ELSEVIER

Journal of Econometrics 86 (1998) 33–54

JOURNAL OF Econometrics

# Posterior simulation and Bayes factors in panel count data models

Siddhartha Chib[a,*], Edward Greenberg[b], Rainer Winkelmann[c]

[a] *John M. Olin School of Business, Washington University, St. Louis, MO 63130, USA*
[b] *Department of Economics, Washington University, St. Louis, MO 63130, USA*
[c] *Department of Economics, University of Canterbury, Christchurch, New Zealand*

Received 1 August 1996; received in revised form 1 September 1997; accepted 22 September 1997

**Abstract**

This paper is concerned with the problems of posterior simulation and model choice for Poisson panel data models with multiple random effects. Efficient algorithms based on Markov chain Monte Carlo methods for sampling the posterior distribution are developed. A new parameterization of the random effects and fixed effects is proposed and compared with a parameterization in common use, and computation of marginal likelihoods and Bayes factors via Chib's (1995) method is also considered. The methods are illustrated with two real data applications involving large samples and multiple random effects. © 1998 Elsevier Science B.V. All rights reserved.

*JEL classification:* C1; C4

*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 estimation and model comparison for panel count data models with multiple random effects. Although we focus on count data, much of our discussion is also relevant for binary data and the class of generalized linear models. We are interested in procedures that allow for the efficient estimation of such models (for which the likelihood function is usually not available) and methods that can be used to compare alternative, potentially non-nested, models. A growing literature has recently begun to

* Corresponding author. E-mail: chib@simon.wustl.edu


*PII* S0304-4076(97)00108-5

address these problems from various numerical perspectives, primarily organized around Markov chain Monte Carlo algorithms (Albert, 1992; Bennett et al., 1996; Gamerman, 1994; Wakefield et al., 1994; Zeger and Karim, 1991). It has also become understood that certain identification problems can severely compromise the performance of the existing simulation methods. Gelfand et al. (1996) discuss one approach for dealing with this problem, but this approach does not appear to be computationally straightforward for the models we study.

This paper advances the existing literature in three important directions. First, we propose a parameterization of the model, related to that in Gelfand et al. (1996), that tackles the identification problem and is simple to implement. In our parameterization the covariate matrices for the fixed effects and the random effects are completely distinct, and the random effects have a non-zero mean. Second, we develop efficient algorithms based on Markov chain Monte Carlo methods for sampling the posterior distribution of the parameters and the random effects. These algorithms in conjunction with the proposed parameterization of the model, provide a substantial improvement over existing methods for sampling the posterior distribution. Finally, we develop an approach for computing Bayes factors (Kass and Raftery, 1995) for alternative panel count models that requires only the simulation routines for sampling the posterior distribution. This approach is based on the work of Chib (1995) and is easy to apply compared to alternative methods described by Carlin and Chib (1995) and Green (1995). As far as we are aware, Bayes factors for panel count data models have not been computed before.

The rest of the paper is organized as follows. In Section 2 we discuss the simulation of the posterior distribution and consider several different routines, each defined by a particular choice of proposal density in the Metropolis–Hastings step. In Section 3 we show how the marginal likelihood may be computed from the Markov chain Monte Carlo output. This section also takes up the calculation of the maximum likelihood estimate through a modification of the Monte Carlo EM (MCEM) algorithm of Wei and Tanner (1991) and the computation of the likelihood function by importance sampling. In Section 4 we consider two applications of the techniques, first to data on the effects of the drug progabide on epileptic patients, and then to patent data on a longitudinal sample of 680 firms in the United States. Concluding remarks appear in Section 5 followed by an Appendix.

## 2. Markov chain Monte Carlo sampling methods

### *2.1. The model and new parameterization*

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

$\beta \in \Re^k$ and random effects $b_i \in \Re^q$ the counts are Poisson, i.e.,

$$y_{it}|\beta, b_i \sim \text{Poisson}(\mu_{it}),$$

where $\mu_{it}$ is the conditional mean

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

$$b_i \sim N_q(\eta, D),$$

the covariates $x_{it}$ and $w_{it}$ contain *no* variables in common, and $N_q$ denotes the $q$-variate normal distribution.

Our parameterization is characterized by two new features: the non-zero mean vector $\eta$ for the random effects and the specification of the covariates $x_{it}$ and $w_{it}$ that are not allowed to have common variables. In previous formulations, $w_{it}$ is a subset of $x_{it}$ and $E(b_i) = 0$ (for example, Laird and Ware, 1982). This parameterization is not recommended in the context of the Markov chain Monte Carlo methods we propose, because of an identification problem. 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 $\mu_{it} = x_{itk}(\beta_k + b_{ik}) + A_{itk}$. But the first term is observationally equivalent to $b_{ik}x_{itk}$, implying that $\beta_k$ is not likelihood identified (O'Hagan, 1994). Identification must therefore be achieved entirely through the *prior* distribution of $b_i$. As a result, if the data contain considerable heterogeneity leading to a large variance $D$, then any Markov chain Monte Carlo algorithm that simulates both $\beta$ and $b_i$ will not mix well. Transferring the 'common' effect of $x_k$ to $\eta_k$ removes the nonidentified parameter $\beta_k$. Our parameterization is related to, but different from, the hierarchical centering introduced by Gelfand et al. (1996).

We complete the model by assuming that the parameters $(\beta, \eta, D)$ follow the prior distributions

$$\beta \sim N(\beta_0, B_0^{-1}), \quad \eta \sim N(\eta_0, M_0^{-1}), \quad D^{-1} \sim \text{Wish}(v_0, R_0),$$

where $(\beta_0, B_0, \eta_0, M_0, v_0, R_0)$ are known hyperparameters and $\text{Wish}(v_0, R_0)$ is the Wishart distribution with $v_0$ degrees of freedom and scale matrix $R_0$ (Press, 1989). These distributions are flexible in representing various prior beliefs about the parameters.

