Stat Methods Appl
DOI 10.1007/s10260-011-0168-x

Sampling schemes for generalized linear Dirichlet process random effects models

Minjung Kyung · Jeff Gill · George Casella

Abstract We evaluate MCMC sampling schemes for a variety of link functions in generalized linear models with Dirichlet process random effects. First, we find that there is a large amount of variability in the performance of MCMC algorithms, with the slice sampler typically being less desirable than either a Kolmogorov–Smirnov mixture representation or a Metropolis–Hastings algorithm. Second, in fitting the Dirichlet process, dealing with the precision parameter has troubled model specifications in the past. Here we find that incorporating this parameter into the MCMC sampling scheme is not only computationally feasible, but also results in a more robust set of estimates, in that they are marginalized-over rather than conditioned-upon. Applications are provided with social science problems in areas where the data can be difficult to model, and we find that the nonparametric nature of the Dirichlet

This study was supported by National Science Foundation Grants DMS-0631632, SES-0631588, DMS-04-05543.

M. Kyung
Department of Statistics, Duksung Women's University, 19 Geunhwagyo-Gil, Dobong Gu, Seoul 132-714, Korea
e-mail: mkyung@duksung.ac.kr

J. Gill
Department of Political Science, Washington University, One Brookings Dr., Seigle Hall, St. Louis, MO, USA
e-mail: jgill@wustl.edu

J. Gill
Department of Biostatistics, Washington University, One Brookings Dr., Seigle Hall, St. Louis, MO, USA

G. Casella (✉)
Department of Statistics, University of Florida, Gainesville, FL 32611, USA
e-mail: casella@ufl.edu

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

process priors for the random effects leads to improved analyses with more reasonable inferences.

Keywords Linear mixed models · Generalized linear mixed models · Hierarchical models · Gibbs sampling · Metropolis–Hastings algorithm · Slice sampling

1 Introduction

Generalized linear models (GLMs) have enjoyed considerable attention over the years, providing a flexible framework for modeling discrete responses using a variety of error structures. If we have observations that are discrete or categorical, $\mathbf{y} = (\mathbf{y_1}, \dots, \mathbf{y_n})$, such data can often be assumed to be independent and from a distribution in the exponential family. The classic book by McCullagh and Nelder (1989) describes these models in detail; see also the more recent developments in Dey et al. (2000) or Fahrmeir and Tutz (2001).

A generalized linear *mixed* model (GLMM) is an extension of a GLM that allows random effects, and can give us flexibility in developing a more suitable model when the observations are correlated, or where there may be other underlying phenomena that contribute to the resulting variability. Thus, the GLMM can be specified to accommodate outcome variables conditional on mixtures of possibly correlated random and fixed effects (Breslow and Clayton 1993; Buonaccorsi 1996; Wang et al. 1998; Wolfinger and O'Connell 1993). Details of such models, covering both statistical inferences and computational methods, can be found in the texts by McCulloch and Searle (2001) and Jiang (2007).

1.1 Sampling schemes for GLMMs

There have been Markov chain Monte Carlo (MCMC) methods developed for the analysis of the GLMMs with random effects modeled with a normal distribution. Although the posteriors of parameters and the random effects are typically numerically intractable, especially when the dimension of the random effects is greater than one, there has been much progress in the development of sampling schemes. For example, Damien et al. (1999) proposed a Gibbs sampler using auxiliary variables for sampling nonconjugate and hierarchical models. Their methods are *slice sampling* methods derived from the full conditional posterior distribution. They mention that the assessment of convergence remains a major problem with the algorithm. However, Neal (2003) provided convergence properties of the posterior for slice sampling. Another sampling scheme was used by Chib et al. (1998) and Chib and Winkelmann (2001), who provided Metropolis–Hastings (M–H) algorithms for various kinds of GLMMs. They proposed a multivariate-t distribution as a candidate density in an M–H implementation, taking the mean equal to the posterior mode, and variance equal to the inverse of the Hessian evaluated at the posterior mode.

To be precise about language, we discuss three types of MCMC algorithms in this work. When we refer to the *slice sampler* we mean a Gibbs sampler on an enlarged state space (augmented by auxiliary variables). When we refer to a *Gibbs sampler*, it is a sampler based on producing automatically accepted candidate values from full conditional distributions that is not the special case of the slice sampler. When *Metropolis–Hastings* algorithms are discussed, these are not the special cases of Gibbs or slice sampling, but instead the more general process of producing candidate values from a separate distribution and deciding to accept them or not using the conventional Metropolis step.

1.2 Sampling schemes for GLMDMs

Another variation of a GLMM was used by Dorazio et al. (2007) and Gill and Casella (2009), where the random effects are modeled with a Dirichlet process, resulting in a Generalized Linear Mixed Dirichlet Process Model (GLMDM). Dorazio et al. (2007) used a GLMDM with a log link for spatial heterogeneity in animal abundance. They proposed an empirical Bayes approach with the Dirichlet process, instead of the regular assumption of normally distributed random effects, because they argued that for some species the sources of heterogeneity in abundance is poorly understood or unobservable. They noted that the Dirichlet process prior is robust to errors in model specification and allows spatial heterogeneity in abundance to be specified in a data-adaptive way. Gill and Casella (2009) suggested a GLMDM with an ordered probit link to model political science data, specifically modeling the stress, from public service, of Senate-confirmed political appointees as a reason for their short tenure. For the analysis, a semi-parametric Bayesian approach was adopted, using the Dirichlet process for the random effect.

Dirichlet process mixture models were introduced by Ferguson (1973) and Antoniak (1974), with further important developments in Blackwell and MacQueen (1973), Korwar and Hollander (1973), and Sethuraman (1994). For estimation, Lo (1984) derived the analytic form of a Bayesian density estimator, and Liu (1996) derived an identity for the profile likelihood estimator of the Dirichlet precision parameter. Kyung et al. (2010) looked at the properties of this MLE and found that the likelihood function can be ill-behaved. They noted that incorporating a gamma prior, and using posterior mode estimation, results in a more stable solution. McAuliffe et al. (2006) used a similar strategy, using a posterior mean for the estimation of the Dirichlet process precision parameter (the term m, which we describe in Sect. 2).

Models with Dirichlet process priors are treated as hierarchical models in a Bayesian framework, and the implementation of these models through Bayesian computation and efficient algorithms has had much attention. Escobar and West (1995) provided a Gibbs sampling algorithm for the estimation of posterior distribution for all model parameters, MacEachern and Müller (1998) presented a Gibbs sampler with non-conjugate priors by using auxiliary parameters, and Neal (2000) provided an extended and more efficient Gibbs sampler to handle general Dirichlet process mixture models. Teh et al. (2006) also extended the auxiliary variable method of Escobar and West (1995) for posterior sampling of the precision parameter with a gamma prior. They developed hierarchical Dirichlet processes, with a Dirichlet prior for the base measure.

Kyung et al. (2010) developed algorithms for estimation of the precision parameter and new MCMC algorithms for a linear mixed Dirichlet process random effects

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

models that had not previously existed. In addition, they showed how to extend the developed framework to a generalized Dirichlet process mixed model with a probit link function. They derived, for the first time, a simultaneous Gibbs sampler for all of the model parameters and the subclusters of the Dirichlet process, and used a new parameterization of the hierarchical model to derive a Gibbs sampler that more fully exploits the structure of the model and mixes very well. Finally they were also able to establish a proof that the proposed sampler is an improvement, in terms of operator norm and efficiency, over other commonly used algorithms.

1.3 Summary

In this paper we look at MCMC sampling schemes for generalized Dirichlet process mixture models, concentrating on logistic and log linear models. For these models, we examine a Gibbs sampling method using auxiliary parameters, based on Damien et al. (1999), and a Metropolis–Hastings sampler where the candidate generating distribution is a Gaussian density from log-transformed count data from a log-linear model (thus producing a form on the correct support). We incorporate the Dirichlet process precision parameter, m, into the Gibbs sampler, through the use of a gamma candidate distribution using a Laplace approximation for the calculation of the mean and variance of m, and use that in the gamma candidate. In the examples analyzed here, we find that the alternative slice sampler typically has higher autocorrelation in logistic regression and loglinear models than the proposed M–H algorithm.

Using the GLMDM with a general link function, Sect. 2 describes the generalized Dirichlet process mixture model. In Sect. 3 we estimate model parameters using a variety of algorithms, and Sect. 4 describes the estimation of the Dirichlet parameters. Section 5 looks at the performance of these algorithms in a variety of simulations, while Sect. 6 analyzes two social science data sets, further illustrating the advantage of the Dirichlet process random effects model. Section 7 summarizes these contributions and adds some perspective, and there is an Appendix with some technical details.

2 A generalized linear mixed Dirichlet process model

Let $\mathbf{X}_i$ be covariates associated with the ith observation, $\boldsymbol{\beta}$ be the coefficient vector, and ψ_i be a random effect accounting for subject-specific deviation from the underlying model. Assume that the $Y_i|\boldsymbol{\psi}$ are conditionally independent, each with a density from the exponential family, where $\boldsymbol{\psi} = (\psi_1, \dots, \psi_n)$. Then, based on the notation of McCulloch and Searle (2001), the GLMDM can be expressed as follows. Start with the generalized linear model,

$$Y_i|\gamma \overset{\text{ind}}{\sim} f_{Y_i|\gamma}(y_i|\gamma), \quad i = 1, \dots, n$$

$$f_{Y_i|\gamma}(y_i|\gamma) = \exp\left[\{y_i\gamma_i - b(\gamma_i)\}/\xi^2 - c(y_i, \xi)\right]. \tag{1}$$

where y_i is discrete valued. Here, we know that $\mathrm{E}[Y_i|\gamma] = \mu_i = \partial b(\gamma_i)/\partial\gamma_i$. Using a link function $g(\cdot)$, we can express the transformed mean of Y_i, $\mathrm{E}[Y_i|\gamma]$, as a linear function, and we add a random effect to create the mixed model:

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

$$g(\mu_i) = \mathbf{X}_i \boldsymbol{\beta} + \psi_i. \tag{2}$$

Here, for the Dirichlet process mixture models, we assume that

$$\begin{aligned} &\psi_i \sim G \\ &G \sim \mathcal{DP}(mG_0), \end{aligned} \tag{3}$$

where $\mathcal{DP}$ is the Dirichlet process with base measure G_0 and precision parameter m. Blackwell and MacQueen (1973) proved that for $\psi_1, \ldots, \psi_n$ iid from $G \sim \mathcal{DP}$, the joint distribution of $\boldsymbol{\psi}$ is a product of successive conditional distributions of the form:

$$\psi_i | \psi_1, \ldots, \psi_{i-1}, m \sim \frac{m}{i-1+m} g_0(\psi_i) + \frac{1}{i-1+m} \sum_{l=1}^{i-1} \delta(\psi_l = \psi_i) \tag{4}$$

where $\delta()$ denotes the Dirac delta function and $g_0()$ is the density function of the base measure.

We define a *partition* C to be a clustering of the sample of size n into k groups, $k = 1, \ldots, n$, and we call these subclusters since the grouping is done nonparametrically rather than on substantive criteria. That is, the partition assigns different distributional parameters across groups and the same parameters within groups; cases are iid only if they are assigned to the same subcluster.

Applying Lo (1984) Lemma 2 and Liu (1996) Theorem 1 to (4), we can calculate the likelihood function, which by definition is integrated over the random effects, as

$$L(\boldsymbol{\beta} \mid \mathbf{y}) = \frac{\Gamma(m)}{\Gamma(m+n)} \sum_{k=1}^{n} m^k \sum_{C:|C|=k} \prod_{j=1}^{k} \Gamma(n_j) \int f(\mathbf{y}_{(j)} \mid \boldsymbol{\beta},\ \psi_j) \mathrm{d}G_0(\psi_j),$$

where C defines the partition of subclusters of size n_j, $|C|$ indicates occupied subclusters, $\mathbf{y}_{(j)}$ is the vector of y_is that are in subcluster j, and ψ_j is the common parameter for that subcluster. There are $\mathcal{S}_{n,k}$ different partitions C, the Stirling Number of the Second Kind (Abramowitz and Stegun 1972, 824–825).

