Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2004-04

Data Augmentation and Gibbs Sampling for Regression Models of Small Counts

Sylvia Frühwirth-Schnatter and Helga Wagner

first version: October 15, 2004
revised version: Februar 3, 2005

Abstract

In this article we consider Bayesian analysis of Poisson regression models. Estimation is carried out within a Bayesian framework using data augmentation and MCMC methods. We suggest a new MCMC sampler, which possesses a Gibbs transition kernel, where we draw from full conditional distributions belonging to standard distribution families, only. This Gibbs sampler is applied to a standard Poisson regression model and to a Poisson regression models dealing with overdispersion.

Key words: Poisson count data, data augmentation, Gibbs sampling, overdispersion

1 Introduction

Applied statisticians commonly have to deal with count data, recording for instance the number of road accidents or disease occurrences. Such data are necessarily nonnegative integers and it is often appropriate to assume that the observed counts follow a Poisson distribution. The Poisson regression model, discussed for instance in McCullagh and Nelder (1999) in the framework of generalized linear models, is an important tool for analyzing the effect of covariates on count data. The basic Poisson regression model has been modified in a number of ways. To account for the dependency likely to be present in sequences of counts data, the covariates may depend on past observations, see for instance Zeger and Qaqish (1988). A couple of extensions deal with overdispersion due to omitted covariates, we mention in particular mixtures of Poisson regressions models (Wang et al. 1996; Hurn et al. 2003), Poisson regression models with additive random effects (Aitkin 1996), panel Poisson regression models with random effects (Chib et al. 1998), and mixtures of Poisson regression models with random effects (Lenk and DeSarbo 2000).

In this paper we consider Bayesian estimation of Poisson regression models, using data augmentation as in Tanner and Wong (1987) and Markov chain Monte Carlo (MCMC) methods, as illustrated first by Zeger and Karim (1991) for generalized linear models with random effects. Since this seminal paper, a number of authors have contributed to MCMC estimation of regression models for count data. We mention here in particular Albert (1992) for Poisson random-effects models, Chib et al. (1998) for panel count data models with multiple random effects, Lenk and DeSarbo (2000) for mixtures of Poisson models with random effects, and Hurn et al. (2003) for mixtures of Poisson regression models.

A major difficulties with any of the existing MCMC approaches is that practical implementation requires the use of a Metropolis-Hastings algorithm at least for part of the unknown parameter vector, which in turns make it necessary to define suitable proposal densities. The present article discusses a new method of data augmentation, and straightforward Gibbs sampling, put forward by Frühwirth-Schnatter and Wagner (2004), in the context of Poisson regression models. Our main result is to show that a Poisson regression model may be regarded as a partially Gaussian regression model in the sense of Shephard (1994), by conditioning on two sequences of suitably chosen artificially missing data. The first sequence are the unobserved inter-arrival times of a suitably chosen Poisson process. This eliminates the non-linearity of the Poisson regression model, and leads to a linear regression model with non-normal errors, that follow a log exponential distribution with mean 1. The log exponential distribution is then approximated by a mixture of normal distributions in a similar way as in Kim et al. (1998) and Chib et al. (2002). We introduce the component indicator of this normal mixture as a second sequence of missing data. By conditioning on both sequences, a Gaussian regression model results. Based on this useful result, we will show that straightforward Gibbs sampling of all regression parameters, and all missing data is possible, requiring only random draws from standard distributions such as multivariate normals, inverse Gamma, exponential and discrete distributions with a few categories.

The rest of the paper is organized as follows. After a short review of the Poisson regression model in Section 2, we discuss in Section 3 in detail data augmentation for Poisson regression models, and implementation of our new Gibbs sampling scheme. Applications to standard regression models as well as to regression models dealing with overdispersion are considered in Section 4, whereas Section 5 concludes.

2 Poisson Regression Models

Let $\mathbf{y} = (y_1, \dots, y_N)$ be a collection of count data. In what follows, we assume that $y_i|\lambda_i$ follows a $\mathcal{P}(\lambda_i)$ distribution, where λ_i depends on an unknown model parameter $\boldsymbol{\vartheta}$ in the following way:

$$y_i|\lambda_i \sim \mathcal{P}(\lambda_i), \qquad \lambda_i = \exp(\mathbf{z}_i\boldsymbol{\vartheta}), \tag{1}$$

where $\mathbf{z}_i$ is a row vector of regressors, including 1 for the intercept. In the present paper we will be interested in two special cases of this model. First, we consider the standard Poisson regression model, where for each count variable y_i, we observe covariates $\mathbf{x}_i$, again including 1 for the intercept. Then in (1), $\mathbf{z}_i = \mathbf{x}_i$, and $\boldsymbol{\vartheta} = \boldsymbol{\beta}$ is a simple regression parameter. We will show below, that our data augmentation method leads to a normal regression model, where the whole regression parameter $\boldsymbol{\beta}$ could be sampled in one sweep from a normal distribution.