### 2.2. Likelihood function

Computational algorithms for estimation are needed because the likelihood function of this model is complicated and intractable. The likelihood function may be expressed formally as follows. Let $y_i = (y_{i1}, \ldots, y_{iT_i})$ denote the observations on the $i$th cluster. Under conditional independence

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

and

$$f(y_i, b_i|\beta, \eta, D) = f(y_i|\beta, b_i)\,\phi(b_i|\eta, D),$$

is the joint density of $(y_i, b_i)$, where $p$ is the Poisson mass function with conditional mean $\mu_{it}$ and $\phi(b_i|\eta, D) \propto |D|^{-1/2}\exp\{-0.5(b_i-\eta)'D^{-1}(b_i-\eta)\}$ is the density of the normal distribution with mean $\eta$ and covariance $D$. The likelihood function of the parameters given $y = (y_1, \ldots, y_n)$ is then given by

$$L(y|\beta, \eta, D) = \prod_{i=1}^{n} \int f(y_i, b_i|\beta, \eta, D)\,\mathrm{d}b_i \equiv \prod_{i=1}^{n} L_i(y_i|\beta, \eta, D), \tag{1}$$

which is the product of the $n$ likelihood contributions $L_i(y_i|\beta, \eta, D)$. The intractability of the likelihood function arises from the integral in Eq. (1).

### 2.3. Sampling the random effects

To develop an operational Markov chain Monte Carlo scheme for simulating the posterior distribution it is necessary to include the random effects in the simulation, an example of data augmentation (Tanner and Wong, 1987). The Markov chain Monte Carlo algorithm is then based on the blocks $b = (b_1, b_2, \ldots, b_n)$, $\beta$, $\eta$, and $D$, and the associated full conditional distributions

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

Given an arbitrary starting point in the parameter space, the conditioning variables are set at their most recent simulated values, and the distributions are sampled recursively a large number of times. Under regularity conditions that are satisfied in this problem, the Markov chain produced by these iterations can be shown to converge to the posterior distribution.

Consider the sampling of the $n$ random effects $b_i$ from the distribution $\pi(b|y,\beta, \eta, D) = \prod_{i=1}^{n}\pi(b_i|y_i, \beta, \eta, D)$, where

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

This density is difficult to simulate by standard rejection-based methods but is amenable to analysis via the versatile Metropolis–Hastings algorithm (Tierney, 1994; Chib and Greenberg, 1995). As a review, recall that, for a given target density $f(\psi)$, the Metropolis–Hastings algorithm is implemented as follows:

1. Given the current value $\psi$, sample a proposal value $\psi^{\dagger}$ from the density $q(\psi, \psi^{\dagger})$.

2. Move to the value $\psi^\dagger$ with probability $\alpha(\psi,\psi^\dagger)$ and stay at the value $\psi$ with probability $1-\alpha(\psi,\psi^\dagger)$, where

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

The behavior of this algorithm (in terms of how well the support of the target density is traversed) depends critically on the choice of $q$. One interesting aspect of our current application is that the Metropolis–Hastings algorithm must be applied to each of the target densities $\pi(b_i|y_i,\beta,\eta,D)$. Monitoring each of these Metropolis-Hastings chains is a difficult matter when one has a large number of clusters, as in the examples below. For this reason, it is extremely important that we identify proposal densities that work well on general grounds and require limited monitoring. The suggestions are compared systematically in the examples.

*Method 1: Random walk proposal.* In this method one defines $q_1(b_i,b_i^\dagger)=\phi(b_i^\dagger|b_i,\tau_1 D)$, where $\tau_1$ 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.* A second approach is to tailor the proposal density to the target density around its modal value $\hat{b}_i=\text{argmax}_{b_i}\ln f(y_i,b_i|\beta,\eta,D)$. The mode is obtained from the Newton–Raphson algorithm using the gradient vector

$$g_{b_i}=-D^{-1}(b_i-\eta)+\sum_{t=1}^{T_i}(y_{it}-\exp(x_{it}'\beta+w_{it}'b_i))w_{it} \tag{3}$$

and Hessian matrix

$$H_{b_i}=-D^{-1}-\sum_{t=1}^{T_i}(\exp(x_{it}'\beta+w_{it}'b_i))w_{it}w_{it}'. \tag{4}$$

Now define

$$q_2=f_T(b_i|\hat{b}_i,\tau_2 V_{b_i},\nu)$$
$$\propto|\tau_2 V_{b_i}|^{-1/2}\left\{1+\frac{1}{\nu}(b_i-\hat{b}_i)'(\tau_2 V_{b_i})^{-1}(b_i-\hat{b}_i)\right\}^{-(\nu+q)/2},$$

where $\tau_2$ and $\nu$ are adjustable constants and $f_T(\cdot|\hat{b}_i,\tau_2 V_{b_i},\nu)$ is the multivariate-$t$ density with $\nu$ degrees of freedom, location parameter $\hat{b}_i$ and scale matrix $\tau_2 V_{b_i}$. We set $V_{b_i}=(-H_{b_i})^{-1}$, which is the negative inverse of the Hessian of $\ln f(y_i,b_i|\beta,\eta,D)$ at the mode. We avoid the use of a similarly matched normal proposal density $\phi(b_i|\hat{b}_i,c\,V_{b_i})$ because the MV$t$ proposal density is more flexible and easier to adjust (because of the extra tuning parameter $\nu$). Furthermore, in our empirical examples, proposal values generated from the normal proposal produce some extreme acceptance rates even after considerable tuning.

*Method 3: Mixture proposal–tailored proposal.* In this method the proposal values are drawn from a mixture of proposal densities $q_1$ and $q_2$. One can select $q_2$ less frequently than $q_1$ to minimize the set-up time within each cycle. The Markov property of the simulation is preserved if the respective densities are selected at fixed, pre-specified intervals. This is a computationally inexpensive way of producing satisfactory proposal values.