Here, we consider an $n \times k$ matrix A defined by

$$A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

where each a_i is a $1 \times k$ vector of all zeros except for a 1 in the position indicating which group the observation is from. Thus, A represents a partition of the sample of size n into k groups, with the column sums giving the subcluster sizes. Note that both the dimension k, and the placement of the 1s, are random, representing the subclustering process.

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

If the partition C has subclusters $\{S_1, \ldots, S_k\}$, then if $i \in S_j$, $\psi_i = \eta_j$ and the random effect can be rewritten as

$$\boldsymbol{\psi} = A\boldsymbol{\eta}, \tag{5}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$ and $\eta_j \overset{iid}{\sim} G_0$ for $j = 1, \ldots, k$. This is the same representation of the Dirichlet process that was used in Kyung et al. (2010), building on the representation in McCullagh and Yang (2006).

In this paper, we consider models for the binary responses with probit and logit link function, and for count data with a log link function. First, for the binary responses,

$$Y_i \sim \text{Bernoulli}(p_i), \quad i = 1, \ldots, n$$

where y_i is 1 or 0, and $p_i = \mathrm{E}(\mathrm{Y_i})$ is the probability of a success for the ith observation. Using a general link function (2) leads to a sampling distribution for the observed outcome variable $\mathbf{y}$:

$$f(\mathbf{y}|A) = \int \prod_{i=1}^{n} \left[g^{-1}(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right]^{y_i} \left[1 - g^{-1}(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right]^{1-y_i} \mathrm{d}G_0(\boldsymbol{\eta}),$$

which typically can only be evaluated numerically. Examples of general link functions for binary outcomes are

$$\begin{aligned}
p_i &= g_1^{-1}(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i) = \Phi(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i) && \text{Probit}\\
p_i &= g_2^{-1}(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i) = (1 + \exp(-\mathbf{X}_i\boldsymbol{\beta} - (\mathbf{A}\boldsymbol{\eta})_i))^{-1} && \text{Logistic}\\
p_i &= g_3^{-1}(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i) = 1 - \exp\left(-\exp(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right) && \text{Cloglog}
\end{aligned}$$

where $\Phi()$ is the cumulative distribution function of a standard normal distribution.

For counting process data,

$$Y_i \sim \text{Poisson}(\lambda_i), \quad i = 1, \ldots, n$$

where y_i is $0, 1, \ldots, \lambda_i = \mathrm{E}(\mathrm{Y_i})$ is the expected number of events for the ith observation. Here, using a log link function

$$\log(\lambda_i) = \mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i,$$

the sampling distribution of $\mathbf{y}$ is

$$f(\mathbf{y}|A) = \prod_{i=1}^{n} \frac{1}{y_i!} \int \prod_{i=1}^{n} \exp\{-\exp(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\} \left[\exp(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right]^{y_i} G_0(\boldsymbol{\eta})\, d\boldsymbol{\eta}.$$

For the base measure of the Dirichlet process, we assume a normal distribution with mean 0 and variance τ^2, $N(0, \tau^2)$. In our experience, the model is not sensitive to this distributional assumption and others, such as the student's-t, could be used.

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

3 Sampling schemes for the model parameters

An overview of the general sampling scheme is as follows. We have three groups of parameters:

(i) m, the precision parameter of the Dirichlet process,
(ii) $\mathbf{A}$, the indicator matrix of the partition defining the subclusters, and
(iii) $(\boldsymbol{\eta}, \boldsymbol{\beta}, \tau^2)$, the model parameters.

We iterate between these three groups until convergence:

1. Conditional on m and A, generate $(\boldsymbol{\eta}, \boldsymbol{\beta}, \tau^2)|\mathbf{A}, m$;
2. Conditional on $(\boldsymbol{\eta}, \boldsymbol{\beta}, \tau^2)$ and m, generate A, a new partition matrix.
3. Conditional on $(\boldsymbol{\eta}, \boldsymbol{\beta}, \tau^2)$ and A, generate m, the new precision parameter.

For the model parameters we add the priors

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2 &\sim N(\mathbf{0}, d^*\sigma^2 I) \\ \tau^2 &\sim \text{Inverted Gamma}(a, b), \end{aligned} \tag{6}$$

where $d^* > 1$ and (a, b) are fixed such that the inverse gamma is diffuse ($a = 1, b$ very small). Thus the partitioning in the algorithm assigns different normal parameters across groups and the same normal parameters within groups. For the Dirichlet process we need the previously stated priors

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k) \quad \text{and} \quad \eta_j \overset{iid}{\sim} G_0 \quad \text{for} \quad j = 1, \ldots, k. \tag{7}$$

We can either fix σ^2 or put a prior on it and estimate it in the hierarchical model with priors; here we will fix a value for σ^2.

In the following sections we consider a number of sampling schemes for the estimation of the model parameters of a GLMDM. We will then turn to generation of the subclusters and the precision parameter.

3.1 Probit models

Albert and Chib (1993) showed how truncated normal sampling could be used to implement the Gibbs sampler for a probit model for binary responses. They use a latent variable V_i such that

$$V_i = \mathbf{X}_i\boldsymbol{\beta} + \psi_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{8}$$

and

$$y_i = 1 \quad \text{if } V_i > 0 \quad \text{and} \quad y_i = 0 \quad \text{if } V_i \leq 0$$

for $i = 1, \ldots, n$. It can be shown that Y_i are independent Bernoulli random variables with the probability of success, $p_i = \Phi((\mathbf{X}_i\boldsymbol{\beta} - \psi_i)/\sigma)$, and without loss of generality, we fix $\sigma = 1$.

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

Details of implementing the Dirichlet process random effect probit model are given in Kyung et al. (2010) and will not be repeated here. We will use this model for comparison, but our main interest is in logistic and loglinear models.

3.2 Logistic models

We look at two samplers for the logistic model. The first is based on the slice sampler of Damien et al. (1999), while the second exploits a mixture representation of the logistic distribution; see Andrews and Mallows (1974) or West (1987).

3.2.1 Slice sampling

The idea behind the slice sampler is the following. Suppose that the density $f(\theta) \propto L(\theta)\pi(\theta)$, where $L(\theta)$ is the likelihood and $\pi(\theta)$ is the prior, and it is not possible to sample directly from $f(\theta)$. Using a latent variable U, define the joint density of θ and U by

$$f(\theta, u) \propto I\{u < L(\theta)\}\pi(\theta).$$

Then, $U|\theta$ is uniform $\mathcal{U}\{0, L(\theta)\}$, and $\theta|U = u$ is π restricted to the set $A_u = \{\theta : L(\theta) > u\}$.

The likelihood function of binary responses with logit link function can be written as

$$L_k(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}|A, \mathbf{y}) = \prod_{i=1}^{n}\left[\frac{1}{1+\exp(-\mathbf{X}_i\boldsymbol{\beta}-(\mathbf{A}\boldsymbol{\eta})_i)}\right]^{y_i}\left[\frac{1}{1+\exp(\mathbf{X}_i\boldsymbol{\beta}+(\mathbf{A}\boldsymbol{\eta})_i)}\right]^{1-y_i}$$
$$\times\prod_{j=1}^{k}\left(\frac{1}{2\pi\tau^2}\right)^{1/2}\exp\left(-\frac{1}{2\tau^2}\eta_j^2\right), \tag{9}$$

and if we introduce latent variables $\mathbf{U} = (U_1, \ldots, U_n)$ and $\mathbf{V} = (V_1, \ldots, V_n)$, we have the likelihood of the model parameters and the latent variables to be

$$L_k(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}|A, \mathbf{y})$$
$$= \prod_{i=1}^{n} I\left[u_i < \left\{\frac{1}{1+\exp(-\mathbf{X}_i\boldsymbol{\beta}-(\mathbf{A}\boldsymbol{\eta})_i)}\right\}^{y_i},\ v_i < \left\{\frac{1}{1+\exp(\mathbf{X}_i\boldsymbol{\beta}+(\mathbf{A}\boldsymbol{\eta})_i)}\right\}^{1-y_i}\right]$$
$$\times\prod_{j=1}^{k}\left(\frac{1}{2\pi\tau^2}\right)^{1/2}\exp\left(-\frac{1}{2\tau^2}\eta_j^2\right) \tag{10}$$

Thus, with priors that are given above, the joint posterior distribution of $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V})$ can be expressed as

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

$$\pi_k(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}|A, \mathbf{y}) \propto L_k(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}|A, \mathbf{y}) \times \left(\frac{1}{\tau^2}\right)^{a+1} \exp\left(-\frac{b}{\tau^2}\right) \exp\left(-\frac{|\boldsymbol{\beta}|^2}{2d^*\sigma^2}\right). \quad (11)$$

Then for fixed m and A, we can implement a Gibbs sampler using the full conditionals. Details are discussed in Appendix A.1.

3.2.2 A mixture representation

Next we consider a Gibbs sampler using truncated normal variables in a manner that is similar to the Gibbs sampler of the probit models, which arise from a mixture representation of the logistic distribution. Andrews and Mallows (1974) discussed necessary and sufficient conditions under which a random variable Y may be generated as the ratio Z/V where Z and V are independent and Z has a standard normal distribution, and establish that when $V/2$ has the asymptotic distribution of the Kolmogorov distance statistic, Y is logistic. West (1987) generalized this result to the exponential power family of distributions, showing these distributional forms to be a subset of the class of scale mixtures of normals. The corresponding mixing distribution is explicitly obtained, identifying a close relationship between the exponential power family and a further class of normal scale mixtures, the stable distributions.

Based on Andrews and Mallows (1974), and West (1987), the logistic distribution is a scale mixture of a normal distribution with a Kolmogorov–Smirnov distribution. From Devroye (1986), the Kolmogorov–Smirnov (K–S) density function is given by

$$f_X(x) = 8\sum_{\alpha=1}^{\infty}(-1)^{\alpha+1}\alpha^2 x e^{-2\alpha^2x^2} \quad x \geq 0, \quad (12)$$

and we define the joint distribution

$$f_{Y,X}(y,x) = (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{y}{2x}\right)^2\right\} f_X(x)\frac{1}{2x}. \quad (13)$$

From the identities in Andrews and Mallows (1974) (see also Theorem 10.2.1 in Balakrishnan 1992), the marginal distribution of Y is then given by

$$f_Y(y) = \int_0^{\infty} f_{Y,X}(y,x)\mathrm{d}x = \sum_{\alpha=1}^{\infty}(-1)^{\alpha+1}\alpha \exp\left(-\alpha|y|\right) = \frac{e^{-y}}{\left(1+e^{-y}\right)^2}, \quad (14)$$

the density function of logistic distribution with mean 0 and variance $\frac{\pi^2}{3}$. Therefore, $Y \sim \Lambda\left(0, \frac{\pi^2}{3}\right)$, where $\Lambda()$ is the logistic distribution.

Journal: 10260 Article No.: 0168 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

Now, using the likelihood function of binary responses with logit link function (9), consider the latent variable W_i such that

$$W_i = \mathbf{X}_i \boldsymbol{\beta} + \psi_i + \boldsymbol{\eta}_i, \quad \boldsymbol{\eta}_i \sim \Lambda\left(0, \frac{\pi^2}{3}\sigma^2\right), \tag{15}$$

with $y_i = 1$ if $W_i > 0$ and $y_i = 0$ if $W_i \leq 0$, for $i = 1, \ldots, n$. It can be shown that Y_i are independent Bernoulli random variables with $p_i = [1+\exp(-\mathbf{X}_i\boldsymbol{\beta} - (\mathbf{A}\boldsymbol{\eta})_i)]^{-1}$, the probability of success, and without loss of generality we fix $\sigma = 1$.

