Posterior Simulation and Bayes Factors in Panel Count Data 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.

July 1996 (revised November 1996)

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 from the simulation output is considered. The methods are illustrated with several real data applications involving large samples and multiple random effects.

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

* *Address for correspondence*: **John M. Olin School of Business, Washington University, Campus Box 1133, 1 Brookings Drive, St. Louis, MO 63130. E-Mail: chib@simon.wustl.edu**

recently begun to address these problems from various numerical perspectives, primarily organized around Bayesian Markov chain Monte Carlo algorithms [Albert (1992), Bennett et al. (1996), Gamerman (1994), Wakefield et al. (1994) and Zeger and Karim (1991)]. It has also become understood that certain identification problems can severely compromise the performance of the existing simulation methods. Gelfand, Sahu, and Carlin (1996) discuss one approach for dealing with this problem, but this approach does not appear to be computationally straightforward for 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, Sahu, and Carlin (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 simulation routines based on Markov chain Monte Carlo methods for sampling the posterior distribution of the parameters and the random effects. These routines 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 three applications of the techniques, first to data on the effects of the drug progabide on epileptic patients, then to patent data on a longitudinal sample of 680 firms in the United

States, and finally to German data on the number of absences from work for a sample of 704 male workers. Concluding remarks appear in Section 5.

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 μ_{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 nonzero mean vector η 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 $E(b_i) = 0$ [Laird and Ware (1982)]. This parameterization is not to be recommended in the context of Markov chain Monte Carlo methods such as those we propose below, 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 β_k is not likelihood identified [O'Hagan (1995)]. 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 β and b_i will not mix well. Transferring the "common" effect of x_k to η_k removes the nonidentified parameter β_k. We mention that our parameterization is related to, but different from, the hierarchical centering introduced by Gelfand, Sahu, and Carlin (1996).

Since our approach to estimation is Bayesian, we complete the model by assuming that the parameters (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 $(\beta_0, B_0, \eta_0, M_0, v_0, R_0)$ are known hyperparameters and Wish (ν_0, R_0) is the Wishart distribution with ν_0 degrees of freedom and scale matrix R_0 [Press (1989)].

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 ith cluster. Under conditional independence

$$\begin{aligned} f(y_i|\beta, b_i) &= \prod_{t=1}^{T_i} p(y_{it}|\beta, b_i) \quad \text{and} \\ f(y_i, b_i|\beta, \eta, D) &= f(y_i|\beta, b_i)\, \phi(b_i|\eta, D) \end{aligned}$$

is the joint density of (y_i, b_i), where p is the Poisson mass function with conditional mean μ_{it} and $\phi(b_i|\eta, D)$ is the density of the normal distribution with mean η and covariance D. The likelihood function of the parameters given $y = (y_1, \ldots, y_n)$ is then given by

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

which is the product of the n likelihood contributions $L_i(y_i|\beta, \eta, D)$. The intractability of the likelihood function arises from the difficulty of evaluating the integral in the first line of (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 $b = (b_1, b_2, \ldots, b_n)$, β, η, and D, and the associated full conditional distributions

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

Given an arbitrary starting point in the parameter space, these distributions are sampled recursively, where the conditioning variables are set at their most recent simulated values. To implement this procedure we need to sample each of the full conditional distributions. We now explain how this can be done.

The main computational problem arises in the sampling of the n random effects b_i from the distribution $\pi(b|y, \beta, \eta, D) = \prod_{i=1}^{n} \pi(b_i|y_i, \beta, \eta, D)$. Specifically, the ith target density is

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

which cannot be sampled by standard methods. As discussed by several authors, such densities can be sampled by the Metropolis-Hastings algorithm [Tierney (1994) and Chib and Greenberg (1995)]. Recall that, for a given target density $f(\psi)$, the Metropolis-Hastings 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 $\alpha(\psi, \psi^{\dagger})$ and if rejected, the next sampled value is taken to be ψ. The key question concerns the choice of the proposal density q. We discuss several choices and provide insight into which are likely to be useful in the commonly occurring context of large data sets and multiple random effects. The methods are systematically compared in our examples.