*Method 4: Acceptance–rejection with tailored proposal.* In this approach the proposal value is obtained by an acceptance–rejection procedure applied to the pseudo-dominating function $c_i f_T(b_i|\hat{b}_i, \tau_2 V_{b_i}, v)$, where $c_i$ is a positive number (its choice is discussed below). Note that we have again utilized the MV$t$ distribution rather than the multivariate normal. Let $b_i^\dagger$ be a value generated from $f_T(b_i|\hat{b}_i, \tau_2 V_{b_i}, v)$ that satisfies the condition

$$u \leqslant f(y_i, b_i^\dagger|\beta, \eta, D)/c_i f_T(b_i^\dagger|\hat{b}_i, \tau_2 V_{b_i}, v),$$

where $u \sim \text{Unif}(0,1)$. Let $C_1 = I[f(y_i, b_i|\beta, \eta, D) \leqslant c_i f_T(b_i|\hat{b}_i, \tau_2 V_{b_i}, v)]$ be an indicator of whether the proposal density dominates the target at the current value $b_i$, and let $C_2 = I[f(y_i, b_i^\dagger|\beta, D) \leqslant c_i f_T(b_i^\dagger|\hat{b}_i, \tau_2 V_{b_i}, v)]$ be an indicator of domination at the proposal value $b_i^\dagger$. Then the probability of move (see, Chib and Greenberg, 1995, p. 332) is defined as

(a) $\alpha(b_i, b_i^\dagger) = 1$ if $C_1 = 1$;
(b) $\alpha(b_i, b_i^\dagger) = c_i f_T(b_i|\hat{b}_i, \tau_2 V_{b_i}, v)/f(y_i, b_i|\beta, D)$ if $C_1 = 0$ and $C_2 = 1$;
(c) $\alpha(b_i, b_i^\dagger) = \min\{f(y_i, b_i^\dagger|\beta, D) f_T(b_i|\hat{b}_i, \tau_2 V_{b_i}, v)/[f(y_i, b_i|\beta, D) f_T(b_i^\dagger|\hat{b}_i, \tau_2 V_{b_i}, v)], 1\}$ if $C_1 = 0$ and $C_2 = 0$.

*Remark.* The quantity $c_i$ used above (the value of $v$ is fixed at 15 in the examples) is determined as follows:

$$c_i = \frac{0.6 \times f(y_i, \hat{b}_i|\beta, \eta, D)}{f_T(\hat{b}_i|\eta, D, v)},$$

which can be explained in the following way. The term $f(y_i, \hat{b}_i|\beta, \eta, D)/f_T(\hat{b}_i|\eta, D, v)$ forces the ordinates of the pseudo-dominating density and the (unnormalized) target density to agree at the mode $\hat{b}_i$. The factor 0.6 (other values might be tried) lowers 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.

### 2.4. Sampling $\beta$, $\eta$, and $D$

Given the random effects, the remaining simulations are quite straightforward, with both $\eta$ and $D$ being simulated from standard distributions. For $\beta$, the sampling requires the use of a Metropolis–Hastings algorithm with an easily

constructed (tailored) proposal density. The target density is proportional to

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

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 latter steps are implemented via the gradient vector

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

and Hessian matrix

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

Analogous to the case of $b_i$ above, one can now define a tailored MV$t$ density for generating proposal values. The density we actually use is further refined so that tailored proposal values that are relatively distant from the current point can be obtained. We do this by reflecting the current value around $\hat{\beta}$ before adding an MV$t$ increment with zero mean and scale matrix $\tau_\beta V_\beta$. The resulting proposal density is given by $q(\beta, \beta^\dagger) = f_T(\beta^\dagger|\hat{\beta} - (\beta - \hat{\beta}), \tau_\beta V_\beta, v)$, which is symmetric in $(\beta, \beta^\dagger)$. There is no need to use the mixture proposal density in this case because the computational burden of finding the tailored density is minimal.

One cycle of the Markov chain Monte Carlo simulation is completed by sampling $\eta$ from

$$\pi(\eta|b, D) = \phi(\eta|\hat{\eta}, M_1^{-1}), \tag{5}$$

where

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

and $D^{-1}$ from

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

where $f_W(\cdot|a, A)$ denotes a Wishart density with $a$ degrees of freedom and scale matrix $A$.

## 3. Marginal likelihood by Markov chain Monte Carlo

From a practical viewpoint, the problem of model choice is one of the most important in fitting panel count data models and similar generalized linear

models. We now show how this problem can be tackled with the posterior simulation 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 Markov chain Monte Carlo output. The marginal likelihood of a given model is the integral of the likelihood with respect to the prior density of the parameters, i.e.,

$$m(y) = \int L(y|\beta, \eta)\pi(\beta, \eta, D)\,\mathrm{d}\beta\,\mathrm{d}\eta\,\mathrm{d}D. \tag{6}$$

On the basis of the marginal likelihood one may compute the Bayes factor (Jeffreys, 1961) in favor of model $\mathscr{M}_k$ (and against model $\mathscr{M}_l$) as

$$B_{k,l} = \frac{m(y|\mathscr{M}_k)}{m(y|\mathscr{M}_l)}. \tag{7}$$

We adopt an approach due to Chib (1995) for computing the model marginal likelihood. First, for some arbitary point $\theta^*$ we note that $m(y)$ can be written as

$$m(y) = \frac{L(y|\theta^*)\pi(\theta^*)}{\pi(\theta^*|y)}. \tag{8}$$

Second, we estimate the posterior ordinate at the point $\theta^*$ as

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

Our estimate of the marginal likelihood on the log scale is then given by

$$\ln \hat{m}(y) = \ln L(y|\theta^*) + \ln \pi(\theta^*) - \Big(\ln \hat{\pi}(D^{-1}{}^*|b, \eta) + \ln \hat{\pi}(\eta|y, D^*) + \ln \hat{\pi}(\beta^*|y, \eta^*, D^*)\Big). \tag{9}$$

Before we discuss how each of the quantities in this expression is obtained we mention that we take the point $\theta^*$ in this calculation to be either the posterior mean or the maximum likelihood estimate.

### 3.1. Likelihood function

We begin with the computation of the likelihood function at the point $\theta^*$. The contribution of $y_i$ to the likelihood at the point $\theta^*$ is

$$L_i(y_i|\theta^*) = \int f(y_i|b_i, \beta^*)\,\phi(b_i|\eta^*, D^*)\,\mathrm{d}b_i, \tag{10}$$

where the normalizing constants for both of the functions that appear under the integral are known and we have suppressed the model indicator $\mathscr{M}$. 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.

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

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

where $b_i^{(j)}$ $(j = 1, \ldots, M)$ are i.i.d. draws from $g(b_i)$. A convenient choice for the latter is $f_T(\cdot|\hat{b}_i, (-H_{b_i})^{-1}, \nu)$. The log-likelihood function is obtained by adding the $\ln \hat{L}_i(y_i|\theta^*)$.

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

We now consider the estimation of each of the three posterior ordinates that appear in the marginal likelihood expression. These ordinates can be estimated from suitably constructed Markov chain samplers via the following steps.