Second, we will consider a Poisson regression model dealing with overdispersion due to omitted covariates. A common way of dealing with this kind of overdispersion is the individual effects model introduced by Aitkin (1996), where the regression intercept varies between the units:

$$\lambda_i = \exp(\alpha_i + \mathbf{x}_i\boldsymbol{\beta}), \tag{2}$$

where $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$. Thus overdispersion is modelled on the same level as the linear predictor. Formally, this model may be written as a special case of (1), where $\boldsymbol{\vartheta} = (\alpha_1, \dots, \alpha_N, \boldsymbol{\beta})$, and

$$\mathbf{z}_i = \begin{pmatrix} \mathbf{e}_i & \mathbf{x}_i \end{pmatrix},$$

where $\mathbf{e}_i$ is a $(1 \times N)$ row vector, containing only zeros, apart from column i, which is equal to 1. Marginally, this model is an infinite mixture of Poisson regression models with no closed form. Aitkin (1996) suggested to approximate the marginal distribution by a mixture Poisson regression models using Gaussian-Hermite quadrature. We will show, how data augmentation leads to a normal random-effects regression model, where the whole sequence $\boldsymbol{\vartheta} = (\alpha_1, \dots, \alpha_N, \boldsymbol{\beta})$ could be sampled simultaneously in an efficient manner.

3 A Bayesian Analysis

3.1 Prior and Posterior Distributions

We assume that the prior distribution $p(\boldsymbol{\vartheta}|\boldsymbol{\delta})$ of $\boldsymbol{\vartheta}$ follows a normal distribution, which is allowed to be indexed by an unknown hyper parameter $\boldsymbol{\delta}$. For the standard regression model $p(\boldsymbol{\beta}|\boldsymbol{\delta})$ typically is normal prior $\mathcal{N}_d\left(\mathbf{b}_0, \mathbf{B}_0\right)$, with known hyperparameters $\mathbf{b}_0$ and $\mathbf{B}_0$. For a regression model with overdispersion, this prior is extended to:

$$p(\boldsymbol{\vartheta}|\boldsymbol{\delta}) = p(\boldsymbol{\beta}|\mathbf{b}_0, \mathbf{B}_0) \prod_{i=1}^{N} p(\alpha_i|\sigma_\alpha^2),$$

where $\alpha_i \sim \mathcal{N}\left(0, \sigma_\alpha^2\right)$. The additional hyperparameter σ_α^2 may be known or unknown.

For both models, these assumptions are sufficient to derive the conditional posterior density $p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \mathbf{y})$ by Bayes' theorem, given all observations $\mathbf{y} = (y_1, \dots, y_N)$:

$$p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}|\boldsymbol{\delta}), \quad p(\mathbf{y}|\boldsymbol{\vartheta}) = \prod_{i=1}^{N} \frac{\exp(\mathbf{z}_i\boldsymbol{\vartheta})^{y_i}}{y_i!} \exp(-\exp(\mathbf{z}_i\boldsymbol{\vartheta})).$$

The resulting posterior density, however, in general does not belong to a density from a well-known distribution family. Markov chain Monte Carlo methods to sample from the posterior distribution of a Poisson regression model were applied by Zeger and Karim (1991), Albert (1992), Chib et al. (1998), Lenk and DeSarbo (2000) and Hurn et al. (2003), among many others. As mentioned in the introduction, any of these methods is based on Metropolis-Hastings sampling.

We are going to demonstrate in the following subsection, that the introduction of two sequences of artificially missing data within a data augmentation scheme leads to a conditional posterior distribution for $\boldsymbol{\vartheta}$ that, in contrast to $p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \mathbf{y})$, is a joint normal distribution, once we conditioned on the artificially missing data.

3.2 Data Augmentation

For each i, the distribution of $y_i|\lambda_i$ may be regarded as the distribution of the number of jumps of an unobserved Poisson process with intensity λ_i, having occurred in the time interval [0,1]. The first step of data augmentation introduces for each i, $i = 1, \dots, N$, the inter-arrival times τ_{ij}, $j = 1, \dots, (y_i + 1)$ of this Poisson process as missing data. From the basic properties of a Poisson process, the inter-arrival times τ_{ij} are known to follow the $\mathcal{E}\left(\lambda_i\right)$-distribution:

$$\tau_{ij}|\boldsymbol{\vartheta} \sim \mathcal{E}\left(\lambda_i\right) = \frac{\mathcal{E}\left(1\right)}{\lambda_i}.$$

Table 1: Normal mixture approximation of the density of the $\log \mathcal{E}\,(1)$-distribution (5 components)

r	1	2	3	4	5
w_r	0.2924	0.2599	0.2480	0.1525	0.0472
m_r	0.0982	-1.5320	-0.7433	0.8303	-3.1428
s_r^2	0.2401	1.1872	0.3782	0.1920	3.2375

As $\lambda_i = \exp(\mathbf{z}_i\boldsymbol{\vartheta})$, this may be reformulated as following linear model:

$$\log \tau_{ij}|\boldsymbol{\vartheta} = -\mathbf{z}_i\boldsymbol{\vartheta} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \log(\mathcal{E}\,(1)). \tag{3}$$