Method 1: Random walk proposal

For this method let $q_1(b_i, b_i^{\dagger}) = \phi(b_i^{\dagger}|b_i, \tau_1 D)$, 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

A second approach is to tailor the proposal density to the target density around its modal value $\hat{b}_i = \arg\max_{b_i} \ln f(y_i, b_i|\beta, \eta, D)$. We now define

$$q_2 = \mathrm{MV}t(b_i|\hat{b}_i, \tau_2 V_{b_i}, \nu),$$

where τ_2 is another scaling factor, $\mathrm{MV}t(\cdot|\ \hat{b}_i, \tau_2 V_{b_i}, \nu)$ is the multivariate-t distribution with location parameter $\hat{b}_i$, scale matrix $\tau_2 V_{b_i}$, and ν degrees of freedom; we set $V_{b_i} = (-H_{b_i})^{-1}$, which is the curvature around the mode. We propose the use of a multivariate-t distribution for the following reason: Since the probability of move is given by

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

it is important that the weight function $w(s)$ be bounded. It can be checked that this is true for the multivariate-t density but not for the Gaussian density that has been recommended in the literature. With the Gaussian density the probability of move is effectively zero where the density is thin, and the chain tends to get stuck.

The mode and the Hessian can be computed approximately by the Newton-Raphson method, for which the gradient 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.

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. The main purpose of this approach is to speed up the computations, which can be achieved by selecting q_2 less frequently than q_1. It is possible to select the respective densities at fixed, pre-specified intervals, thus preserving the Markov property of the simulation. This combination of proposal densities is a useful way of producing satisfactory proposal values at moderate computational expense.

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 $c_i \text{MV}t(b_i|\hat{b}_i, \tau_2 V_{b_i}, \nu)$, 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 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, p. 332)] is defined as

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

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

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

Given the random effects, the remaining simulations are quite straightforward, with both η and D being simulated from standard distributions. For β, 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 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 MVt increment with zero mean and scale matrix $\tau_\beta V_\beta$. It is easy to check that the resulting proposal density is given by $q(\beta, \beta^\dagger) = \text{MV}t(\beta^\dagger|\hat{\beta} - (\beta - \hat{\beta}), \tau_\beta V_\beta, \nu)$, which is symmetric in $(\beta, \beta^\dagger)$. Chib and Greenberg (1995) have documented the value 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 Markov chain Monte Carlo simulation one samples η from

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

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 Markov chain Monte Carlo sampling.

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 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 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. Chib (1995) suggests that a high density point, such as the posterior mean of θ or the maximum likelihood estimate (whose computation is discussed below), be used as the point θ^*.

We next consider the calculation of each term in (10) from the Markov chain Monte Carlo 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. The contribution of y_i to the likelihood at the point θ^* is

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

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

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

where $\hat{b}_i$ denotes the mode of $\ln\{f(y_i|b_i, \beta^*)\phi(b_i|\eta^*, D^*)\}$, H_{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 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 MV$t(\cdot|\hat{b}_i, (-H_{b_i})^{-1}, \nu)$. The log-likelihood function is obtained by adding the ln $\hat{L}_i(y_i|\theta^*)$.

For the simulated data set described above we let $M = 2000$ and specify 10 degrees of freedom for the MVt 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, we continue to suppress $\mathcal{M}$ and 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

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

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

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

Second, each of these ordinates is estimated from the Markov chain Monte Carlo output. A little reflection shows that to estimate (13) one requires output from the initial Markov chain Monte Carlo 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 obtained by averaging the Wishart density $\pi(D^{-1*}|b, \eta)$ in (13) over these simulated draws. Next, a reduced Markov chain Monte Carlo 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 can 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. The ML estimate may also serve as a starting point for the full Markov chain Monte Carlo 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

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

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

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

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 Metropolis-Hastings 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.

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 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

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

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

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

Poisson link. In the (β, η)-parameterization, we let

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

and in the β-parameterization

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

Thus η corresponds to (β_1, β_3) since the intercept and x_{ij2} (time) variables are random effects.

Focusing first on the (β, η)-parameterization, we experiment with the four alternative proposal generating densities for b discussed in Section 2 under the following vague priors for β, η, and D:

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

Tuning constants in these methods (such as τ_1 and τ_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 (β, η)-parameterization. The table contains the posterior means, the posterior standard deviations, the autocorrelation at lag 20 of the generated sample for each of the alternative methods, and the acceptance rates in the b_i and β steps. Because there are a large number of b_i, we report only the minimum and maximum acceptance rates achieved in the sampling. We have found that this diagnostic is a useful summary of the performance of the Metropolis-Hastings simulations, given that the acceptance rate for each random effect cannot be easily 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 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 β 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 $\tau_\beta = 1.5$ and $\tau_2 = 1.5$. We summarize the results 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 β-parameterization (third column). From these figures we conclude that the Q-Q plots are linear and that the

	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 1,000 cycles.

chain displays less serial correlation in the (β, η)-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 (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 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 (β, η)-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.