First, use the draws $\{b^{(g)}, \eta^{(g)}\}$ from the initial run consisting of the distributions $[\beta|y, b]$, $[b|y, \beta, \eta, D]$, $[\eta|b, D]$, and $[D^{-1}|\eta, b]$ to form the estimate

$$\hat{\pi}(D^{-1}*|y) = c\, G^{-1} \sum_{g=1}^{G} \frac{|D^{-1}*|^{(n+\nu_0-q-1)/2}}{|R_n^{(g)}|^{(n+\nu_0)/2}} \exp\{0.5\, \mathrm{tr}(R_n^{(g)-1} D^{-1}*)\},$$

where $R_n^{(g)} = [R_0^{-1} + \sum_{i=1}^{n} (b_i^{(g)} - \eta^{(g)})(b_i^{(g)} - \eta^{(g)})']^{-1}$ and $c$ is the normalizing constant of the Wishart density (see, Press, 1982, p. 108).

Second, continue the sampling with the reduced set of distributions $[\beta|y, b]$, $[b|y, \beta, \eta, D^*]$, and $[\eta|b, D^*]$, where $D$ is set equal to $D^*$, and use the draws of $\{b^{(g)}\}$ from this run to form the estimate

$$\hat{\pi}(\eta^*|y, D^*) = G^{-1} \sum_{g=1}^{G} \phi(\eta^*|\hat{\eta}^{(g)}, M_1^{*-1}),$$

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

Finally, use the draws $\{\beta^{(g)}\}$ from the final reduced Markov chain run involving $[\beta|y, b]$ and $[b|y, \beta, \eta^*, D^*]$ to form the Gaussian kernel estimate

$$\hat{\pi}(\beta^*|y, \eta^*, D^*) = G^{-1} \sum_{g=1}^{G} \phi(\beta^*|\beta^{(g)}, H),$$

where $H = \mathrm{diag}(h_1, \ldots, h_k)$ is a diagonal window-width matrix.

The accuracy of the posterior ordinate estimate can be gauged by calculating the numerical standard error, as discussed in the appendix. The numerical standard error is generally small if $G$ is large.

### 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 $\theta^*$ for the marginal likelihood calculation. The ML estimate may also serve as a starting point for the full Markov chain Monte Carlo iterations.

The EM algorithm (Dempster et al., 1977) 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

$$Q(\theta^{(j)}, \theta) = \int \sum_{i=1}^{n} [\ln \Pr(y_i|\beta, b_i) + \ln \phi(b_i|\eta, D)]\pi(b|y, \theta^{(j)})\, db. \tag{11}$$

Although $Q$ 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 $K$ (about 1,000) is used at the start of the iterations and increased (to about 5,000) as the maximizer is approached. Then

$$\hat{Q}(\theta^{(j)}, \theta) = K^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n} \{\ln \Pr(y_i|\beta, b_i^{(k)}) + \ln \phi(b_i^{(k)}|\eta, D)\} \tag{12}$$

is an ergodic average that, under regularity conditions, converges to $Q$ as $K \to \infty$. 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 two conditional maximization steps:

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

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

and

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

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 as

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

The calculation of $\hat{Q}$ and the maximization over $\theta$ are terminated when the change in successive parameter values is sufficiently small. Standard errors of the estimate $\theta^*$ can be obtained from a Monte Carlo version of Louis's (1982) formula for the information matrix

$$-J^{-1} \sum_{k=1}^{J} \frac{\partial^2 \ln f(y, b^{(k)}|\theta^*)}{\partial\theta\partial\theta'} - J^{-1} \sum_{k=1}^{J} \left(\frac{\partial \ln f(y, b^{(k)}|\theta^*)}{\partial\theta} - m\right) \times \left(\frac{\partial \ln f(y, b^{(k)}|\theta^*)}{\partial\theta} - m\right)', \tag{13}$$

where $m = J^{-1}\sum_{k=1}^{J} \partial \ln f(y, b^{(k)}|\hat{\theta})/\partial\theta$ and $b^{(j)} \sim [b|y, \theta^*]$.

## 4. Examples

We next present two applications of the methods developed above to count data. The first is to data on a treatment for epilepsy and the second to patent data.

### 4.1. Epilepsy data

Diggle et al. (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 complete data set appears in Table 1. The covariates are

$$x_{ij1} = \begin{cases} 1 & \text{if treatment group,} \\ 0 & \text{if control,} \end{cases} \qquad x_{ij2} = w_{ij1} = \begin{cases} 1 & \text{if visit } j = 1, 2, 3, \text{ or } 4, \\ 0 & \text{if baseline} \end{cases}$$

and $t_{ij}$ (the offset term), which equals 8 if $j = 0$ and 2 if $j = 1$, 2, 3, or 4. Following Diggle et al., we model the counts by a Poisson link. In the $(\beta, \eta)$-parameterization, we let

$$\log E(y_{ij}|\beta, b_i) = \log t_{ij} + \beta_2 x_{ij1} + \beta_4 x_{ij1} x_{ij2} + b_{i1} + b_{i2} w_{ij1}, \; b_i \sim N_2(\eta, D)$$

Table 1
Epilepsy data

| 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 | | | | | | | |

and in the $\beta$-parameterization

$$\log E(y_{ij}|\beta, b_i) = \log t_{ij} + \beta_1 + \beta_2 x_{ij1} + \beta_3 x_{ij2} + \beta_4 x_{ij1} x_{ij2} + b_{i1} + b_{i2} w_{ij1},\ b_i \sim N_2(0, D).$$

Thus, $\eta$ corresponds to $(\beta_1, \beta_3)$ since the intercept and $x_{ij2}$ (time) variables are random effects.