Let $\boldsymbol{\tau} = \{\tau_{ij}, j = 1, \ldots, (y_i + 1), i = 1, \ldots, N\}$ denote the collection of all inter-arrival times. Our first data augmentation step introduces the inter-arrival times $\boldsymbol{\tau}$ as missing data, with two effects. First, the full-conditional posterior distribution $p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \boldsymbol{\tau}, \mathbf{y})$ of $\boldsymbol{\vartheta}$, where additionally to $\boldsymbol{\delta}$ and $\mathbf{y}$ the inter-arrival times $\boldsymbol{\tau}$ appear as conditioning argument, is independent of $\mathbf{y}$, $p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \boldsymbol{\tau}, \mathbf{y}) = p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \boldsymbol{\tau})$. Second, conditional on $\boldsymbol{\tau}$, we are dealing with model (3), which is non-normal, but where the mean of the observation equation is linear in the unknown model parameters $\boldsymbol{\vartheta}$. Thus, the first augmentation steps eliminates the non-linearity of the Poisson regression model, the non-normality of the error term, however, remains.

It is important to realize that the error term in (3) follows a $\log \mathcal{E}\,(1)$-distribution which is independent of any unknown model parameter. To obtain a model that is conditionally Gaussian, we start by approximating the non-normal density of $\varepsilon_{ij} \sim \log(\mathcal{E}\,(1))$ by a normal mixture of 5 components with parameters m_r and s_r for the r-th component:

$$p(\varepsilon_{ij}) = \exp\{\varepsilon_{ij} - e^{\varepsilon_{ij}}\} \approx \sum_{r=1}^{5} w_r f_{\mathcal{N}}(\varepsilon_{ij}; m_r, s_r^2). \tag{4}$$

This idea is influenced by the related articles of Kim et al. (1998) and Chib et al. (2002), who used a normal mixture approximation of the density of a $\log \chi^2$-distribution in the context of stochastic volatility models. The appropriate parameters $(w_r, m_r, s_r^2), r = 1, \ldots, 5$, however, are different for our problem and are tabulated in Table 1 for 5 components, a number that we found to be sufficiently large in practice. The parameters of the mixture approximation were determined by minimizing the Kullback-Leibler distance.

Following Kim et al. (1998) and Chib et al. (2002), the mixture distribution (4) is regarded as the marginal distribution of a problem where additional to ε_{ij} the component indicators r_{ij} are observed. The second step of our data augmentation scheme introduces for each ε_{ij} the latent component indicator r_{ij} as missing data. Let $\mathbf{R} = \{r_{ij}, j = 1, \ldots, (y_i{+}1), i = 1, \ldots, N\}$ denote the collection of all component indicators r_{ij}. Conditional on $\boldsymbol{\tau}$ and $\mathbf{R}$ the Poisson regression model (1) reduces to a Gaussian regression model with heteroscedastic errors with known variance:

$$\log \tau_{ij}|\boldsymbol{\vartheta}, r_{ij} = -\mathbf{z}_i\boldsymbol{\vartheta} + m_{r_{ij}} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}\left(0, s_{r_{ij}}^2\right). \tag{5}$$

For such a model the conditional posterior $p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \boldsymbol{\tau}, \mathbf{R}, \mathbf{y})$, is a multivariate normal density, which is easy to sample from. This result is the basis for our new two-block Gibbs sampler, that will be described in the next subsection.

3.3 The Basic Two-block Gibbs Sampler

When the hyperparameter $\boldsymbol{\delta}$ of the prior $p(\boldsymbol{\vartheta}|\boldsymbol{\delta})$ are assumed to be known, a two-block Gibbs sampler results, if data augmentation as described in the previous section is applied. Select a starting value for the component indicators $\mathbf{R} = \{r_{ij}, j = 1, \ldots, y_i + 1, i = 1, \ldots, N\}$, and the inter-arrival times $\boldsymbol{\tau} = \{\tau_{ij}, j = 1, \ldots, y_i + 1, i = 1, \ldots, N\}$ and repeat the following steps:

(a) Multi-move sampling of $\boldsymbol{\vartheta}$ conditional on knowing $\boldsymbol{\tau}$, $\mathbf{R}$, and $\mathbf{y}$, based the normal regression model (5).

(b) Sampling of the inter-arrival times $\boldsymbol{\tau}$ and $\mathbf{R}$ conditional on knowing $\boldsymbol{\vartheta}$ and $\mathbf{y}$. For $i = 1, \ldots, N$ run through the following steps (b1) to (b3) with $\lambda_i = \exp(\mathbf{z}_i\boldsymbol{\vartheta})$ and $n = y_i$:

(b1) If $y_i > 0$, sample the order statistics $u_{t,(1)}, \ldots, u_{t,(n)}$ of n uniformly distributed random variables, see e.g. Robert and Casella (1999, p.47) for details, and define the inter-arrival times τ_{ij} as their increments: $\tau_{ij} = u_{i,(j)} - u_{i,(j-1)}, j = 1, \ldots, n$, where $u_{i,(0)} := 0$.