4.2 Patent data

These patent data have 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. With y_{it} denoting the number of patents received by firm i in year t, the model of interest specifies, in the β-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 (β, η)-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.

	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 1,000 cycles.

It is also interesting to mention that these data clearly illustrate the advantages of using an MVt 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 β-parameterization by fitting the above model with Method 3 and setting $\tau_\beta = .7$, $\tau_1 = 1$, and $\tau_2 = 1.5$. For simplicity we compare the marginal posterior distributions of the intercept and the coefficient of log R_0 from the alter-

native 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 quite close. It appears that the distribution of the intercept in the β-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 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.

4.3 Absence data

Our final illustration is with data 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.

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 days 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 job tenure and job satisfaction variables 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 whom the (4×3) matrix containing the constant, job tenure, and job satisfaction has full column rank.

We specify four models for these data. The first three 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, job satisfaction.
- Model 4. Job tenure, job satisfaction.

For each model we simulate the posterior density in the (β, 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.

The prior densities and the Markov chain Monte Carlo design are similar to those 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 increase it to 1,000. The tuning constants are adjusted to produce acceptance rates between 0.3 and 0.5.

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 from $M = 1,000$ random effects draws after convergence.

Table 4 displays the results from the MCEM estimation, and Table 5 shows the posterior means, the posterior standard deviations, and the log marginal likelihoods from the Markov chain Monte Carlo simulation. We note that both maximum likelihood estimators 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 intercepts is better than Model 4: the included time invariant covariates are not able to explain the between-individual variation in absences.

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

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 with Method 3. Results are based on $G = 10,000$ samples beyond an initial transient stage of 2,000 cycles.

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.

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 (1996), "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.

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.

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., 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.