Focusing first on the $(\beta, \eta)$-parameterization, we experiment with the four alternative proposal generating densities for $b$ discussed in Section 2 under the following vague priors for $\beta$, $\eta$, and $D^{-1}$:

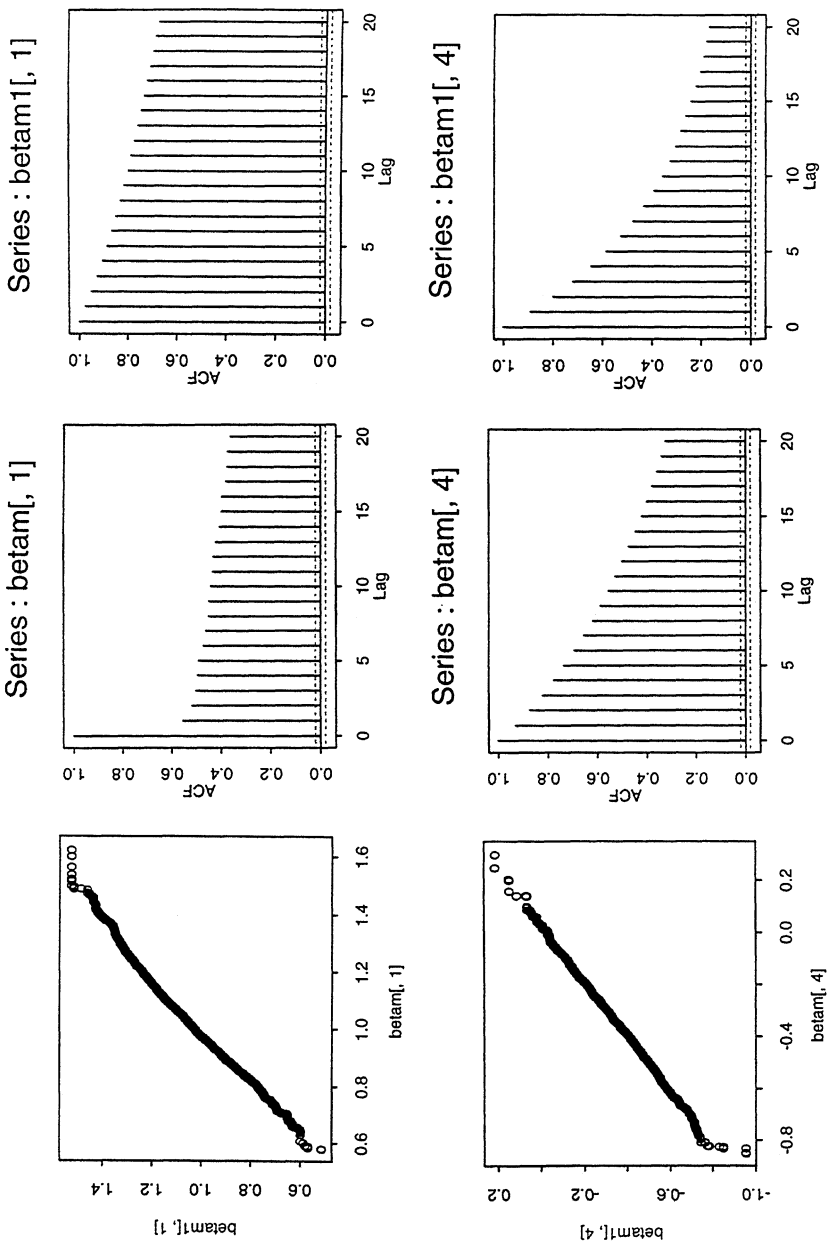
$$\beta \sim N_2(0, 10^{-2} \times I), \quad \eta \sim N_2(0, 10^{-2} \times I), \quad D^{-1} \sim W(4, I).$$

Tuning constants in these methods (such as $\tau_1$ and $\tau_2$) are obtained in short preliminary runs by examining the acceptance rates and the serial correlations of the output. The values of these adjustable constants are included in our tabular output. The final Markov chain Monte Carlo iterations are then run for 10,000 cycles beyond a burn-in of 1,000 iterations.

Table 2 contains results for these data in the $(\beta, \eta)$-parameterization. The table contains the posterior means, the posterior standard deviations, the autocorrelation at lag 20 of the generated sample, and the acceptance rates in the $b_i$ and $\beta$ steps. Because there are a large number of $b_i$, we report only the minimum and maximum acceptance rates achieved in the sampling. These are useful summaries of the performance of the Metropolis–Hastings simulations; acceptance rates for each random effect cannot be easily monitored in real time.

Table 2
Epilepsy data: M–H tuning constants, posterior means (standard deviations) and performance summaries in the $(\beta, \eta)$-parameterization. Results are based on $G = 10{,}000$ samples beyond an initial transient stage of 1,000 cycles

| | Method 1 | Method 2 | Method 3 | Method 4 |
|---|---|---|---|---|
| M–H constants | | | | |
| $\tau_\beta^{1/2}$ | 1.5 | 1.5 | 1.5 | 1.5 |
| $\tau_1^{1/2}$ | 0.7 | n.a. | 0.7 | n.a. |
| $\tau_2^{1/2}$ | 1.5 | 1.5 | 1.5 | 1.5 |
| Parameters | | | | |
| Const | 1.093 (0.128) | 1.076 (0.134) | 1.080 (0.143) | 1.066 (0.134) |
| Treat | − 0.051 (0.170) | − 0.023 (0.180) | − 0.029 (0.204) | − 0.002 (0.185) |
| Time | 0.017 (0.101) | 0.016 (0.115) | 0.021 (0.108) | 0.013 (0.114) |
| Interact | − 0.370 (0.133) | − 0.363 (0.166) | − 0.373 (0.147) | − 0.360 (0.159) |
| $D_{11}$ | 0.474 (0.099) | 0.478 (0.100) | 0.481 (0.100) | 0.476 (0.100) |
| $D_{21}$ | 0.017 (0.056) | 0.015 (0.058) | 0.013 (0.058) | 0.014 (0.057) |
| $D_{22}$ | 0.241 (0.062) | 0.245 (0.065) | 0.244 (0.063) | 0.246 (0.064) |
| Acf(20) | | | | |
| Const | 0.429 | 0.435 | 0.395 | 0.368 |
| Treat | 0.872 | 0.779 | 0.804 | 0.721 |
| Time | 0.421 | 0.276 | 0.362 | 0.195 |
| Interact | 0.686 | 0.471 | 0.580 | 0.321 |
| $D_{11}$ | 0.042 | 0.010 | 0.024 | 0.024 |
| $D_{21}$ | 0.096 | 0.017 | 0.018 | 0.003 |
| $D_{22}$ | 0.124 | 0.005 | 0.045 | 0.012 |
| M-H acceptance | | | | |
| $\beta$ | 0.392 | 0.401 | 0.401 | 0.399 |
| $b_i$ min | 0.084 | 0.587 | 0.187 | 0.895 |
| $b_i$ max | 0.429 | 0.610 | 0.466 | 0.911 |