(b2) Sample the final arrival time as $\tau_{i,n+1} = 1 - \sum_{j=1}^{n} \tau_{ij} + \xi_i$, where $\xi_i \sim \mathcal{E}(\lambda_i)$.

(b3) For each $j = 1, \ldots, y_i + 1$, sample the component indicators r_{ij} conditional on τ_{ij} and $\boldsymbol{\vartheta}$ from the following discrete density:

$$\Pr(r_{ij} = k|\tau_{ij}, \boldsymbol{\vartheta}) \propto p(\tau_{ij}|r_{ij} = k, \boldsymbol{\vartheta})w_k, \tag{6}$$

where

$$\ln p(\tau_{ij}|r_{ij} = k, \boldsymbol{\delta}, \boldsymbol{\vartheta}) \propto -\ln s_k - \frac{1}{2}\left(\frac{\ln \tau_{ij} + \mathbf{z}_i\boldsymbol{\vartheta} - m_k}{s_k}\right)^2.$$

The quantities $(w_k, m_k, s_k^2), k = 1, \ldots, 5$ are the parameters of the finite mixture approximation tabulated in Table 1.

Note that step (b) involves only draws from standard densities. Thus sampling scheme (a) and (b) is actually a Gibbs sampler without any tuning.

Comments

It is easy to verify the different sampling steps (b1) to (b3). The joint posterior $p(\mathbf{R}, \boldsymbol{\tau}|\mathbf{y}, \boldsymbol{\vartheta})$ is decomposed as:

$$p(\mathbf{R}, \boldsymbol{\tau}|\mathbf{y}, \boldsymbol{\vartheta}) = p(\mathbf{R}|\boldsymbol{\tau}, \mathbf{y}, \boldsymbol{\vartheta})p(\boldsymbol{\tau}|\mathbf{y}, \boldsymbol{\vartheta}).$$

The inter-arrival times $\{\tau_{ij}, j = 1, \ldots, y_i + 1\}$ are independent for different time points i, given $\mathbf{y}$, and $\boldsymbol{\vartheta}$:

$$p(\boldsymbol{\tau}|\mathbf{y}, \boldsymbol{\vartheta}) = \prod_{i=1}^{N} p(\tau_{i1}, \ldots, \tau_{i,y_i}, \tau_{i,y_i+1}|y_i, \boldsymbol{\vartheta}).$$

For fixed i, the inter-arrival times $\tau_{i1}, \dots, \tau_{i,n+1}$, where $n = y_i$, are stochastically dependent, and the joint distribution factorizes as:

$$\begin{aligned} & p(\tau_{i1}, \dots, \tau_{in}, \tau_{i,n+1} | y_i = n, \boldsymbol{\vartheta}) \\ & = p(\tau_{i,n+1} | y_i = n, \boldsymbol{\vartheta}, \tau_{i1}, \dots, \tau_{in}) p(\tau_{i1}, \dots, \tau_{in} | y_i = n). \end{aligned}$$

The first n inter-arrival times are independent of $\boldsymbol{\vartheta}$ and the component indicator $\mathbf{R}$, and are determined only by the observed number of counts y_i. Due to well-known properties of a Poisson process, the n arrival times occurring in [0,1] are distributed as the order statistics of $n\, \mathcal{U}\,[0,1]$-distributed random variables, and step (b1) follows immediately. The final inter-arrival time $\tau_{i,n+1}$ depends on the actual model parameters $\boldsymbol{\vartheta}$ through the risk λ_i, but is also independent of the component indicator $\mathbf{R}$. Conditionally on $y_i = n$ and $\tau_{i1}, \dots, \tau_{in}$, the last arrival time $\tau_{i,n+1}$ has an exponential distribution with mean $1/\lambda_i$, truncated at $1 - \sum_{j=1}^{n} \tau_{ij}$, thus step (b2) follows.

The component indicators r_{ij} are mutually independent for different i as well as for different j, given $\boldsymbol{\tau}$, $\boldsymbol{\vartheta}$ and $\mathbf{y}$:

$$p(\mathbf{R}|\boldsymbol{\tau}, \mathbf{y}, \boldsymbol{\vartheta}) = \prod_{i=1}^{N} \prod_{j=1}^{y_i+1} p(r_{ij}|\tau_{ij}, \boldsymbol{\vartheta}).$$

For i, j fixed, the posterior of each component indicator r_{ij} depends on the data only through τ_{ij} and on the model parameters $\boldsymbol{\vartheta}$ only through the risk λ_i, thus step (b3) follows immediately.