For given A, the likelihood function of model parameters and the latent variable is given by

$$\begin{aligned} L_k(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}|A, \mathbf{y}, \sigma^2) = &\prod_{i=1}^{n} \{I(U_i > 0)I(y_i = 1) + I(U_i \leq 0)I(y_i = 0)\} \\ &\times \int_0^\infty \left(\frac{1}{2\pi\sigma^2(2\xi)^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2(2\xi)^2}|\mathbf{U}-\mathbf{X}\boldsymbol{\beta}-A\boldsymbol{\eta}|^2} \\ &\times 8\sum_{\alpha=1}^{\infty}(-1)^{\alpha+1}\alpha^2\xi e^{-2\alpha^2\xi^2}\mathrm{d}\xi \left(\frac{1}{2\pi\tau^2}\right)^{k/2} e^{-\frac{1}{2\tau^2}|\boldsymbol{\eta}|^2}, \end{aligned}$$

where $\mathbf{U} = (U_1, \ldots, U_n)$, and U_i is the truncated normal variable which is described in (8).

Let m and A be considered fixed for the moment. Thus, with priors given in (6) and (7), the joint posterior distribution of $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U})$ given the outcome $\mathbf{y}$ is

$$\pi_k^L \propto L_k(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}|A, \mathbf{y}, \sigma^2) e^{-\frac{1}{2d^*\sigma^2}|\boldsymbol{\beta}|^2} \left(\frac{1}{\tau^2}\right)^{a+1} e^{-\frac{b}{\tau^2}}.$$

This representation avoids the problem of generating samples from the truncated logistic distribution, which is not easy to implement. As we now have the logistic distribution expressed as a normal mixture with the K–S distribution, we now only need to generate samples from the truncated normal distribution and the K–S distribution, and we can get a Gibbs sampler for the model parameters. The details are left to Appendix A.1.2.

3.3 Log linear models

Similar to Sect. 3.2, we look at two samplers for the loglinear model. The first is again based on the slice sampler of Damien et al. (1999), while the second is an M–H algorithm based on using a Gaussian density from log-transformed data as a candidate.

3.3.1 Slice sampling

The likelihood function of the counting process data with log link function can be written as

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

$$L_k(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}|A, \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{y_i!} e^{-\exp(\mathbf{X}_i\boldsymbol{\beta}+(\mathbf{A}\boldsymbol{\eta})_i)} \left[\exp(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right]^{y_i}$$
$$\times \prod_{j=1}^{k} \left(\frac{1}{2\pi\tau^2}\right)^{1/2} \exp\left(-\frac{1}{2\tau^2}\eta_j^2\right), \tag{16}$$

and the joint posterior distribution of $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta})$ can be obtained by appending the priors for τ^2 and $\boldsymbol{\beta}$. As in Sect. 3.2.1 we introduce latent variables $\mathbf{U} = (U_1, \ldots, U_n)$ and $\mathbf{V} = (V_1, \ldots, V_n)$, yielding a likelihood of the model parameters and the latent variables, $L_k(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}|A, \mathbf{y})$, similar to (10). Setting up the Gibbs sampler is now straightforward, with details in Appendix A.2.1.

3.3.2 Metropolis–Hastings

The primary challenge in setting up an efficient Metropolis–Hastings algorithm is specifying practical candidate generating functions for each of the unknown parameters in the sampler. This involves both stipulating a distributional form close to the target *and* variances that provide a reasonable acceptance rate. Starting with the likelihood and priors described at (16), for the candidate distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, we consider the model:

$$\log(Y_i) = \mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2).$$

which is a linear mixed Dirichlet process model (LMDPM). Sampling these model parameters is straightforward, and this enables us to have high-quality candidate values for the accept/reject stage of the Metropolis–Hastings algorithm for the log linear setup here. Using a similar model with the same parameter support but different link function as a way to generate M–H candidate values is a standard trick in the MCMC literature (Robert and Casella 2004). Details about this process are provided in Appendix A.2.2.

3.3.3 Comparing slice sampling to Metropolis–Hastings

In a special case it is possible to directly compare slice sampling and independent Metropolis–Hastings. If we have a Metropolis–Hastings algorithm with target density π and candidate h, we can compare it to the slice sampler

$$U|X = x \sim \text{Uniform}\{u : 0 < u < \pi(x)/h(x)\},$$
$$X|U = u \sim h(x)\{x : 0 < u < \pi(x)/h(x)\}.$$

In this setup Mira and Tierney (2002) show that the slice sampler dominates the Metropolis–Hastings algorithm in the efficiency ordering, meaning that all asymptotic variances are smaller, as well as first-order covariances.

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

At first look this result seems to be in opposition with what we will see in Sect. 5; we find that Metropolis–Hastings outperforms slice sampling with respect to autocorrelations. The resolution of this discrepancy is simple; the Mira-Tierney result applies when slice sampling and Metropolis–Hastings have the relationship described above—the candidate densities must be the same. In practice, and in the examples that we will see, the candidates are chosen in each case based on ease of computation, and in the case of the Metropolis–Hastings algorithm, to try to mimic the target. Under the demanding circumstances required of our Metropolis–Hastings algorithm for the real-world data and varied link functions used, it would be a very difficult task to produce candidate generating distributions that might match a slice sampler.

As an illustration of where we can actually match candidate generating distributions, consider the parameterization of Mira and Tierney (2002), where

$$\pi(x) = e^{-x} \quad \text{and} \quad h(x) = qe^{-qx}, \quad 0 < q < 1. \tag{17}$$

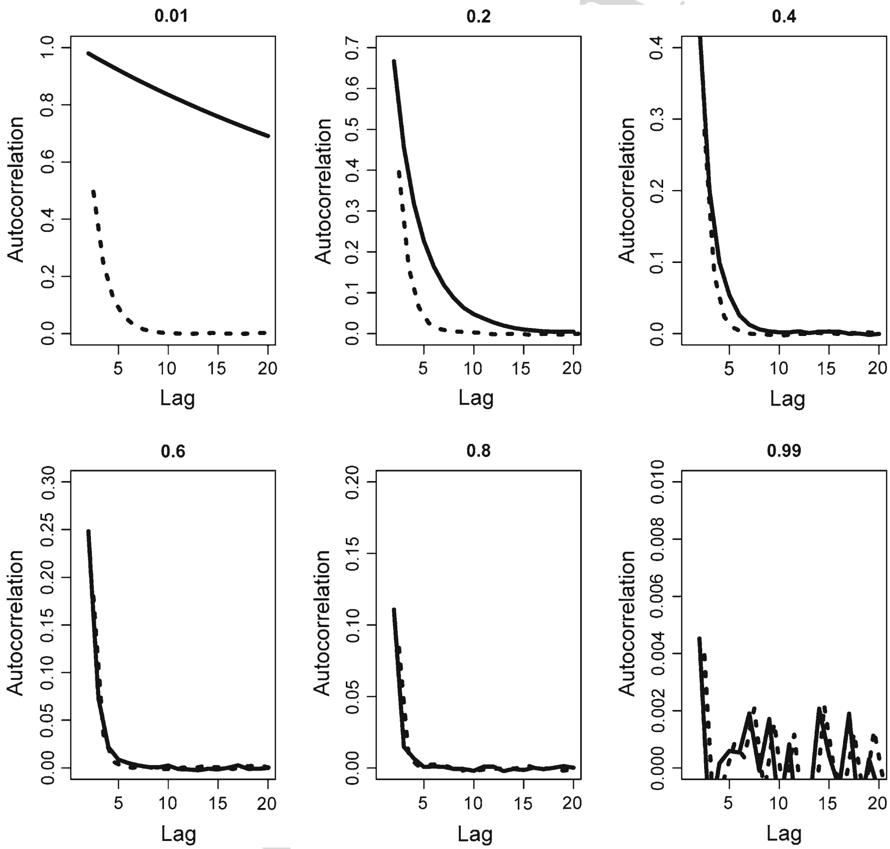

Fig. 1 Autocorrelations for both the slice sampler (*dashed*) and the Metropolis–Hastings algorithm (*solid*), for different values of q, for the model in (17). Note that the *panels* have different scales on the y-axis

Journal: **10260** Article No.: **0168** TYPESET DISK LE CP Disp.:**2011/10/22** Pages: **32** Layout: **Small-X**

If both slice and Metropolis–Hastings use the same value of q, then the slice sampler dominates. But if the samplers use different values of q, it can be the case that Metropolis–Hastings dominates the slice sampler. This is illustrated in Fig. 1, where we show the autocorrelations for both the slice sampler and the Metropolis–Hastings algorithm, for different values of q. Compare Metropolis–Hastings with large values of q, where the candidate gets closer to the target, with a slice sampler having a smaller value of q (Note that the different plots have different scales). We see that in these cases the Metropolis–Hastings algorithm can dominate the slice sampler.

4 Sampling schemes for the Dirichlet process parameters

4.1 Generating the partitions

We use a Metropolis–Hastings algorithm with a candidate taken from a multinomial/Dirichlet. This produces a Gibbs sampler that converges faster than the popular "stick-breaking" algorithm of Ishwaran and James (2001). See Kyung et al. (2010) for details on comparing stick-breaking versus "restaurant" algorithms.

For $t = 1, \dots T$, at iteration t

1. Starting from $(\boldsymbol{\theta}^{(t)}, \mathbf{A}^{(t)})$,

$$\boldsymbol{\theta}^{(t+1)} \sim \pi(\boldsymbol{\theta} \mid \mathbf{A}^{(t)}, \mathbf{y}),$$

   where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta})$ and the updating methods are discussed above.

2. If $\boldsymbol{q} = (q_1, \dots, q_n) \sim \text{Dirichlet}(r_1, \dots, r_n)$, then for any k and $k+1 \le n$

$$\mathbf{q}^{(t+1)} = \left(q_1^{(t+1)}, \dots q_n^{(t+1)}\right) \sim \text{Dirichlet}\left(n_1^{(t)} + r_1, \dots, n_k^{(t)} + r_k, r_{k+1}, \dots, r_n\right) \tag{18}$$

3. Given $\boldsymbol{\theta}^{(t+1)}$,

$$\mathbf{A}^{(t+1)} \sim P(\mathbf{A}) f(\mathbf{y}|\boldsymbol{\theta}^{(t+1)}, \mathbf{A}) \binom{n}{n_1 \cdots n_k} \prod_{j=1}^{k} \left[q_j^{(t+1)}\right]^{n_j} \tag{19}$$

where $\mathbf{A}$ is $n \times k$ with column sums $n_j > 0$, $n_1 + \dots + n_k = n$.

Based on the value of the $q_j^{(t+1)}$ in (18) we generate a candidate $\mathbf{A}$ that is an $n \times n$ matrix where each row is a multinomial, and the effective dimension of the matrix, the size of the partition, k, are the non-zero column sums. Deleting the columns with column sum zero is a marginalization of the multinomial distribution. The probability of the candidate is given by

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

$$P\left(\mathbf{A}^{(t+1)}\right) = \frac{\Gamma\left(\sum_{j=1}^{n} r_j\right)}{\prod_{j=1}^{k^{(t+1)}-1} \Gamma(r_j)\Gamma\left(\sum_{j=k^{(t+1)}}^{n} r_j\right)}$$

$$\times \frac{\prod_{j=1}^{k^{(t+1)}-1} \Gamma\left(n_j^{(t+1)} + r_j\right) \Gamma\left(n_{k^{(t+1)}}^{(t+1)} + \sum_{j=k^{(t+1)}}^{n} r_j\right)}{\Gamma\left(n + \sum_{j=1}^{n} r_j\right)}$$

and a Metropolis–Hastings step is then done.