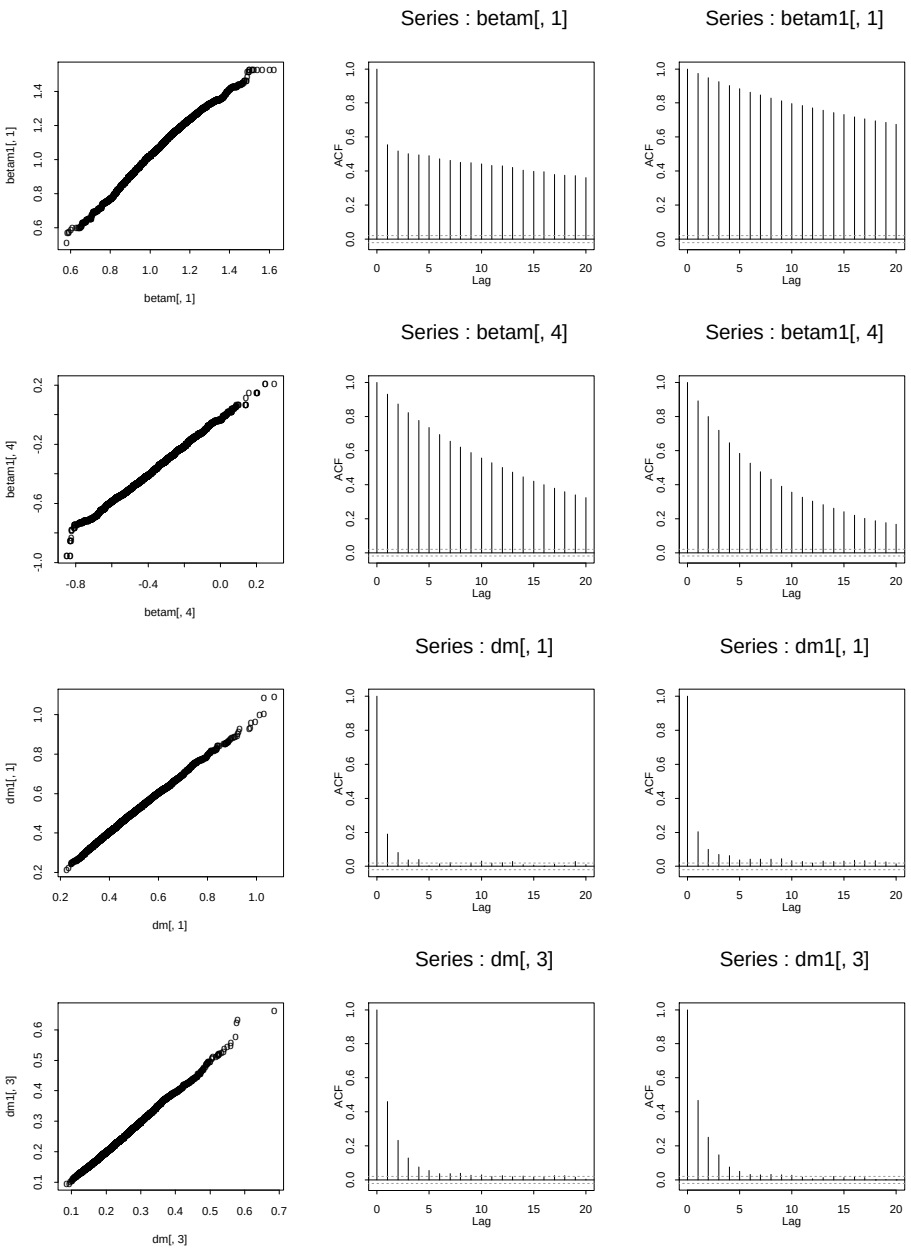

Figure 1: Epilepsy data. Q-Q plots and autocorrelation functions for selected parameters under alternative parameterizations: output from (β, η)-parameterization is in second column.

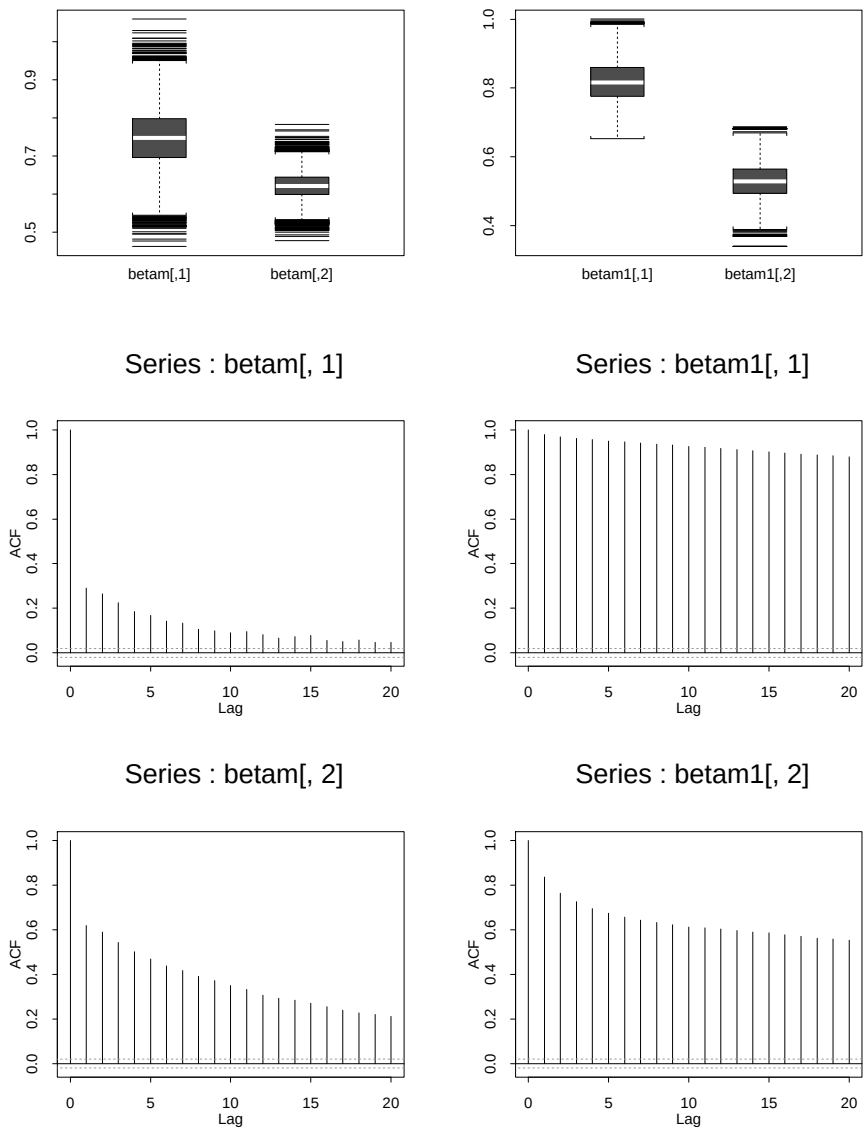

Figure 2: Patents data. Posterior densities and acf's for (β_1, β_2) under alternative parameterizations: output from (β, η)-parameterization is in first column.