Step (b1) could be used to sample starting values for $\tau_{i1}, \dots, \tau_{in}$ for each i, given the observed counts y_i. To obtain a starting value for $\tau_{i,n+1}$, we use (b2) and sample ξ_i from $\mathcal{E}\,(\lambda_i)$ with $\lambda_i = y_i$. For all i, where $y_i = 0$, λ_i can be set to a "small" value for λ_i, in our examples we used $\lambda_i = 0.1$. Starting values for each component indicator r_{ij} are obtained as random draws from 1 to 5.

4 Applications

4.1 Application to the Standard Poisson Regression Model

For the Poisson regression model $y_i \sim \mathcal{P}\,(\lambda_i)\,, \log \lambda_i = \mathbf{x}_i \boldsymbol{\beta}$, data augmentations by $\boldsymbol{\tau}$ and $\mathbf{R}$ as described above, leads to a normal regression model with $n_i = y_i + 1$ repeated measurements, and heteroscedastic errors with known variance. For each i, this model reads:

$$\log \tau_{ij} | \boldsymbol{\vartheta}, r_{ij} = -\mathbf{x}_i \boldsymbol{\beta} + m_{r_{ij}} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}\left(0, s_{r_{ij}}^2\right), \tag{7}$$

where $j = 1, \dots, y_i + 1$.

Under the normal prior $\boldsymbol{\beta} \sim \mathcal{N}_d\,(\mathbf{b}_0, \mathbf{B}_0)$, the conditional posterior $p(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R}, \mathbf{y}) = p(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R})$ is equal to the multivariate normal density $\mathcal{N}_d\,(\mathbf{b}_N, \mathbf{B}_N)$ with following posterior moments:

$$\mathbf{B}_N^{-1} = \mathbf{B}_0^{-1} + \sum_{i=1}^{N} \mathbf{x}_i^{'} \mathbf{x}_i w_i, \qquad \mathbf{b}_N = \mathbf{B}_N (\mathbf{B}_0^{-1} \mathbf{b}_0 - \sum_{i=1}^{N} \mathbf{x}_i^{'} \tilde{y}_i), \tag{8}$$

Table 2: Parameter estimates for the regression model

Parameter	Gibbs Sampling			Maximum Likelihood	
	Mean	Std.dev	95%H.P.D. regions	Mean	Std.dev
β	0.7028	0.0756	[0.5501, 0.8466]	0.7177	0.0730
δ	-0.3553	0.1080	[-0.5629, -0.1379]	-0.3543	0.1077

where

$$w_i = \sum_{j=1}^{y_i+1} \frac{1}{s^2_{r_{ij}}}, \qquad \tilde{y}_i = \sum_{j=1}^{y_i+1} \frac{\log \tau_{ij} - m_{r_{ij}}}{s^2_{r_{ij}}}. \tag{9}$$

This result is derived in a straightforward manner from a Bayesian analysis of the normal regression model (7), see Zellner (1971), among many other references.

4.1.1 Application to Road Safety Data

In this application we use the Gibbs sampler to analyze a data set provided by the Austrian Road Safety Board. These data are monthly counts of killed or injured pedestrians, aged 6-10 in Linz, which is the third largest town in Austria. The time period covered was 1987 to 2002. A legal intervention intended to increase road safety took place during the observation period. On October 1, 1994 an amendment increasing priority for pedestrians became effective: since then pedestrians who want to use a crosswalk have to be allowed crossing without risk. We model these data using a regression model with fixed seasonal effects s_i for the different months and the intervention effect is modelled as a level shift at the time point $t = t_{int}$ when legal amendments became effective. Our model parameters are an intercept β, the seasonal dummies $s_1, \dots, s_{11}$ and the intervention effect δ, the covariate vector is of dimension 1×13.

The Gibbs sampler described in Subsection 3.3 was run 12000 times with a burn in of 2000 runs. Table 2 reports point estimates as well as 95%-H.P.D. regions for the intercept and the intervention effect and compares them to the maximum likelihood estimates. The intervention effect is significantly negative. ML estimates and MCMC estimates are very similar which is also true for the seasonal effects.

Figure 1 shows the observed counts with the exponentiated estimated level including the intervention effect and pointwise 95% credible intervals and the seasonal pattern $\exp(s_t)$. The main feature of the seasonal pattern in the children series is a significant decrease in holiday months July and August.

Table 3 shows the parameter estimates obtained from a Gibbs sampler based on a five component mixture approximation in comparison to the maximum likelihood estimates. In the same table we study the effect of using less than five components in the mixture approximation to the distribution of ε_{ij}. Not surprisingly, we find a large improvement of choosing a mixture approximation with at least two components, rather than the one component normal approximation. The small differences between the different higher component estimates are only due to the Monte Carlo integration error.