4.2 Gibbs sampling the precision parameter

To estimate the precision parameter of the Dirichlet process, m, we start with the profile likelihood,

$$L(m \mid \boldsymbol{\theta}, \mathbf{A}, \mathbf{y}) = \frac{\Gamma(m)}{\Gamma(m+n)} m^k \prod_{j=1}^{k} \Gamma(n_j) f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{A}). \tag{20}$$

Rather than estimating m, a better strategy is to include m directly in the Gibbs sampler, as the maximum likelihood estimate from (20) can be very unstable (Kyung et al. 2010). Using the prior $g(m)$ we get the posterior density

$$\pi(m \mid \boldsymbol{\theta}, \mathbf{A}, \mathbf{y}) = \frac{\frac{\Gamma(m)}{\Gamma(m+n)} g(m) m^k}{\int_0^\infty \frac{\Gamma(m)}{\Gamma(m+n)} g(m) m^k \mathrm{d}m}, \tag{21}$$

where $\int \pi(m \mid \boldsymbol{\theta}, \mathbf{A}, \mathbf{y})\ dm < \infty$ must be finite for this to be proper. Note also how far removed m is from the data, as the posterior only depends on the number of groups k. We consider a gamma distribution as a prior, $g(m) = m^{a-1} e^{-m/b} / \Gamma(a) b^a$, and generate m using an M–H algorithm with another gamma density as a candidate.

We choose the gamma candidate by using a approximate mean and variance of $\pi(m)$ to set the parameters of the candidate. To get the approximate mean and variance, we will use the Laplace approximation of Tierney and Kadane (1986). Applying their results and using the log-likelihood, $\ell()$ in place of the likelihood, $L()$, we have:

$$\frac{\int m^\nu \frac{\Gamma(m)}{\Gamma(m+n)} g(m) m^k \mathrm{d}m}{\int \frac{\Gamma(m)}{\Gamma(m+n)} g(m) m^k \mathrm{d}m} \approx \sqrt{\frac{\ell''(\hat{m})}{\ell''_\nu(\hat{m}_\nu)}} \exp\left\{n\left[\ell_\nu(\hat{m}_\nu) - \ell(\hat{m})\right]\right\}, \tag{22}$$

where

$$\ell = \log \frac{m^{a-1} e^{-m/b}}{\Gamma(a) b^a} + \frac{1}{n}\left\{\log \frac{\Gamma(m)}{\Gamma(m+n)} + k \log m\right\}$$

$$\ell_\nu = \ell + \nu \log m$$

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

$$\ell' = \frac{\partial}{\partial m}\ell = \frac{1}{bm}\left[b\left(\frac{k}{n}+a-1\right) - m - \frac{bm}{n}\sum_{i=1}^{n}\frac{1}{m+i-1}\right]$$

$$\ell''(\hat{m}) = \left.\frac{\partial^2}{\partial m^2}\ell\right|_{m=\hat{m}} = \frac{1}{\hat{m}}\left[-\frac{1}{\hat{m}}\left(\frac{k}{n}+a-1\right) + \frac{\hat{m}}{n}\sum_{i=1}^{n}\frac{1}{\left(\hat{m}+i-1\right)^2}\right]$$

$$\ell_\nu{}' = \ell' + \frac{\nu}{m}, \quad \ell_\nu{}''(\hat{m}_\nu) = \left.\frac{\partial^2}{\partial m^2}\ell_\nu\right|_{m=\hat{m}_\nu} = \ell''(\hat{m}_\nu) - \frac{\nu}{\hat{m}_\nu^2}$$

where we get a simplification because the second derivative is evaluated at the zero of the first derivative. We use these approximations as the first and second moments of the candidate gamma distribution. Note that if $\hat{m} \approx \hat{m}_\nu$, then a crude approximation, which should be enough for Metropolis–Hastings, is $\mathrm{E}m^\nu \approx (\hat{m})^\nu$.

5 Simulation study

We evaluate our sampler through a number of simulation studies. We need to generate outcomes from Bernoulli or Poisson distributions with random effects that follow the Dirichlet process. To do this we fix K, the true number of clusters (which is unknown in actual circumstances), then we set the parameter m according to the relation

$$K = \sum_{i=1}^{n}\frac{m}{m+i-1}, \tag{23}$$

where we note that even if $\hat{m}$ is quite variable, there is less variability in $\hat{K} = \sum_{i=1}^{n}\frac{\hat{m}}{\hat{m}+i-1}$. When we integrate over the Dirichlet process (as done algorithmically according to Blackwell and McQueen 1973), the right-hand-side of (23) is the expected number of clusters, given the prior distribution on m. Neal (2000, p. 252) shows this as the probability in the limit, of a unique table seating, conditional on the previous table seatings, which makes intuitive sense since this expectation depends on individuals sitting at unique tables to start a new (sub)cluster in the algorithm.

5.1 Logistic models

Using the GLMDM with the logistic link function of Sect. 3.2, we set the parameters: $n = 100, K = 40, \tau^2 = 1$, and $\boldsymbol{\beta} = (1, 2, 3)$. Our Dirichlet process for the random effect has precision parameter m and base distribution $G_0 = N(0, \tau^2)$. Setting $K = 40$, yields $m = 24.21$. We then generated X_1 and X_2 independently from $N(0, 1)$, and used the fixed design matrix to generate the binary outcome Y. Then the Gibbs sampler was iterated 200 times to get values of $m, A, \boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}$. This procedure was repeated 1,000 times saving the last 500 draws as simulations from the posterior.

We compare the slice sampler (**Slice**) to the Gibbs sampler with the K–S distribution normal scale mixture (**K–S Mixture**) with the prior distribution of $\boldsymbol{\beta}$ from

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

Table 1 Estimation of the coefficients of the GLMDM with logistic link function and the estimate of K, with true values $K = 40$ and $\boldsymbol{\beta} = (1, 2, 3)$

Estimation method	β_0	β_1	β_2	K
Slice	2.2796 (0.4628)	3.2709 (0.5558)	4.7529 (0.7208)	43.0423 (4.2670)
K–S mixture	0.4900 (0.2024)	1.0494 (0.2468)	1.7787 (0.2491)	43.4646 (4.0844)

Standard errors are in parentheses

$\boldsymbol{\beta}|\sigma^2 \sim N\left(\mu\mathbf{1}, d^*\sigma^2 I\right)$ and $\mu \sim \pi(\mu) \propto c$, a flat prior for μ. For the estimation of K, we use the posterior mean of m, $\hat{m}$ and calculate $\hat{K}$ by using Eq. (23). The starting points of $\boldsymbol{\beta}$ come from the maximum likelihood (ML) estimates using iteratively reweighted least squares. All summaries in the tables are posterior means and standard deviations calculated from the empirical draws of the chain in its apparent converged (stationary) distribution.

The numerical summary of this process is given in Table 1. The estimates of K were 43.0423 with standard error 4.2670 from **Slice** and 43.4646 with standard error 4.0844 from **K–S Mixture**. Obviously these turned out to be good estimates of the true $K = 40$. The estimate of $\boldsymbol{\beta}$ with **K–S Mixture** is closer to the true value than those with **Slice**, with smaller standard deviation. To evaluate the convergence of $\boldsymbol{\beta}$, we consider the autocorrelation function (ACF) plots that are given in Fig. 2. The Gibbs sampler of $\boldsymbol{\beta}$ from **Slice** exhibits strong autocorrelation, implying poor mixing.

5.2 Log linear models

We now look at the GLMDM with the log link function of Sect. 3.3. The setting for the data generation is the same as the procedure that we discussed in the previous section except that we take $\boldsymbol{\beta} = (3, 0.5, 1)$. With $K = 40$, the solution of m from Eq. (23) is 24.21. As before, we generated X_1 and X_2 independently from $N(0, 1)$, and used the fixed design matrix to generate count data Y. The Gibbs sampler was iterated 200 times to produce draws of $m, A, \boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}$. This procedure was repeated 1,000 times, saving the last 500 values as draws from the posterior.

In this section, we compare the Gibbs sampler with the auxiliary variables (**Slice**) and the M–H sampler with a candidate density from the log-linear model (**M–H Sampler**). We use the posterior mean of m, $\hat{m}$, and calculate $\hat{K}$ by using (23) for the estimation of K. The starting points of $\boldsymbol{\beta}$ are set to the maximum likelihood (ML) estimates by using iterative reweighted least squares. The numerical summary is given in Table 2 and the ACF plots of $\boldsymbol{\beta}$ are given in Fig. 3. The resulting estimates for K are 43.5188(4.1398) from **Slice** and 43.516(4.1274) from the **M–H Sampler**, which are fairly close to the true $K = 40$. The estimated βs from the **M–H Sampler**, while not right on target, are much better than that of the slice sampler which, by standard diagnostics, has not yet converged. Once again, the consecutive draws of $\boldsymbol{\beta}$ of **Slice** from the Gibbs sampler are strongly autocorrelated. The convergence of $\boldsymbol{\beta}$ of **Slice** and **M–H Sampler** can be assessed by viewing the ACF plots in Fig. 3. The M–H

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

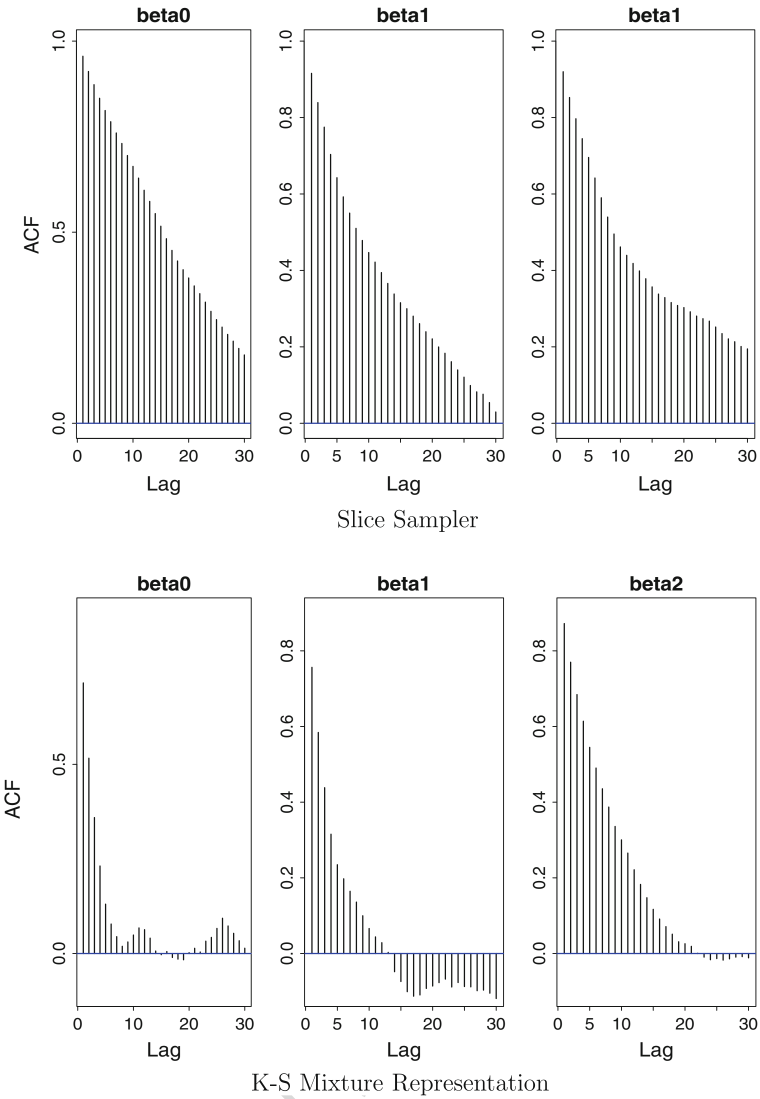

Fig. 2 ACF Plots of $\boldsymbol{\beta}$ for the GLMDM with logistic link. The *top panel* are the plots for $(\beta_0, \beta_1, \beta_2)$ from the slice sampler, and the *bottom panel* are the plots for $(\beta_0, \beta_1, \beta_2)$ from the K–S/normal mixture sampler

chain with candidate densities from log-linear models mixes better, giving additional confidence about convergence.

