Poon, and A. E. Gelfand (1994), "Bayesian analysis of linear and non-linear population models by using the Gibbs sampler," *Applied Statistics*, 43, 201–221.

Wei, G. C. G. and M. A. Tanner, (1990), "A Monte Carlo implementation of the EM Algorithm and the Poor Man's data augmentation algorithm," *Journal of the American Statistical Association*, 85, 699–704.

Zeger, S. L. and M. R. Karim (1991), "Generalized linear models with random effects: A Gibbs sampling approach," *Journal of the American Statistical Association* 86, 79–86.

LIST OF DISCUSSION PAPERS*

No. 9201 Testing for Arch-Garch Errors in a Mis-specified Regression, by David E. A. Giles, Judith A. Giles, and Jason K. Wong.

No. 9202 Quasi Rational Consumer Demand Some Positive and Normative Surprises, by John Fountain.

No. 9203 Pre-test Estimation and Testing in Econometrics: Recent Developments, by Judith A. Giles and David E. A. Giles.

No. 9204 Optimal Immigration in a Model of Education and Growth, by K-L. Shea and A. E. Woodfield.

No. 9205 Optimal Capital Requirements for Admission of Business Immigrants in the Long Run, by K-L. Shea and A. E. Woodfield.

No. 9206 Causality, Unit Roots and Export-Led Growth: The New Zealand Experience, by David E. A. Giles, Judith A. Giles and Ewen McCann.

No. 9207 The Sampling Performance of Inequality Restricted and Pre-Test Estimators in a Mis-specified Linear Model, by Alan T. K. Wan.

No. 9208 Testing and Estimation with Seasonal Autoregressive Mis-specification, by John P. Small.

No. 9209 A Bargaining Experiment, by Michael Carter and Mark Sunderland.

No. 9210 Pre-Test Estimation in Regression Under Absolute Error Loss, by David E. A. Giles.

No. 9211 Estimation of the Regression Scale After a Pre-Test for Homoscedasticity Under Linex Loss, by Judith A. Giles and David E. A. Giles.

No. 9301 Assessing Starmer's Evidence for New Theories of Choice: A Subjectivist's Comment, by John Fountain.

No. 9302 Preliminary-Test Estimation in a Dynamnic Linear Model, by David E. A. Giles and Matthew C. Cunneen.

No. 9303 Fans, Frames and Risk Aversion: How Robust is the Common Consequence Effect? by John Fountain and Michael McCosker.

No. 9304 Pre-test Estimation of the Regression Scale Parameter with Multivariate Student-t Errors and Independent Sub-Samples, by Juston Z. Anderson and Judith A. Giles

No. 9305 The Exact Powers of Some Autocorrelation Tests When Relevant Regressors are Omitted, by J. P. Small, D. E. Giles and K. J. White.

No. 9306 The Exact Risks of Some Pre-Test and Stein-Type Regression Estimators Under Balanced Loss*, by J. A. Giles, D. E. A. Giles, and K. Ohtani.

No. 9307 The Risk Behavior of a Pre-Test Estimator in a Linear Regression Model with Possible Heteroscedasticity under the Linex Loss Function, by K. Ohtani, D. E. A. Giles and J. A. Giles.

No. 9308 Comparing Standard and Robust Serial Correlation Tests in the Presence of Garch Errors, by John P. Small.

No. 9309 Testing for Serial Independence in Error Components Models: Finite Sample Results, by John P. Small.

No. 9310 Optimal Balanced-Growth Immigration Policy for Investors and Entrepeneurs, by A. E. Woodfield and K-L. Shea.

No. 9311 Optimal Long-Run Business Immigration Under Differential Savings Functions, by A. E. Woodfield and K-L. Shea.

No. 9312 The Welfare Cost of Taxation in New Zealand Following Major Tax Reforms, by P. McKeown and A. Woodfield.

No. 9313 The Power of the Goldfeld-Quandt Test when the errors are autocorrelated, by J.P. Small and R.J. Dennis.

No. 9314 The Nucleolus Strikes Back, by M. Carter and P. Walker.

(Continued on next page)

No. 9315 The Output-Inflation Tradeoff in the United States: New evidence on the New Classical vs. New Keynesian Debate, by Alfred V. Guender

No. 9401 Insurance Market Equilibrium and the Welfare Costs of Gender-Neutral Insurance Pricing under Alternative Regulatory Regimes by Alan E. Woodfield

No. 9402 Labour Market Signalling and the Welfare Costs of Regulated Insurance Market Equilibria under Gender-neutral Pricing, by Alan E. Woodfield.

No. 9403 The New Classical Vs The New Keynesian debate On The Output – Inflation tradeoff: Evidence From Four industrialized Countries, by Alfred V. Guender

No. 9404 Yield Spreads & Real Economic Activity: The Case of New Zealand & Australia, by Alfred V. Guender and Mathias Moersch.

No. 9405 Periodic Integration & cointegration with applications to the New Zealand Aggregate Consumption Function, by Robin Harrison and Aaron Smith.

No. 9406 Linear Programming with Mathematica, by Michael Carter

No. 9407 Are People Really Risk Seeking When Facing Losses? by John Fountain, Michael McCosker & Dean Morris

No. 9501 Pricing Internet: The New Zealand Experience by Michael Carter and Graeme Guthrie

No. 9502 Long term and Short Term Aggregate Uncertainty and the Effect on Real Output, by Alfred V. Guender and Robin Young

No. 9503 Framing and Incentive Effects on Risk Attitudes When Facing Uncertain Losses, by John Fountain, Michael McCosker and Bruce Macfarlane

No. 9504 An Outline of the History of Game Theory, by Paul Walker

No. 9505 A Drunk, Her Dog and a Boyfriend: An Illustration of Multiple Cointegration and Error Correction, by Aaron Smith and Robin Harrison

No. 9506 Optimal Markup Responses to Trade Policy Reform in a Small, Open Economy: Evidence from New Zealand, by Liliana Winkelmann and Rainer Winkelmann

No. 9507 Unemployment: Where Does It Hurt? by Liliana Winkelmann and Rainer Winkelmann

No. 9508 Apprenticeship And After: Does It Really Matter? by Rainer Winkelmann

No. 9601 Another Look at Work Contracts and Absenteeism, by Rainer Winkelmann

No. 9602 Markov Chain Monte Carlo Analysis of Underreported Count Data With an Application to Worker Absenteeism, by Rainer Winkelmann

No. 9603 Count Data Models With Selectivity, by Rainer Winkelmann

No. 9604 Optimal Revenue Smoothing: The Case of New Zealand, by Alfred V. Guender and Kirdan Lees

No. 9605 Posterior Simulation and Model Choice in Longitudinal Generalized Linear Models, by Siddhartha Chib, Edward Greenberg, Rainer Winkelmann

* Copies of these Discussion Papers may be obtained for $4 (including postage, price changes occasionally) each by writing to the Secretary, Department of Economics, University of Canterbury, Christchurch, New Zealand. A list of the Discussion Papers prior to 1991 is available on request.

This paper is circulated for discussion and comments. It should not be quoted without the prior approval of the author. It reflects the views of the author who is responsible for the facts and accuracy of the data presented. Responsibility for the application of material to specific cases, however, lies with any user of the paper and no responsibility in such cases will be attributed to the author or to the University of Canterbury.