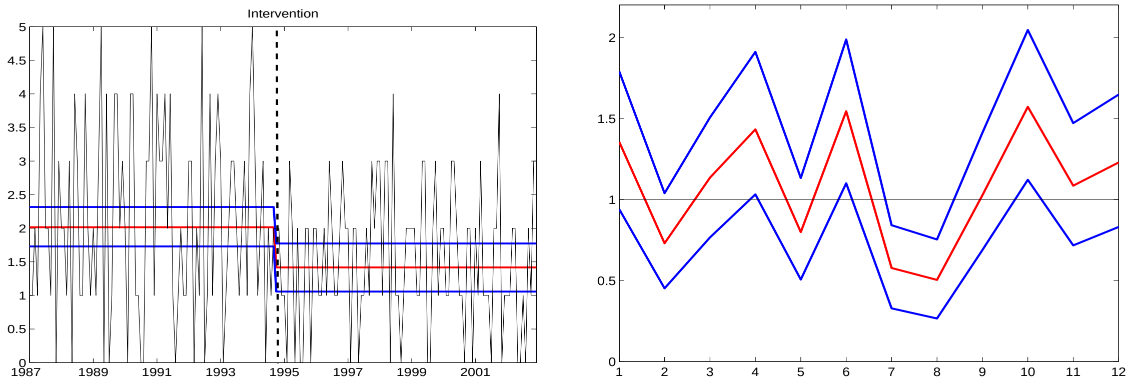

Figure 1: Counts of killed or injured children with estimated rate(posterior means, left) and seasonal pattern (posterior means,right) within 95% credible region

Table 3: Comparison of parameter estimates for different methods

		β		μ	
Method		Mean	Std.dev	Mean	Std.dev
Maximum Likelihood		0.7177	0.0730	-0.3543	0.1077
Gibbs Sampling	1 component	0.6938	0.0888	-0.4014	0.1307
	2 components	0.7096	0.0719	-0.3611	0.1055
	3 components	0.6978	0.0735	-0.3535	0.1070
	4 components	0.6971	0.0747	-0.3544	0.1087
	5 components	0.7028	0.0756	-0.3553	0.1080

4.2 Poisson Regression Models Dealing With Overdispersion

A common problem in Poisson regression models is the presence of overdispersion due to omitted covariates which may cause bias and loss of efficiency in estimating the remaining regression parameters, see for instance Cox (1983). A simple, but useful way of dealing with this kind of overdispersion is to introduce individual effects $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$ into the linear predictor:

$$y_i \sim \mathcal{P}(\lambda_i), \quad \log \lambda_i = \alpha_i + \mathbf{x}_i \boldsymbol{\beta}.$$

The Bayesian estimation methods discussed in Subsection 4.1 are easily extended to this more general setting. Data augmentations by $\boldsymbol{\tau}$ and $\mathbf{R}$ as described above leads to a normal individual effects regression model with $n_i = y_i + 1$ repeated measurements, and heteroscedastic errors with known variance. For each i, this model reads:

$$\log \tau_{ij} | \alpha_i, \boldsymbol{\beta}, r_{ij} = -\alpha_i - \mathbf{x}_i \boldsymbol{\beta} + m_{r_{ij}} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}\left(0, s_{r_{ij}}^2\right), \tag{10}$$

where $j = 1, \ldots, y_i + 1$. The joint conditional posterior $p(\boldsymbol{\vartheta}|\boldsymbol{\tau}, \mathbf{R}, \mathbf{y})$ of $\boldsymbol{\vartheta} = (\alpha_1, \ldots, \alpha_N, \boldsymbol{\beta})$ could easily be derived as in Subsection 4.1, however, joint sampling from this extremely high-dimensional normal density through the Cholesky decomposition of the corresponding variance-covariance matrix is not very efficient. We discuss an efficient multi move sampler.

4.2.1 Efficient multi move sampling

It is possible to write the joint conditional posterior of $\boldsymbol{\vartheta} = (\alpha_1, \ldots, \alpha_N, \boldsymbol{\beta})$ as

$$p(\boldsymbol{\vartheta}|\boldsymbol{\tau}, \mathbf{R}, \mathbf{y}) = p(\boldsymbol{\vartheta}|\boldsymbol{\tau}, \mathbf{R}) = p(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R}) \prod_{i=1}^{N} p(\alpha_i|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{R}).$$

$p(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R})$ is derived from a regression model, where the individual effects appearing in (10) are integrated out.

For each $i, i = 1, \ldots, N$, all inter arrival times τ_{ij}, generated for a specific count observation y_i, for $j = 1, \ldots, n_i$, where $n_i = (y_i + 1)$, share the same individual effect α_i and are independent only conditional on knowing α_i. The marginal model for $\log \boldsymbol{\tau}_i = (\log \tau_{i1} \ldots \log \tau_{i,n_i})'$ given only $\boldsymbol{\beta}$, with the individual effect being integrated out, is

$$\log \boldsymbol{\tau}_i = -\mathbf{1}\mathbf{x}_i \boldsymbol{\beta} + \boldsymbol{m}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{V}_i), \tag{11}$$

where $\boldsymbol{m}_i = (m_{r_{i1}} \cdots m_{r_{i,n_i}})'$ and

$$\boldsymbol{V}_i = \mathbf{1}\mathbf{1}'\sigma_\alpha^2 + \mathrm{Diag}\left(s_{r_{i1}}^2, \ldots, s_{r_{i,n_i}}^2\right).$$