5.3 Probit models

For completeness, we also generated data, similar to that described in Sect. 3.2, for a probit link. In Fig. 4 we only show the ACF plot from a latent variable Gibbs sampler

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

Table 2 Estimation of the coefficients of the GLMDM with log link function and the estimate of K, with true values $K = 40$ and $\boldsymbol{\beta} = (3, 0.5, 1)$

Estimation method	β_0	β_1	β_2	K
Slice	2.7984 (0.0099)	0.0907 (0.0196)	0.8350 (0.0184)	43.5188 (4.1398)
M–H Sampler	2.3107 (0.1407)	0.8493 (1.1309)	0.9492 (1.0637)	43.5161 (4.1274)

Standard errors are in parentheses

as described in Sect. 3.1. where we see that the autocorrelations are not as good as the M–H algorithm, but better than those of the slice sampler.

6 Data analysis

In this section we provide two real data examples that highlight the workings of generalized linear Dirichlet process random effects models, using both logit and probit link functions. Both examples are drawn from important questions in social science research: voting behavior and terrorism studies. The voting behavior study, of social attitudes in Scotland, is fit using a logit link, while the terrorism data is fit with a probit link.

6.1 Social attitudes in Scotland

The data for this example come from the Scottish Social Attitudes Survey, 2006 (UK Data Archive Study Number 5840). This study is based on face-to-face interviews conducted using computer assisted personal interviewing and a paper-based self-completion questionnaire, providing 1,594 data points and 669 covariates. However, to highlight the challenge in identifying consistent attitudes with small data sizes we restrict the sample analyzed to females 18–25 years-old, giving 44 cases. This is a politically interesting group in terms of their interaction with the government, particularly with regard to healthcare and Scotland's voice in UK public affairs. The general focus was on attitudes towards government at the UK and national level, feelings about racial groups including discrimination, views on youth and youth crime, as well as exploring the Scottish sense of national identity.

Respondents were asked whether they favored full independence for Scotland with or without membership in the European Union versus remaining in the UK under varying circumstances. This was used as a dichotomous outcome variable to explore the factors that contribute to advocating secession for Scotland. The explanatory variables used are: `househld` measuring the number of people living in the respondent's household, `relgsums` indicating identification with the Church of Scotland versus another or no religion, `ptyallgs` measuring party allegiance with the ordering of parties given from more conservative to more liberal, `idlosem` a dichotomous variable equal to one if the respondent agreed with the statement that increased numbers of Muslims in Scotland would erode the national identity, `marrmus` another

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

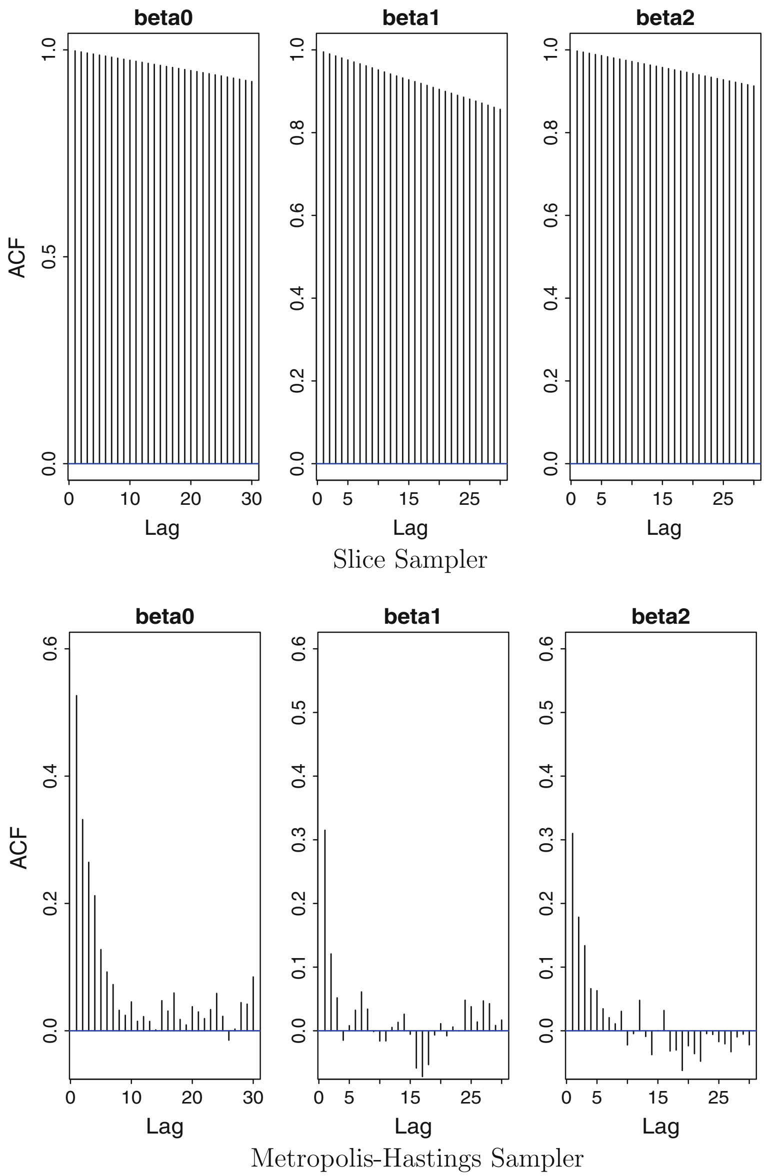

Fig. 3 ACF Plots of $\boldsymbol{\beta}$ for the GLMDM with log link. The *top panel* are the plots for $(\beta_0, \beta_1, \beta_2)$ from the slice sampler, and the *bottom panel* are the plots for $(\beta_0, \beta_1, \beta_2)$ from the M–H sampler

dichotomous variable equal to one if the respondent would be unhappy or very unhappy if a family member married a Muslim, `ukintnat` for agreement that the UK government works in Scotland's long-term interests, `natinnat` for agreement that the Scottish Executive works in Scotland's long-term interests, `voiceuk3` indicating that the respondent believes that the Scottish Parliament gives Scotland a greater voice in the UK, `nhssat` indicating satisfaction (1) or dissatisfaction (0) with the National

Journal: **10260** Article No.: **0168** TYPESET DISK LE CP Disp.:**2011/10/22** Pages: **32** Layout: **Small-X**

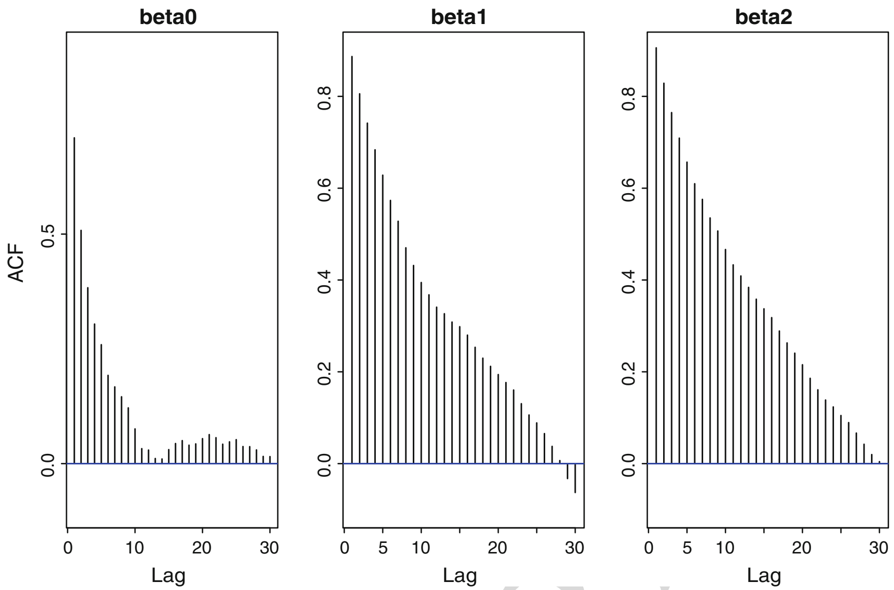

Fig. 4 ACF Plots for $(\beta_0, \beta_1, \beta_2)$ for the GLMDM with probit link, using the simulated data of Sect. 5.1

Health Service, `hincdif2`, a seven-point Likert scale showing the degree to which the respondent is living comfortably on current income or not (better in the positive direction), `unionsa` indicating union membership at work, `whrbrn` a dichotomous variable indicating birth in Scotland or not, and `hedqual2` the respondent's education level. We retain the variable names from the original study for ease of replication by others. All coding decisions (along with code for the models and simulations) are documented on the webpage http://www.jgill.wustl.edu/replication.html.

We ran the Markov chain for 10,000 iterations saving the last 5,000 for analysis. All indications point towards convergence using empirical diagnostics (Geweke, Heidelberger & Welsh, graphics, etc.). The results in Table 3 are interesting in a surprising way. Notice that there are very similar results for the standard Bayesian logit model with flat priors (estimated in `JAGS`, see http://www-fis.iarc.fr/~martyn/software/jags/) and the GLMDM logit model, save for one coefficient (discussed below). This indicates that the nonparametric component does not affect all of the marginal posterior distributions and the recovered information is confined to specific aspects of the data. Figure 5 graphically displays the credible intervals, and makes it easier to see the agreement of the analyses in this case.

Several of the coefficients point towards interesting findings from these results. There is reliable evidence from the Dirichlet process results that women under 25 believe that increased numbers of Muslims in Scotland would erode the Scottish national identity. This is surprising since anecdotally and journalistically one would expect this group to be among the most welcoming in the country. There is modest evidence (the two models differ slightly here) that this group is dissatisfied by the service provided by the National Health Service. In addition, these young Scottish

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

Table 3 Logit models for attitudes of females 18–25 years in Scotland

Coefficient	Standard logit				GLMDM logit			
	COEF	SE	95% CI		COEF	SE	95% CI	
`Intercept`	0.563	1.358	−2.133	3.274	0.351	1.396	−2.321	3.075
`househld`	0.281	0.303	−0.293	0.912	0.239	0.299	−0.342	0.830
`relgsums`	−2.006	1.604	−5.352	0.899	−1.840	1.614	−5.175	1.114
`ptyallgs`	−0.066	0.089	−0.239	0.114	−0.035	0.091	−0.207	0.150
`idlosem`	2.381	1.432	−0.101	5.498	2.663	1.343	0.219	5.487
`marrmus`	1.281	1.469	−1.552	4.164	1.089	1.528	−1.818	4.151
`ukintnat`	0.403	0.616	−0.799	1.638	0.347	0.582	−0.752	1.553
`natinnat`	−0.194	0.487	−1.179	0.739	−0.304	0.446	−1.174	0.575
`voiceuk3`	−0.708	0.433	−1.597	0.095	−0.637	0.443	−1.573	0.159
`nhssat`	−1.677	0.841	−3.347	−0.056	−1.405	0.812	−3.018	0.152
`hincdif2`	−1.219	0.446	−2.175	−0.415	−1.205	0.448	−2.114	−0.387
`unionsa`	0.521	0.723	−0.867	1.970	0.247	0.718	−1.117	1.692
`whrbrn`	1.494	0.944	−0.398	3.336	1.229	0.861	−0.461	2.924
`hedqual2`	−0.082	0.233	−0.532	0.374	−0.036	0.235	−0.493	0.434