Series : betam1[, 1]
ACF
Lag
Series : betam[, 1]
ACF
Lag
betam1[, 1]
betam[, 1]
Series : betam1[, 4]
ACF
Lag
Series : betam[, 4]
ACF
Lag
betam1[, 4]
betam[, 4]

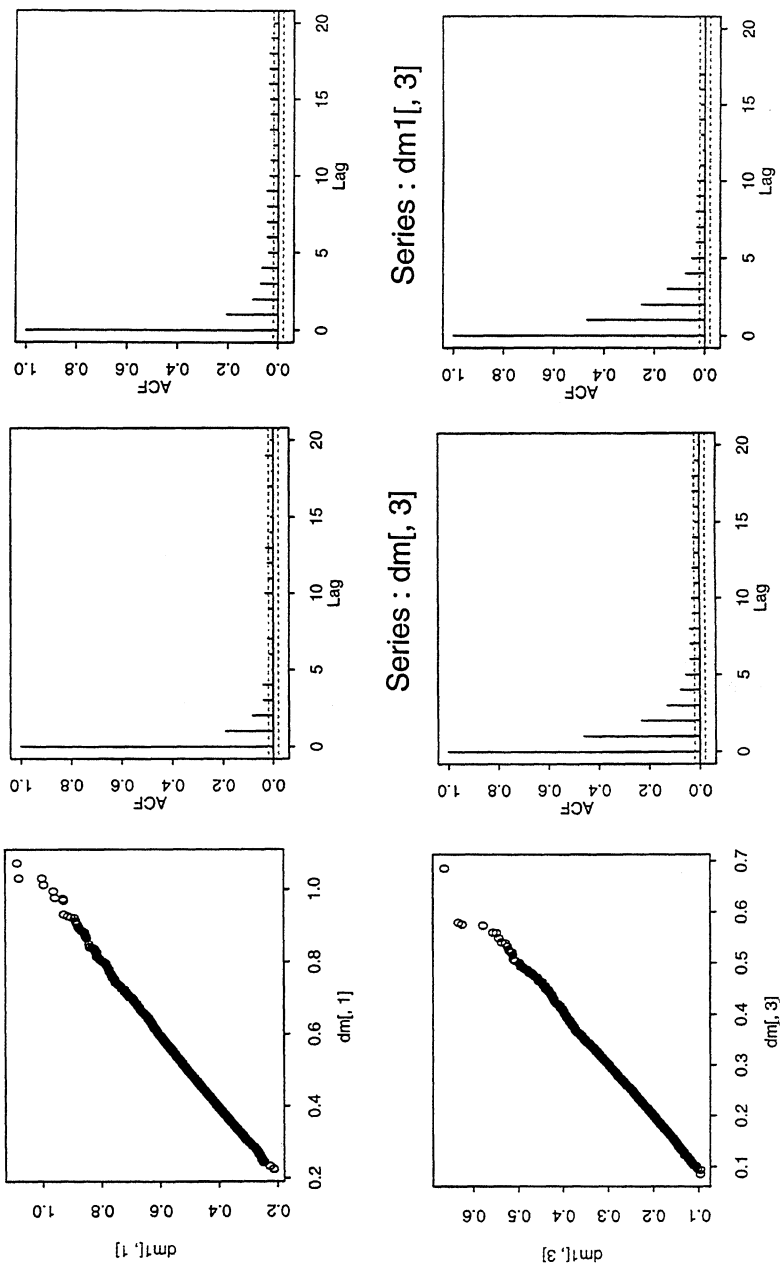



Fig. 1. Epilepsy data. Q–Q plots and autocorrelation functions for selected parameters ($\beta_1$, $\beta_4$, $D_{11}$, $D_{22}$) under alternative parameterizations: output from the ($\beta$, $\eta$)-parameterization (indexed by m) is in the middle column and output from the $\beta$-parameterization (indexed by m1) is in the third column.

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 et al. (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 next 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 $\beta$ in these runs is $N_4(0,10^{-2} \times I)$. For brevity we focus on method 4 and simulate 10,000 draws from the posterior distribution, setting $\tau_\beta^{1/2} = 1.5$ and $\tau_2^{1/2} = 1.5$. We summarize the results in Fig. 1 for ($\beta_1$, $\beta_4$, $D_{11}$, $D_{22}$). The figure contains Q–Q and autocorrelation plots for output from the recommended ($\beta$, $\eta$)-parameterization (second column) and from the $\beta$-parameterization (third column). From these figures we conclude that the Q–Q plots are linear and that the chain displays less serial correlation in the ($\beta$, $\eta$)-parameterization.

The best overall results are obtained when the random effects are simulated by the accept–reject method with a pseudo-dominating density (Method 4) in the ($\beta$, $\eta$) 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 by computing the log marginal likelihoods for the model discussed above ($\mathcal{M}_1$) and for an alternative model ($\mathcal{M}_2$) in which the intercept is the only random effect. The marginal likelihoods are computed from the ($\beta$, $\eta$)-parameterization. Method 4 is used to simulate the random effects. Each of the reduced Markov chain Monte Carlo iterations is 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 $\mathcal{M}_1$ and $-969.824$ for $\mathcal{M}_2$. This is very strong evidence in favor of including the second random effect. The numerical standard error (calculated using the expression in the Appendix) of the former estimate is 0.1, which is negligible compared to $-915.404$.

### 4.2. Patent data

These patent data have previously been analyzed by Hausman et al. (1984) and Blundell et al. (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. With $y_{it}$ denoting the number of patents received by firm $i$ in year $t$, the model of interest specifies,

in the $\beta$-parameterization, that

$$\log E(y_{ij}|\beta, b_i) = \beta_1 + \beta_2 x_{ij1} + \beta_3 x_{ij2} + \beta_4 x_{ij3} + \beta_5 x_{ij4} + b_{i1} + b_{i2} w_{ij1},$$

where $E(b_i) = 0$, $x_{ij1} = w_{ij1}$ 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. In the ($\beta$, $\eta$)-parameterization the model is written as

$$\log E(y_{ij}|\beta, b_i) = \beta_3 x_{ij2} + \beta_4 x_{ij3} + \beta_5 x_{ij4} + b_{i1} + b_{i2} w_{ij1},$$

where $E(b_i) = \eta$ and the variables are defined as above. The model also contains time dummies for 1976–1979, which are suppressed here and in the output for convenience. The data set contains additional variables – a dummy variable for whether 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 included as covariates in the model, because they exhibit no within-firm variation and hence are indistinguishable from the random intercept.