From the marginal model (11) we obtain that the posterior $p(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R})$ is equal to a multivariate normal density $\mathcal{N}_d(\mathbf{b}_N, \mathbf{B}_N)$ with the posterior moments $\mathbf{b}_N$ and $\mathbf{B}_N$ given by

$$\mathbf{B}_N^{-1} = \mathbf{B}_0^{-1} + \sum_{i=1}^{N} \mathbf{x}_i^{'} \mathbf{1}' \boldsymbol{V}_i^{-1} \mathbf{1} \mathbf{x}_i, \qquad \mathbf{b}_N = \mathbf{B}_N (\mathbf{B}_0^{-1} \mathbf{b}_0 - \sum_{i=1}^{N} \mathbf{x}_i^{'} \mathbf{1}' \boldsymbol{V}_i^{-1} (\log \boldsymbol{\tau}_i - \boldsymbol{m}_i)). \tag{12}$$

For each $i = 1, \dots, N$, the conditional posterior $p(\alpha_i|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{R})$ is derived from regression model (10), with $\boldsymbol{\beta}$ considered to be known. From the results of Subsection 4.1, under the normal prior $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$, the conditional posterior $p(\alpha_i|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{R})$ is equal to a univariate normal density $\mathcal{N}(a_i(\boldsymbol{\beta}), A_i)$ with following posterior moments:

$$A_i^{-1} = \frac{1}{\sigma_\alpha^2} + \sum_{j=1}^{y_i+1} \frac{1}{s_{r_{ij}}^2}, \qquad a_i(\boldsymbol{\beta}) = A_i \left(\sum_{j=1}^{y_i+1} \frac{\log \tau_{ij} + \mathbf{x}_i\boldsymbol{\beta} - m_{r_{ij}}}{s_{r_{ij}}^2} \right). \tag{13}$$

Thus step (a) in the Gibbs sampler is implemented by sampling $\boldsymbol{\beta}$ from $\mathcal{N}_d(\mathbf{b}_N, \mathbf{B}_N)$, with the moments given by (12), and then sampling α_i from $\mathcal{N}(a_i(\boldsymbol{\beta}), A_i)$ for $i = 1, \dots, N$.

4.2.2 Dealing with unknown hyperparameters

If the hyperparameter σ_α^2 is unknown, then an additional step (c) has to be added, to sample σ_α^2 from the inverted Gamma posterior distribution $\mathcal{G}^{-1}(c_N/2, C_N/2)$ with $c_N = c_0 + N$, and $C_N = C_0 + \sum_{i=1}^N \alpha_i^2$, under the inverted Gamma prior distribution $\sigma_\alpha^2 \sim \mathcal{G}^{-1}(c_0/2, C_0/2)$.

4.2.3 Application to Simulated Data

For illustration, we consider two simulated data sets of size $N = 200$, generated according to the heterogeneity model

$$y_i \sim \mathcal{P}(\lambda_i), \quad \log \lambda_i = \alpha_i + \beta_0 + \beta_1 x_i,$$

where x_i corresponds to a linear trend and $\boldsymbol{\beta} = (\beta_0, \beta_1) = (0.5, -0.3)$. The variance of the random effects is $\sigma_\alpha^2 = 0.0001$ for the first data set, which implies a low degree of unobserved heterogeneity, whereas $\sigma_\alpha^2 = 1$ for the second data set, causing a high degree of unobserved heterogeneity.

A two-block Gibbs sampler based on the five component mixture approximation was used to estimate the unknown parameters. Conditional on knowing σ_α^2, we sample the whole parameter $\boldsymbol{\vartheta} = (\alpha_1, \dots, \alpha_N, \boldsymbol{\beta})$ jointly within one move using the results of Subsection 4.2.1. Conditional on knowing $\boldsymbol{\vartheta}$, the variance σ_α^2 is samples as described in Subsection 4.2.2. The Gibbs sampler was run 12000 times with a burn in of 2000 runs.

Estimation results are presented in Table 4, autocorrelation functions of the estimated values of $\boldsymbol{\beta}$ and σ_α^2 for both the low and high heterogeneity case are shown in Figure 2. As for Gaussian random-effects models, see for instance Gelfand et al. (1995) and van Dyk and Meng (2001), the convergence behavior of Gibbs sampling if worse in the low heterogeneity case (σ_α^2 close to 0), than for the high heterogeneity case ($\sigma_\alpha^2 = 1$).