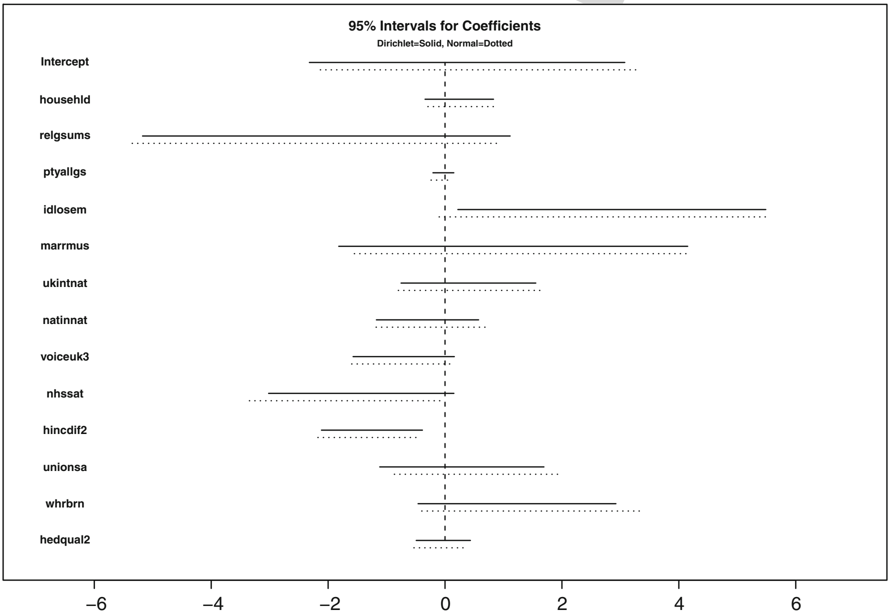

Fig. 5 Lengths and placement of credible intervals for the coefficients of the logit model fit for the Scottish Social Attitudes Survey on Females 18–25 years using Dirichlet process random effects (*black*) and normal random effects (*dotted lines*)

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

Table 4 Probit models for terrorism incidents

Coefficient	Standard probit				GLMDM probit			
	COEF	SE	95% CI		COEF	SE	95% CI	
`Intercept`	0.249	0.337	−0.412	0.911	0.123	0.187	−0.244	0.490
`DEM`	0.109	0.035	0.041	0.177	0.058	0.019	0.020	0.095
`FED`	0.649	0.469	−0.270	1.567	0.253	0.254	−0.245	0.750
`SYS`	−0.817	0.252	−1.312	−0.323	−0.418	0.136	−0.685	−0.151
`AUT`	1.619	0.871	−0.088	3.327	0.444	0.369	−0.279	1.167

women have a negative effect of increasing income on support for secession. It is also interesting here that the prior information provided by the GLMDM model is overwhelmed by the data as evidenced by the similarity between the two models. In line with Kyung (2010), most of the credible intervals of the GLMDM model are slightly shorter.

6.2 Terrorism targeting

In this example we look at terrorist activity in 22 Asian democracies over 8 years (1990–1997) with data subsetted from Koch and Cranmer (2007). Data problems (a persistent issue in the empirical study of terrorism) reduce the number of cases to 162 and make fitting any standard model difficult due to the generally poor level of measurement. The outcome of interest is dichotomous, indicating whether or not there was at least one violent terrorist act in a country/year pair. In order to control for the *level* of democracy (`DEM`) in these countries we use the Polity IV 21-point democracy scale ranging from -10 indicating a hereditary monarchy to $+10$ indicating a fully consolidated democracy (Gurr et al. 2003). The variable `FED` is assigned zero if sub-national governments do not have substantial taxing, spending, and regulatory authority, and one otherwise. We look at three rough classes of government structure with the variable `SYS` coded as: (0) for direct presidential elections, (1) for strong president elected by assembly, and (2) dominant parliamentary government. Finally, `AUT` is a dichotomous variable indicating whether or not there are autonomous regions not directly controlled by central government. The key substantive question evaluated here is whether specific structures of government and sub-governments lead to more or less terrorism.

We ran the Markov chain for 50,000 iterations disposing of the first half. There is no evidence of non-convergence in these runs using standard diagnostic tools. Table 4 again provides results from two approaches: a standard Bayesian probit model with flat priors, and a Dirichlet process random effects model. Notice first that while there are no changes in sign or statistical reliability for the estimated coefficients, the magnitudes of the effects are uniformly smaller with the enhanced model: four of the estimates are roughly twice as large and the last one is about three times as large in the standard model. This is clearly seen in Fig. 6, which is a graphical display of

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

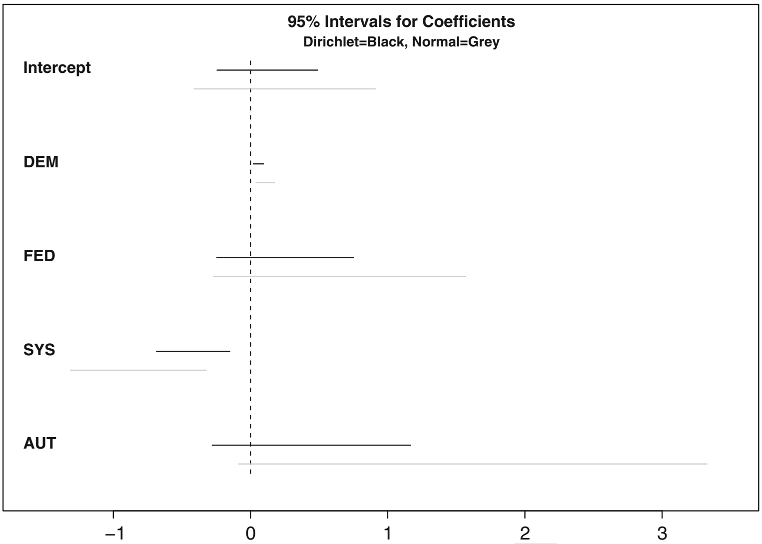

Fig. 6 Lengths and placement of credible intervals for the coefficients of the probit model fit for the terrorist activity data using Dirichlet process random effects (*black*) and normal random effects (*grey*)

Table 4. We feel that this indicates that there is extra variability in the data detected by the Dirichlet process random effect that tends to dampen the size of the effect of these explanatory variables on explaining incidences of terrorist attacks. Specifically, running the standard probit model would find an *exaggerated* relationship between these explanatory variables and the outcome.

The results are also interesting substantively. The more democratic a country is, the more terrorist attacks they can expect. This is consistent with the literature in that autocratic nations tend to have more security resources per capita and fewer civil rights to worry about. Secondly, the more the legislature holds central power, the fewer expected terrorist attacks. This also makes sense, given what is known; disparate groups in society tend to have a greater voice in government when the legislature dominates the executive. Two results are puzzling and are therefore worth further investigation. Strong sub-governments and the presence of autonomous regions both lead to more expected terrorism. This may result from strong separatist movements and typical governmental responses, an observed endogenous and cycling effect that often leads to prolonged struggles and intractable relations. We further investigate the use of Dirichlet process priors for understanding latent information in terrorism data in Kyung et al. (2011) with the goal of sorting out such effects.

7 Discussion

In this paper we demonstrate how to set up and run sampling schemes for the generalized linear mixed Dirichlet process model with a variety of link functions. We focus on the mixed effects model with a Dirichlet process prior for the random effects instead of the normal assumption, as in standard approaches. We

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

are able to estimate model parameters as well as the Dirichlet process parameters using convenient MCMC algorithms, and to draw latent information from the data. Simulation studies and empirical studies demonstrate the effectiveness of this approach.

The major methodological contributions here are the derivation and evaluation of strategies of estimation for model parameters in Sect. 3 and the inclusion of the precision parameter directly into the Gibbs sampler for estimation in Sect. 4.2. In the latter case, including the precision parameter in the Gibbs sampler means that we are marginalizing over the parameter rather than conditioning on it leading to a more robust set of estimates. Moreover, we have seen a large amount of variability in the performance of MCMC algorithms, with the slice sampler typically being less optimal than either a K–S mixture representation or a Metropolis–Hastings algorithm.

The relationship of credible intervals that is quite evident in Fig. 6, and less so in Fig. 5, that the Dirichlet intervals tend to be shorter than those based on normal random effects, persists in other data that we have analyzed. We have found that this in not a data anomaly, but has a explanation in that the Dirichlet process random effects model results in posterior variances that are smaller than that of the normal. Kyung et al. (2009) are able to prove this first in a special case of the linear model (when $\mathbf{X} = \mathbf{I}$), and then for almost all data vectors. The intuition follows the logic of multilevel (hierarchical) models whereby some variability at the individual-level is moved to the heterogeneous group-level thus producing a better model fit. Here, the group-level is represented by the nonparametric assignment to latent categories through the process of the Gibbs sampler.

Finally, we observed that the additional effort needed to include a Dirichlet process prior for the random effects in two empirical examples with social science data, which tends to be more messy and interrelated than that in other fields, added significant value to the data analysis. We found that the GLMDM model can detect additional variability in the data which affects parameter estimates. In particular, in the case of social attitudes in Scotland the GLMDM model improved estimates over the usual probit analysis. For the second example, we found that the GLMDM specification dampened-down over enthusiastic findings from a conventional model. In both cases either non-Bayesian or Bayesian models with flat priors would have reported results that had substantively misleading findings.

A Appendix: Generating the model parameters

A.1 A logistic model

A.1.1 Slice sampling

For fixed m and A, a Gibbs sampler of $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V})$ is

- for $d = 1, \ldots, p$,

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

$$\beta_d|\boldsymbol{\beta}_{-d}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}, A, \mathbf{y} \sim$$

$$\begin{cases} N\left(0, d^*\sigma^2\right) & \text{if } \beta_d \in \begin{matrix} \left[\left\{\max\left(\max_{X_{id}>0}\left(\frac{\alpha_{id}}{X_{id}}\right)\right), \left(\max_{X_{id}\le 0}\left(\frac{\gamma_{id}}{X_{id}}\right)\right)\right\},\right. \\ \left.\left\{\min\left(\min_{X_{id}\le 0}\left(\frac{\alpha_{id}}{X_{id}}\right)\right), \left(\min_{X_{id}>0}\left(\frac{\gamma_{id}}{X_{id}}\right)\right)\right\}\right] \end{matrix} \\ 0 & \text{otherwise} \end{cases}$$

where

$$\alpha_{id} = -\log\left(u_i^{-\frac{1}{y_i}} - 1\right) - \sum_{l\neq d} X_{il}\beta_l - (\mathbf{A}\boldsymbol{\eta})_i \quad \text{for } i \in S$$

$$\gamma_{id} = \log\left(v_i^{\frac{1}{y_i-1}} - 1\right) - \sum_{l\neq d} X_{il}\beta_l - (\mathbf{A}\boldsymbol{\eta})_i \quad \text{for } i \in F.$$

Here, $S = \{i : y_i = 1\}$ and $F = \{i : y_i = 0\}$.