The Markov chain Monte Carlo 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) appears in Table 3. 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 current R&D expenditures have a smaller effect on firms with large intercepts. 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 these data clearly illustrate the advantages of using an MV$t$ tailored proposal as opposed to the Gaussian tailored proposal in the generation of the random effects. The latter proposal was found to yield minimum acceptance rates of 0 and poor mixing in some cases.

Next we report on the results from the $\beta$-parameterization by fitting the above model with Method 3 and setting $\tau_\beta^{1/2} = 0.7$, $\tau_1^{1/2} = 1$, and $\tau_2^{1/2} = 1.5$. For simplicity we compare the marginal posterior distributions of the intercept and the coefficient of $\log R_0$ from the alternative parameterizations. We also examine the autocorrelation plots of the sampled values. The results appear in Fig. 2, where the first column corresponds to the recommended parameterization. It can be seen that the marginal posterior distributions for $\beta_1$ are different, but those of $\beta_2$ are quite close. It appears that the distribution of the intercept in the $\beta$-parameterization has not converged even after 12,000 iterations due to the

Table 3
Patent data: M–H tuning constants, posterior means (standard deviations) and performance summaries in the $(\beta, \eta)$-parameterization. Results are based on $G = 10{,}000$ samples beyond an initial transient stage of 1,000 cycles

| | Method 1 | Method 2 | Method 3 | Method 4 |
|---|---|---|---|---|
| M–H const | | | | |
| $\tau_\beta^{1/2}$ | 0.7 | 1.0 | 0.7 | 1.0 |
| $\tau_1^{1/2}$ | 0.7 | n.a. | 1.0 | n.a. |
| $\tau_2^{1/2}$ | 1.0 | 2.5 | 1.5 | 2.0 |
| Param | | | | |
| constant | 0.776 (0.075) | 0.772 (0.077) | 0.747 (0.076) | 0.733 (0.076) |
| $\log R_0$ | 0.694 (0.030) | 0.697 (0.040) | 0.621 (0.035) | 0.572 (0.036) |
| $\log R_{-1}$ | − 0.043 (0.031) | − 0.055 (0.033) | 0.005 (0.032) | 0.046 (0.033) |
| $\log R_{-2}$ | 0.128 (0.036) | 0.130 (0.036) | 0.138 (0.038) | 0.144 (0.037) |
| $\log R_{-3}$ | 0.092 (0.030) | 0.089 (0.030) | 0.113 (0.030) | 0.129 (0.031) |
| $D_{11}$ | 2.588 (0.259) | 2.668 (0.256) | 2.594 (0.248) | 2.547 (0.252) |
| $D_{21}$ | − 0.578 (0.072) | − 0.618 (0.079) | − 0.597 (0.076) | − 0.585 (0.076) |
| $D_{22}$ | 0.215 (0.027) | 0.293 (0.035) | 0.287 (0.034) | 0.282 (0.032) |
| Acf(20) | | | | |
| Constant | 0.153 | 0.026 | 0.048 | 0.031 |
| $\log R_0$ | 0.480 | 0.322 | 0.221 | 0.171 |
| $\log R_{-1}$ | 0.186 | 0.263 | 0.045 | 0.155 |
| $\log R_{-2}$ | 0.034 | 0.034 | − 0.009 | 0.007 |
| $\log R_{-3}$ | 0.182 | 0.083 | 0.050 | 0.042 |
| $D_{11}$ | 0.515 | 0.117 | 0.204 | 0.011 |
| $D_{21}$ | 0.550 | 0.182 | 0.253 | 0.019 |
| $D_{22}$ | 0.630 | 0.290 | 0.385 | 0.032 |
| M-H acceptance rate | | | | |
| $\beta$ | 0.377 | 0.222 | 0.387 | 0.233 |
| $b_i$ min | 0.015 | 0.259 | 0.121 | 0.818 |
| $b_i$ max | 0.590 | 0.291 | 0.482 | 0.925 |

high serial correlation. For each parameter, the autocorrelation patterns are better behaved in the $(\beta,\eta)$-parameterization. This is the kind of improvement we expected given the high degree 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 Method 3 can deliver an order of magnitude reduction in computing time for large data sets.