5 Concluding Remarks

The Gibbs sampler studied in this paper provides an important step toward operational MCMC estimation for a broad class of regression models for Poisson counts,

Table 4: Parameter estimates for the heterogeneity model

	low heterogeneity		high heterogeneity	
	$\sigma^2_\alpha = 0.0001$		$\sigma^2_\alpha = 1$	
Parameter	Mean	Std.dev	Mean	Std.dev
β_0	0.5391	0.1196	0.4450	0.1895
β_1	-0.3035	0.1107	-0.3339	0.1659
σ^2_α	0.0070	0.0110	1.0020	0.2037

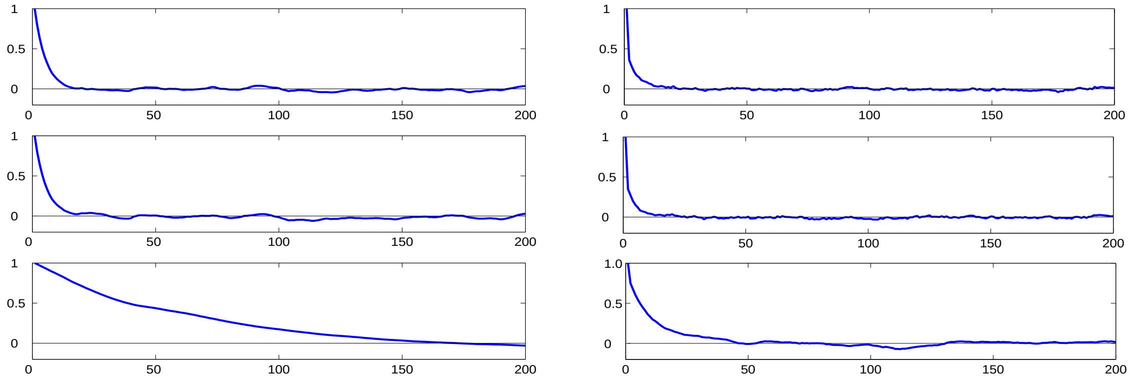

Figure 2: ACF of β_0 (above), β_1 (middle) and σ^2_α (below); low heterogeneity case (σ^2_α=0.01) left side, high heterogeneity case (σ^2_α=1) right side

as our sampler requires only draws from standard densities, without tuning of proposal densities. Gibbs sampling is feasible for most of the Poisson regression models suggested in the literature so far. The Gibbs sampler discussed in this paper for a standard regression model and a Poisson regression model with additive random effects is easily modified to deal with various extensions of the Poisson regression model, in particular with mixtures of Poisson regressions models (Wang et al. 1996; Hurn et al. 2003), panel Poisson regression models with random effects (Chib et al. 1998), and mixtures of Poisson regression models with random effects (Lenk and DeSarbo 2000). Space limits do not allows us to give all the details, which are, however, pretty straightforward, and are worked out in detail for time-varying Poisson regression models in Frühwirth-Schnatter and Wagner (2004).

Our new data augmentation scheme seems to be a promising step toward solving further issue in Bayesian estimation for generalized linear models based on the Poisson distribution. First, for random effects models some care must be exercised with respect to parameterization issues, as Gibbs sampling often leads to convergence problems, if σ^2_α is close to 0. Such problems are well-known for Gaussian random-effects model, see for instance Gelfand et al. (1995) and van Dyk and Meng (2001). For Poisson count data parameterization issues are also addressed in Chib et al. (1998). Our new data augmentation may be helpful in this regard, as a noncentered parameterization similar to the one studied in van Dyk and Meng (2001) for Gaussian models, is easily available.

Finally, we would like to mention that our data augmentation scheme leads to straightforward computation of marginal likelihoods based on the candidate's formula (Chib 1995).

Acknowledgement

We are grateful to Rudi Frühwirth for providing the normal mixture approximation to the log exponential distribution which is in fact an essential ingredient for our Gibbs sampling scheme. Furthermore we want to thank Stefan Höglinger from the KfV, the Austrian Road Safety Board, for providing the data analyzed in Subsection 4.

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing 6*, 251–262.

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

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

Chib, S., E. Greenberg, and R. Winkelmann (1998). Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics 86*, 33–54.

Chib, S., F. Nardari, and N. Shephard (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics 108*, 281–316.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika 70*, 269–274.

Frühwirth-Schnatter, S. and H. Wagner (2004). Gibbs sampling for parameter-driven models of time series of small counts with applications to state space modelling.

Gelfand, A., S. Sahu, and B. Carlin (1995). Efficient parametrisations for normal linear mixed models. *Biometrika 82*, 479–488.

Hurn, M., A. Justel, and C. P. Robert (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics 12*, 55–79.

Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies 65*, 361–393.

Lenk, P. J. and W. S. DeSarbo (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika 65*, 93–119.

McCullagh, P. and J. A. Nelder (1999). *Generalized linear models.* Chapman & Hall Ltd.

Robert, C. P. and G. Casella (1999). *Monte Carlo statistical methods.* Springer Series in Statistics. New York/Berlin/Heidelberg: Springer-Verlag Inc.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika 81*, 115–131.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*, 528–540.

van Dyk, D. and X.-L. Meng (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics 10*, 1–50.

Wang, P., M. L. Puterman, I. Cockburn, and N. Le (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics 52*, 381–400.

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

Zeger, S. L. and B. Qaqish (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics 44*, 1019–1031.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: John Wiley & Sons.