- $\tau^2|\boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}, A, \mathbf{y} \sim$ Inverted Gamma $\left(\frac{k}{2} + a, \frac{1}{2}|\boldsymbol{\eta}|^2 + b\right)$
- for $j = 1, \ldots, k$,

$$\eta_j|\boldsymbol{\beta}, \tau^2, \mathbf{U}, \mathbf{V}, A, \mathbf{y} \sim \begin{cases} N\left(0, \tau^2\right) & \text{if } \eta_j \in \left(\max_{i\in S_j}\left\{\alpha_i^*\right\}, \min_{i\in S_j}\left\{\gamma_i^*\right\}\right) \\ 0 & \text{otherwise} \end{cases},$$

where

$$\alpha_i^* = -\log\left(u_i^{-1} - 1\right) - \mathbf{X}_i\boldsymbol{\beta} \quad \text{for } i \in S$$

$$\gamma_i^* = \log\left(v_i^{-1} - 1\right) - \mathbf{X}_i\boldsymbol{\beta} \quad \text{for } i \in F$$

- for $i = 1, \ldots, n$,

$$\pi_k(U_i|\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{V}, A, \mathbf{y}) \propto I\left[u_i < \left\{\frac{1}{1 + \exp(-\mathbf{X}_i\boldsymbol{\beta} - \eta_j)}\right\}^{y_i}\right] \quad \text{for } i \in S$$

$$\pi_k(V_i|\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, A, \mathbf{y}) \propto I\left[v_i < \left\{\frac{1}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta} + \eta_j)}\right\}^{1-y_i}\right] \quad \text{for } i \in F.$$

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

A.1.2 K–S mixture

Given ξ, for fixed m and A, a Gibbs sampler of $(\mu, \boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U})$ is

$$
\begin{aligned}
&\boldsymbol{\eta}|\mu, \boldsymbol{\beta}, \tau^2, \mathbf{U}, A, \mathbf{y}, \sigma^2 \sim N_k\left(\frac{1}{\sigma^2(2\xi)^2}\left(\frac{1}{\tau^2}I + \frac{1}{\sigma^2(2\xi)^2}A'A\right)^{-1} A'(\mathbf{U} - X\boldsymbol{\beta}),\right.\\
&\times \left.\left(\frac{1}{\tau^2}I + \frac{1}{\sigma^2(2\xi)^2}A'A\right)^{-1}\right)\\
&\mu|\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, A, \mathbf{y}, \sigma^2 \sim N\left(\frac{1}{p}\mathbf{1}'_p\boldsymbol{\beta}, \frac{d^*}{p}\sigma^2\right)\\
&\boldsymbol{\beta}|\mu, \tau^2, \boldsymbol{\eta}, \mathbf{U}, A, \mathbf{y}, \sigma^2 \sim N_p\left(\left(\frac{1}{d^*}I + \frac{1}{(2\xi)^2}X'X\right)^{-1}\right.\\
&\quad \times \left.\left(\frac{1}{d^*}\mu\mathbf{1}_p + \frac{1}{(2\xi)^2}X'(\mathbf{U} - A\boldsymbol{\eta})\right), \sigma^2\left(\frac{1}{d^*}I + \frac{1}{(2\xi)^2}X'X\right)^{-1}\right)\\
&\tau^2|\mu, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{U}, A, \mathbf{y}, \sigma^2 \sim \text{Inverted Gamma}\left(\frac{k}{2} + a, \frac{1}{2}|\boldsymbol{\eta}|^2 + b\right)\\
&U_i|\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, A, y_i, \sigma^2 \sim \begin{cases} N\left(\mathbf{X}_i\boldsymbol{\beta} + (A\boldsymbol{\eta})_i, \sigma^2(2\xi)^2\right) I(U_i > 0) & \text{if } y_i = 1\\ N\left(\mathbf{X}_i\boldsymbol{\beta} + (A\boldsymbol{\eta})_i, \sigma^2(2\xi)^2\right) I(U_i \le 0) & \text{if } y_i = 0\end{cases}
\end{aligned}
$$

Then we update ξ from

$$
\xi|\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, A, \mathbf{y} \sim \left(\frac{1}{(2\xi)^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2(2\xi)^2}|\mathbf{U} - X\boldsymbol{\beta} - A\boldsymbol{\eta}|^2} 8\sum_{\alpha=1}^{\infty}(-1)^{\alpha+1}\alpha^2\xi e^{-2\alpha^2\xi^2}.
$$

The conditional posterior density of ξ is the product of a inverted gamma with parameters $\frac{\alpha}{2} - 1$ and $-\frac{1}{8\sigma^2}|\mathbf{U} - X\boldsymbol{\beta} - A\boldsymbol{\eta}|^2$, and the infinite sum of the sequence $(-1)^{\alpha+1}\alpha^2\xi e^{-2\alpha^2\xi^2}$. To generate samples from this target density, we consider the alternating series method that is proposed by Devroye (1986). Based on his notation, we take

$$
\begin{aligned}
ch(\xi) &= 8\left(\frac{1}{\xi^2}\right)^{n/2} e^{-\frac{1}{8\sigma^2\xi^2}|\mathbf{U} - X\boldsymbol{\beta} - A\boldsymbol{\eta}|^2}\xi e^{-2\xi^2}\\
a_n(\xi) &= (\alpha + 1)^2 e^{-2\xi^2\{(\alpha+1)^2 - 1\}}.
\end{aligned}
$$

Here, we need to generate sample from $h(\xi)$, and we use accept-reject sampling with candidate $g(\xi^*) = 2e^{-2\xi^*}$, the exponential distribution with $\lambda = 2$, where $\xi^* = \xi^2$. Then we follow Devroye's method.

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

A.2 A log link model

A.2.1 Slice sampling

Starting from the likelihood $L(\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V})$, and the priors on $(\boldsymbol{\beta}, \tau^2)$, we have the following Gibbs sampler of the model parameters.

- The conditional posterior distribution of $\boldsymbol{\beta}$:

$$\pi_K(\boldsymbol{\beta}|\tau^2, \boldsymbol{\eta}, A, \mathbf{y}, \mathbf{U}, \mathbf{V}) \propto e^{-\frac{1}{2d^*\sigma^2}|\boldsymbol{\beta}|^2}$$
$$\times \prod_{i=1}^{n} I\left[u_i < \exp\left\{y_i(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right\}, v_i > \exp(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right].$$

For $d = 1, \ldots, p$,

$$\pi_K(\beta_d|\boldsymbol{\beta}_{-d}\tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}, A, \mathbf{y}) \propto e^{-\frac{1}{2d^*\sigma^2}\beta_d^2}$$
$$\times \prod_{i=1}^{n} I\left[u_i < \exp\left\{y_i(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right\}, v_i > \exp(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right],$$

which can be expressed as:

$$\pi_K(\beta_d|\boldsymbol{\beta}_{-d}\tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}, A, \mathbf{y}) \propto e^{-\frac{1}{2d^*\sigma^2}\beta_d^2}$$
$$\times \prod_{i=1}^{n} I\left[X_{id}\beta_d < \frac{1}{y_i}\log(u_i) - \sum_{l\neq j} X_{il}\beta_l - (\mathbf{A}\boldsymbol{\eta})_i, X_{id}\beta_d < \log(v_i)\right.$$
$$\left. - \sum_{l\neq j} X_{il}\beta_l - (\mathbf{A}\boldsymbol{\eta})_i\right],$$

where $\boldsymbol{\beta}_{-d} = (\beta_1, \ldots, \beta_{d-1}, \beta_{d+1}, \ldots, \beta_p)$. The full conditional posterior of β_d for $d = 1, \ldots, p$ is

$$\pi_k(\beta_d|\boldsymbol{\beta}_{-j}\tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}, A, \mathbf{y}) \propto e^{-\frac{1}{2d^*\sigma^2}\beta_d^2}\beta_d$$
$$\in \left[\left\{\max\left(\max_{X_{id}>0}\left(\frac{\alpha_{id}^*}{X_{id}}\right)\right), \left(\max_{X_{id}\leq 0}\left(\frac{\gamma_{id}^*}{X_{id}}\right)\right)\right\},\right.$$
$$\left.\left\{\min\left(\min_{X_{id}\leq 0}\left(\frac{\alpha_{id}^*}{X_{id}}\right)\right), \left(\min_{X_{id}>0}\left(\frac{\gamma_{id}^*}{X_{id}}\right)\right)\right\}\right],$$

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

where

$$\alpha_{id}^{*} = \frac{1}{y_i}\log(u_i) - \sum_{l\neq d} X_{il}\beta_l - (\mathbf{A}\boldsymbol{\eta})_i \quad \text{for } i \in S$$
$$\gamma_{id}^{*} = \log(v_i) - \sum_{l\neq d} X_{il}\beta_l - (\mathbf{A}\boldsymbol{\eta})_i \quad \text{for } i \in F.$$

Thus, for $d = 1, \ldots, p$,

$$\beta_d | \boldsymbol{\beta}_{-d}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}, A, \mathbf{y} \sim$$
$$\begin{cases} N\left(0, d^{*}\sigma^2\right) & \text{if } \beta_d \in \begin{array}{l}\Big[\Big\{\max\left(\max_{X_{id}>0}\left(\frac{\alpha_{id}^{*}}{X_{id}}\right)\right), \left(\max_{X_{id}\le 0}\left(\frac{\gamma_{id}^{*}}{X_{id}}\right)\right)\Big\}, \\ \Big\{\min\left(\min_{X_{id}\le 0}\left(\frac{\alpha_{id}^{*}}{X_{id}}\right)\right), \left(\min_{X_{id}>0}\left(\frac{\gamma_{id}^{*}}{X_{id}}\right)\right)\Big\}\Big]\end{array} \\ 0 & \text{otherwise.} \end{cases}$$

- The conditional posterior distribution of τ^2:

$$\pi_k(\tau^2 | \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}, A, \mathbf{y}) \propto \left(\frac{1}{\tau^2}\right)^{k/2+a+1} e^{-\frac{1}{\tau^2}\left(\frac{1}{2}|\boldsymbol{\eta}|^2+b\right)}.$$

Thus,

$$\tau^2 | \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{U}, \mathbf{V}, A, \mathbf{y} \sim \text{Inverted Gamma}\left(\frac{k}{2}+a, \frac{1}{2}|\boldsymbol{\eta}|^2+b\right).$$

- The conditional posterior distribution of $\boldsymbol{\eta}$:

$$\pi_k(\boldsymbol{\eta} | \boldsymbol{\beta}, \tau^2, \mathbf{U}, \mathbf{V}, A, \mathbf{y}) \propto \prod_{j=1}^{k} e^{-\frac{1}{2\tau^2}\eta_j^2} \prod_{i\in S_j} I\left[u_i < \exp\left\{y_i(\mathbf{X}_i\boldsymbol{\beta}+\eta_j)\right\},\right.$$
$$\left. v_i > \exp(\mathbf{X}_i\boldsymbol{\beta}+\eta_j)\right].$$

For $j = 1, \ldots, k$,

$$\pi_k(\eta_j | \boldsymbol{\beta}, \tau^2, \mathbf{U}, \mathbf{V}, A, \mathbf{y}) \propto e^{-\frac{1}{2\tau^2}\eta_j^2} \prod_{i\in S_k} I\left[u_i < \exp\left\{y_i(\mathbf{X}_i\boldsymbol{\beta}+\eta_j)\right\},\right.$$
$$\left. v_i > \exp(\mathbf{X}_i\boldsymbol{\beta}+\eta_j)\right]$$
$$\propto e^{-\frac{1}{2\tau^2}\eta_j^2} I\left[\eta_j \in \left(\max_{i\in S_k}\left\{\gamma_i^{*}\right\}, \min_{i\in S_k}\left\{\alpha_i^{*}\right\}\right)\right],$$

where

$$\alpha_i^* = \frac{1}{y_i}\log(u_i) - \mathbf{X}_i\boldsymbol{\beta}$$
$$\gamma_i^* = \log(v_i) - \mathbf{X}_i\boldsymbol{\beta}.$$

Thus, for $j = 1, \ldots, k$,

$$\eta_j|\boldsymbol{\beta}, \tau^2, \mathbf{U}, \mathbf{V}, A, \mathbf{y} \sim \begin{cases} N\left(0, \tau^2\right) & \text{if } \eta_j \in \left(\max_{i\in S_k}\left\{\gamma_i^*\right\}, \min_{i\in S_k}\left\{\alpha_i^*\right\}\right) \\ 0 & \text{otherwise.} \end{cases}$$

- The conditional posterior distribution of **U** and **V**:

$$\pi_k(U_i|\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{V}, A, \mathbf{y}) \;\propto\; I\left[u_i < \exp\left\{y_i(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right\}\right]$$
$$\pi_k(V_i|\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, \mathbf{U}, A, \mathbf{y}) \;\propto\; e^{-v_i} I\left[v_i > \exp(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right].$$

A.2.2 Metropolis–Hastings

Let $Z_i \equiv \log(Y_i)$, then Z_i is a linear mixed Dirichlet model (LMDM). For this model,

- the conditional posterior distribution of $\boldsymbol{\beta}$ in the LMDM:

$$\boldsymbol{\beta}|\mu, \tau^2, \boldsymbol{\eta}, A, \mathbf{Z}, \sigma^2 \sim N_p\Bigg(\left(\frac{1}{d^*}I + X'X\right)^{-1} \times \left(\frac{1}{d^*}\mu\mathbf{1}_p + X'(\mathbf{Z} - A\boldsymbol{\eta})\right), \sigma^2\left(\frac{1}{d^*}I + X'X\right)^{-1}\Bigg). \tag{24}$$

- the conditional posterior distribution of $\boldsymbol{\eta}$ in the LMDM:

$$\boldsymbol{\eta}|\boldsymbol{\beta}, \mu, \tau^2, \mathbf{Z}, A, \mathbf{y}, \sigma^2 \sim N_k\left(\frac{1}{\sigma^2}\left(\frac{1}{\tau^2}I + \frac{1}{\sigma^2}A'A\right)^{-1} A'(\mathbf{Z} - X\boldsymbol{\beta}), \left(\frac{1}{\tau^2}I + \frac{1}{\sigma^2}A'A\right)^{-1}\right). \tag{25}$$

Therefore, (24) is considered as a candidate density of $\boldsymbol{\beta}$ and (25) for $\boldsymbol{\eta}$.