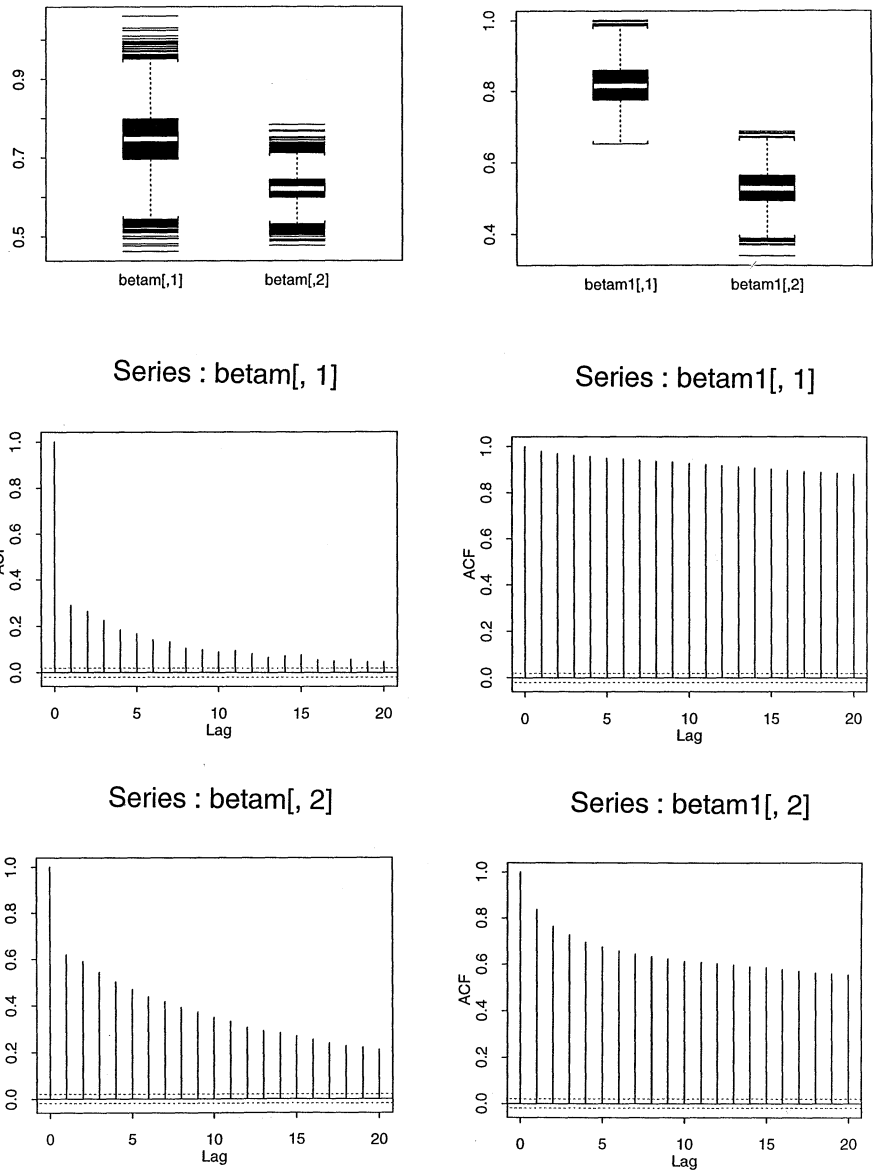



Fig. 2. Patents data. Posterior densities and acf's for $(\beta_1,\beta_2)$ under alternative parameterizations: output from the $(\beta,\eta)$-parameterization is in the first column, and output from the $\beta$-parameterization is in the third column.

## 5. Conclusions

This paper has shown how Markov chain Monte Carlo methods make possible the analysis of rather complex variants of the Poisson panel count model with random effects. We have discussed several different Metropolis–Hastings 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 multivariate-$t$ density. We have found that it is important to use a multivariate-$t$ distribution instead of the Gaussian distribution for this purpose. We have discussed the use of a Metropolis–Hastings accept–reject algorithm with a pseudo-dominating density and documented the value of a new parameterization of the random effects and the fixed effects.

In addition, we have considered the problems of ML estimation and model choice and have developed the first 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 should prove useful and important.

## Acknowledgements

The authors would like to thank two referees for helpful comments and suggestions

## Appendix A. Numerical standard error of the marginal likelihood estimate

In this appendix we briefly discuss how the numerical standard error of the posterior ordinate estimate in Section 3.2 can be derived. The numerical standard error gives the variation that can be expected in the posterior ordinate estimate if the simulation were to be repeated.

Following Chib (1995), let

$$h^{(g)} = \begin{pmatrix} f_{\mathrm{W}}(D^{*-1}|n+v_0,\, R_n^{(g)-1}) \\ \phi(\eta^*|\hat{\eta}^{(g)},\, M_1^{*-1}) \\ \phi(\beta^*|\beta^{(g)},\, H) \end{pmatrix},$$

where $f_{\mathrm{W}}$ is the Wishart density, and note that

$$\hat{h} = G^{-1}\sum_{g=1}^{G} h^{(g)} = \begin{pmatrix} \hat{\pi}(D^{-1}*|y) \\ \hat{\pi}(\eta^*|y,\, D^*) \\ \hat{\pi}(\beta^*|y,\, \eta^*,\, D^{-1}*) \end{pmatrix}.$$

Hence, from Newey and West (1987),

$$\mathrm{Var}(\hat{h}) = G^{-1}\left[\Omega_0 + \sum_{s=1}^{m}\left(1 - \frac{s}{m+1}\right)(\Omega_s + \Omega_s')\right],$$

where

$$\Omega_s = G^{-1}\sum_{g=s+1}^{G}(h^{(g)} - \hat{h})(h^{(g)} - \hat{h})'$$

and $m$ is a constant that represents the lag at which the autocorrelation function of $h^{(g)}$ tapers off. By the delta method the variance of

$$\ln \hat{\pi}(D^{-1}*|y) + \ln \hat{\pi}(\eta*|y, D*) + \ln \hat{\pi}(\beta*|y, \eta*, D*)$$

is given by

$$a' \mathrm{Var}(\hat{h})a,$$

where $a = (\hat{h}_1^{-1}, \hat{h}_2^{-1}, \hat{h}_3^{-1})$. The numerical standard error is the square root of this quantity.

## References


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

Bennett, J.E., Racine-Poon, A., Wakefield, J.C., 1996. MCMC for nonlinear hierarchical models. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), Markov Chain Monte Carlo in Practice, Chapman Hall, London.

Blundell, R., Griffith, R., Van Reenan, J., 1995. Dynamic count models of technological innovation. Economic Journal 105, 333–344.

Carlin, B., Chib, S., 1995. Bayesian model choice via Markov chain Monte Carlo. Journal of the Royal Statistical Society Ser. B 57, 473–484.

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

Chib, S., Greenberg, E., 1995. Understanding the Metropolis–Hastings algorithm. American Statistician 49, 327–335.

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

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

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

Gelfand, A.E., Sahu, S.K., Carlin, B.P., 1996. Efficient parametrizations for generalized linear mixed models (with discussion). In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 5, Oxford University Press, Oxford, 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., Hall, B.H., Griliches, Z., 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 ed., Oxford University Press, New York.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.
Louis, T.A., 1982. Finding the observed information matrix using the EM algorithm. Journal of the Royal Statistical Society Ser B 44, 226–233.
Newey, W.K., West, K.D., 1987. A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica 55, 703–708.
O'Hagan, A., 1994. Kendall's Advanced Theory of Statistics, vol. 2B, Bayesian Inference, Halsted Press, New York.
Press, 1982. Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference, Krieger, Malabar.
Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–549.
Tierney, L., 1994. Markov chains for exploring posterior distributions (with discussion). Annals of Statistics 22, 1701–1762.
Tierney, L., Kadane, J., 1986. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 81, 82–86.
Wakefield, J.C., Smith, A.F.M., Racine Poon, A., Gelfand, A.E., 1994. Bayesian analysis linear and non-linear population models by using the Gibbs sampler. Applied Statistics 43, 201–221.
Wei, G.C.G., Tanner, M.A., 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., Karim, M.R., 1991. Generalized linear models with random effects: A Gibbs sampling approach. Journal of the American Statistical Association 86, 79–86.