The Metropolis–Hastings sampler of $(\boldsymbol{\beta}, \mu, \tau^2, \boldsymbol{\eta})$ follows.

- The conditional posterior distribution of $\boldsymbol{\beta}$ in the log linear model:

$$\pi_k(\boldsymbol{\beta}|\mu, \tau^2, \boldsymbol{\eta}, A, \mathbf{y}, \sigma^2) \propto e^{-\frac{1}{2d^*\sigma^2}|\boldsymbol{\beta}-\mu\mathbf{1}_p|^2} \times \prod_{i=1}^{n} e^{-\exp(\mathbf{X}_i\boldsymbol{\beta}+(\mathbf{A}\boldsymbol{\eta})_i)}\left[\exp(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right]^{y_i}.$$

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

Let

$$\pi_k^+(\boldsymbol{\beta}) \equiv e^{-\frac{1}{2d^*\sigma^2}|\boldsymbol{\beta}-\mu\mathbf{1}_p|^2} \prod_{i=1}^{n} e^{-\exp(\mathbf{X}_i\boldsymbol{\beta}+(\mathbf{A}\boldsymbol{\eta})_i)} \left[\exp(\mathbf{X}_i\boldsymbol{\beta}+(\mathbf{A}\boldsymbol{\eta})_i)\right]^{y_i}.$$

For given $\boldsymbol{\beta}^{(t)}$,

1. Generate $\boldsymbol{\beta}^* \sim N_p\left(\left(\frac{1}{d^*}I + X'X\right)^{-1}\left(\frac{1}{d^*}\mu\mathbf{1}_p + X'(\mathbf{Z}-A\boldsymbol{\eta})\right), \sigma^2\left(\frac{1}{d^*}I\right.\right.$ $\left.\left.+X'X\right)^{-1}\right)$.
2. Take

$$\boldsymbol{\beta}^{(t+1)} = \begin{cases} \boldsymbol{\beta}^* & \text{with probability } \min\left\{\left(\frac{\pi^+(\boldsymbol{\beta}^*)}{\pi_k^+(\boldsymbol{\beta}^{(t)})}\frac{q(\boldsymbol{\beta}^{(t)})}{q(\boldsymbol{\beta}^*)}\right), 1\right\} \\ \boldsymbol{\beta}^{(t)} & \text{otherwise} \end{cases},$$

where $q(\cdot)$ is a density of N_p distribution in (24), and recall that $\pi^+(\theta) = l(\theta)\pi(\theta)$.

- The conditional posterior distribution of μ in the log linear model:

$$\pi_k(\mu|\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, A, \mathbf{y}, \sigma^2) \propto \exp\left\{-\frac{p}{2d^*\sigma^2}\left(\mu - \frac{1}{p}\mathbf{1}_p'\boldsymbol{\beta}\right)^2\right\}.$$

Thus,

$$\mu|\boldsymbol{\beta}, \tau^2, \boldsymbol{\eta}, A, \mathbf{y}, \sigma^2 \sim N\left(\frac{1}{p}\mathbf{1}_p'\boldsymbol{\beta}, \frac{d^*}{p}\sigma^2\right).$$

- The conditional posterior distribution of τ^2 in the log linear model:

$$\pi_k(\tau^2|\boldsymbol{\beta}, \mu, \boldsymbol{\eta}, A, \mathbf{y}, \sigma^2) \propto \left(\frac{1}{\tau^2}\right)^{k/2+a+1} e^{-\frac{1}{\tau^2}\left(\frac{1}{2}|\boldsymbol{\eta}|^2+b\right)}.$$

Thus,

$$\tau^2|\boldsymbol{\beta}, \mu, \boldsymbol{\eta}, A, \mathbf{y}, \sigma^2 \sim \text{Inverted Gamma}\left(\frac{k}{2}+a, \frac{1}{2}|\boldsymbol{\eta}|^2+b\right).$$

- The conditional posterior distribution of $\boldsymbol{\eta}$ in the log linear model:

$$\pi_k(\boldsymbol{\eta}|\boldsymbol{\beta}, \mu, \tau^2, A, \mathbf{y}, \sigma^2)$$
$$\propto \prod_{k=1}^{K} e^{-\frac{1}{2\tau^2}\eta_k^2} \prod_{i\in S_k} e^{-\exp(\mathbf{X}_i\boldsymbol{\beta}+(\mathbf{A}\boldsymbol{\eta})_i)} \left[\exp(\mathbf{X}_i\boldsymbol{\beta}+(\mathbf{A}\boldsymbol{\eta})_i)\right]^{y_i}.$$

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

For $j = 1, \ldots, k$, let

$$\pi_k^{+}(\eta_j) \equiv e^{-\frac{1}{2\tau^2}\eta_j^2} \prod_{i \in S_j} e^{-\exp(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)} \left[\exp(\mathbf{X}_i\boldsymbol{\beta} + (\mathbf{A}\boldsymbol{\eta})_i)\right]^{y_i}$$

$$= e^{-\frac{1}{2\tau^2}\eta_j^2} \exp\left[\eta_j \sum_{i \in S_j} y_i - e^{\eta_j} \sum_{i \in S_j} e^{\mathbf{X}_i\boldsymbol{\beta}}\right].$$

For given $\boldsymbol{\eta}^{(t)}$,

1. Generate $\boldsymbol{\eta}^* \sim N_k\left(\frac{1}{\sigma^2}\left(\frac{1}{\tau^2}I + \frac{1}{\sigma^2}A'A\right)^{-1} A'(\mathbf{Z} - X\boldsymbol{\beta}), \left(\frac{1}{\tau^2}I + \frac{1}{\sigma^2}A'A\right)^{-1}\right)$.
2. Take

$$\boldsymbol{\eta}^{(t+1)} = \begin{cases} \boldsymbol{\eta}^* & \text{with probability } \min\left\{\left(\frac{\pi_k^{+}(\boldsymbol{\eta}^*)}{\pi_k^{+}(\boldsymbol{\eta}^{(t)})} \frac{q^*(\boldsymbol{\eta}^{(t)})}{q^*(\boldsymbol{\eta}^*)}\right), 1\right\} \\ \boldsymbol{\eta}^{(t)} & \text{otherwise} \end{cases},$$

where $q^*(\cdot)$ is a density of N_k distribution in (25).

References

Abramowitz M, Stegun IA (1972) Stirling numbers of the second kind. Section 24.1.4. In: Handbook of mathematical functions with formulas, graphs, and mathematical tables, 9th printing. Dover, New York, pp 824–825.

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88:669–679

Andrews DF, Mallows CL (1974) Scale mixtures of normal distributions. J R Stat Soc Ser B 36:99–102

Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat 2:1152–1174

Balakrishnan N (1992) Handbook of the logistic distribution. CRC Press, Boca Raton

Blackwell D, MacQueen JB (1973) Discreteness of Ferguson selections. Ann Stat 1:358–365

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:9–25

Buonaccorsi JP (1996) Measurement error in the response in the general linear model. J Am Stat Assoc 91:633–642

Chib S, Greenberg E, Chen Y (1998) MCMC methods for fitting and comparing multinomial response models. Technical Report, Economics Working Paper Archive, Washington University at St. Louis, http://129.3.20.41/econ-wp/em/papers/9802/9802001.pdf

Chib S, Winkelmann R (2001) Markov chain Monte Carlo analysis of correlated count data. J Bus Econ Stat 19:428–435

Damien P, Wakefield J, Walker S (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. J R Stat Soc Ser B 61:331–344

Devroye L (1986) Non-uniform random variate generation. Springer, New York

Dey DK, Ghosh SK, Mallick BK (2000) Generalized linear models: a Bayesian perspective. Marcel Dekker, New York

Dorazio RM, Mukherjee B, Zhang L, Ghosh M, Jelks HL, Jordan F (2007) Modelling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. Biometrics (online publication August 3, 2007)

Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90:577–588

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X

Fahrmeir L, Tutz G (2001) Multivariate statistical modelling based on generalized linear models. 2. Springer, New York

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230

Gill J, Casella G (2009) Nonparametric priors for ordinal Bayesian social science models: specification and estimation. J Am Stat Assoc 104:453–464

Gurr TR, Marshall MG, Jaggers K (2003) PolityIV, http://www.cidcm.umd.edu/inscr/polity/

Ishwaran H, James LF (2001) Gibbs sampling methods for stick-breaking priors. J Am Stat Assoc 96: 161–173

Jiang J (2007) Linear and generalized linear mixed models and their applications. Springer, New York

Koch MT, Cranmer S (2007) Terrorism than governments of the right? Testing the 'Dick Cheney' hypothesis: do governments of the left attract more than governments of the right?. Conflict Manage Peace Sci 24:311–326

Korwar RM, Hollander M (1973) Contributions to the theory of Dirichlet processes. Ann Probab 1:705–711

Kyung M, Gill J, Casella G (2009) Characterizing the variance improvement in linear Dirichlet random effects models. Stat Probab Lett 79:2343–2350

Kyung M, Gill J, Casella G (2010) Estimation in Dirichlet random effects models. Ann Stat 38:979–1009

Kyung M, Gill J, Casella G (2011) New findings from terrorism data: Dirichlet process random effects models for latent groups. J R Stat Soc Ser C (Forthcoming)

Liu JS (1996) Nonparametric hierarchical Bayes via sequential imputations. Ann Stat 24:911–930

Lo AY (1984) On a class of Bayesian nonparametric estimates: I. Density estimates. Ann Stat 12:351–357

MacEachern SN, Müller P (1998) Estimating mixture of Dirichlet process model. J Comput Graph Stat 7:223–238

McAuliffe JD, Blei DM, Jordan MI (2006) Nonparametric empirical Bayes for the Dirichlet process mixture model. Stat Comput 16:5–14

McCullagh P, Nelder JA (1989) Generalized linear models. 2. Chapman & Hall, New York

McCullagh P, Yang J (2006) Stochastic classification models. Int Congr Math III:669–686

McCulloch CE, Searle SR (2001) Generalized, linear, and mixed models. Wiley, New York

Mira A, Tierney L (2002) Efficiency and convergence properties of slice samplers. Scand J Stat 29:1–12

Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. J Comput Graph Stat 9:249–265

Neal RM (2003) Slice sampling. Ann Stat 31:705–741

Robert C, Casella G (2004) Monte Carlo statistical methods. Springer, New York

Sethuraman J (1994) A constructive definition of Dirichlet priors. Statistica Sinica 4:639–650

Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. J Am Stat Assoc 101: 1566–1581

Tierney L, Kadane JB (1986) Accurate approximations for posterior moments and marginal densities. J Am Stat Assoc 81:82–86

Wang N, Lin X, Gutierrez RG, Carroll RJ (1998) Bias analysis and SIMEX approach in generalized linear mixed measurement error models. J Am Stat Assoc 93:249–261

West M (1987) On scale mixtures of normal distributions. Biometrika 74:646–648

Wolfinger R, O'Connell M (1993) Generalized linear mixed models: a pseudolikelihood approach. J Stat Comput Simul 48:233–243

Journal: 10260 Article No.: 0168 TYPESET DISK LE CP Disp.:2011/10/22 Pages: 32 Layout: Small-X