Nonparametric Bayesian Modeling for Multivariate Ordinal Data

Athanasios Kottas, Peter Müller and Fernando Quintana*

August 18, 2004

Abstract

We propose a probability model for k-dimensional ordinal outcomes, i.e., we consider inference for data recorded in k-dimensional contingency tables with ordinal factors. The proposed approach is based on full posterior inference, assuming a flexible underlying prior probability model for the contingency table cell probabilities. We use a variation of the traditional multivariate probit model, with latent scores that determine the observed data. In our model, a mixture of normals prior replaces the usual single multivariate normal model for the latent variables. By augmenting the prior model to

*Athanasios Kottas is Assistant Professor, Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California at Santa Cruz, Santa Cruz, CA 95064 (E-mail: thanos@ams.ucsc.edu). Peter Müller is Professor, Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030 (E-mail: pm@odin.mdacc.tmc.edu). Fernando Quintana is Associate Professor, Departamento de Estadística, Pontificia Universidad Católica de Chile, Avenida Vicuña Mackenna 4860, Santiago, CHILE (E-mail: quintana@mat.puc.cl). Quintana's research was partially supported by grant FONDECYT 1020712. The order of author names is strictly alphabetical. The authors wish to thank an Associate Editor and three referees for comments that greatly improved the manuscript.

a mixture of normals we generalize inference in two important ways. First, we allow for varying local dependence structure across the contingency table. Second, inference in ordinal multivariate probit models is plagued by problems related to the choice and resampling of cutoffs defined for these latent variables. We show how the proposed mixture model approach entirely removes these problems. We illustrate the methodology with two examples, one simulated data set and one data set of interrater agreement.

Key Words: Contingency tables; Dirichlet process; Markov chain Monte Carlo; Polychoric correlations.

1 Introduction

We consider inference for k-dimensional ordinal outcomes. We define a mixture of multivariate probits model that can represent any set of k-dimensional contingency table cell probabilities. The proposed approach generalizes the traditional multivariate probit model, and at the same time allows for significant simplification of computational complexity. Computational simplicity is achieved by avoiding the need to impute cutoffs for the latent scores. Increased modeling flexibility is provided by allowing arbitrarily accurate approximation of any given set of probabilities on the outcomes.

Assume that for each of n experimental units the values of k ordinal categorical variables $V_1, \ldots, V_k$ are recorded. Let $C_j \geq 2$ represent the number of categories for the jth variable, $j = 1, \ldots, k$, and denote by $n_{\ell_1 \cdots \ell_k}$ the number of observations with $\boldsymbol{V} = (V_1, \ldots, V_k) = (\ell_1, \ldots, \ell_k)$. Denote by $p_{\ell_1 \cdots \ell_k} = P(V_1 = \ell_1, \ldots, V_k = \ell_k)$ the classification probability for the $(\ell_1, \ldots, \ell_k)$ cell. The data can be summarized in a multidimensional contingency table with $C = \prod_{j=1}^{k} C_j$ cells, with frequencies $\{n_{\ell_1 \cdots \ell_k}\}$ constrained by $\sum_{\ell_1 \cdots \ell_k} n_{\ell_1 \cdots \ell_k} = n$.

Inference for such data structures is of interest in many applications. The related statistical literature is correspondingly diverse and extensive. Many examples, applications and technical details can be found in Bishop et al. (1975), Goodman (1985), Read and Cressie (1988) and references therein. Log-linear models are a popular choice for the analysis of this data structure. However, the typically large number of parameters gives rise to a number of difficulties related to interpretation, prior elicitation and assessment of association between categorical variables.

An alternative modeling strategy involves the introduction of latent variables. Examples include Albert and Chib (1993), Cowles et al. (1996), Chib and Greenberg (1998), Bradlow and Zaslavsky (1999), Chen and Dey (2000) and Chib (2000) for ordinal regression models, Johnson and Albert (1999) for the analysis of data from multiple raters and Newton et al. (1995) for semiparametric binary regression. The common idea in these approaches is to introduce cutoffs $-\infty = \gamma_{j,0} < \gamma_{j,1} < \cdots < \gamma_{j,C_j-1} < \gamma_{j,C_j} = \infty$, for each $j = 1, \ldots, k$, and a k-dimensional latent variable vector $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$ such that for all $\ell_1, \ldots, \ell_k$

$$p_{\ell_1 \cdots \ell_k} = P\left(\bigcap_{j=1}^{k} \left\{ \gamma_{j,\ell_j - 1} < Z_j \leq \gamma_{j,\ell_j} \right\} \right). \tag{1}$$

A common distributional assumption is $\boldsymbol{Z} \sim N_k(\boldsymbol{0}, \boldsymbol{S})$, a k-dimensional normal distribution. One advantage of this model is the parsimony compared to the saturated log-linear model. In addition, $\rho_{st} = \text{corr}(Z_s, Z_t) = 0$, $s \neq t$, implies independence of the corresponding categorical variables. The coefficients ρ_{st}, $s \neq t$, are known as *polychoric correlation coefficients* and are traditionally used in the social sciences as a measure of the association between pairs of the (observed) categorical variables. See for instance, Olsson (1979) and more recently, Ronning and Kukuk (1996) and references therein.

However, the described model is not an appropriate choice for contingency tables that concentrate most of the data near the borders or corners and are rather sparse in the central cells. Also, the multivariate normal probit model implicitly assumes that the same polychoric correlations are a globally meaningful summary statistic. It does not allow for the dependence structure to vary across the contingency table. A typical example where such heterogeneity might arise is interrater agreement data. Different raters might agree about extremely high or low scores. But there might be considerable disagreement about scoring average observations. These limitations motivate the introduction of more flexible families of distributions for the latent variables $\boldsymbol{Z}$. Literature on fully Bayesian inference in this context is rather limited. We are only aware of Albert (1992), involving bivariate log-normal and t distributions, and Chen and Dey (2000), where certain scale mixtures of multivariate normals are considered.

An important practical issue is related to the choice of the cutoffs γ_{j,ℓ_j}. First, identifiability constraints complicate inference. Second, if the cutoffs are considered unknown parameters, inference is complicated by the fact that they are highly correlated with the latent variables $\boldsymbol{Z}$. In particular, when abundant data are available, the values of $\boldsymbol{Z}$ can become tightly clustered around a given γ_{j,ℓ_j} leaving little room for the cutoff to move when implementing a Markov chain Monte Carlo (MCMC) posterior simulation scheme. In this case, the corresponding full conditional posterior distribution becomes nearly degenerate. Johnson and Albert (1999) handle this problem via hybrid MCMC samplers.

In this article we propose a nonparametric probability model for the latent variables $\boldsymbol{Z}$ employing a Dirichlet process mixture of normals prior. We show that this provides the required flexibility to accommodate virtually any desired pattern in k-dimensional contin-

gency tables. At the same time we argue that, under the proposed model, we can without loss of generality fix the cutoffs. Therefore, from a modeling and inferential perspective, we provide a general framework for the analysis of contingency tables, while, from a practical perspective, we provide an approach that is easier to implement than existing Bayesian methods. Similar semiparametric models for univariate ordinal and binary data have been proposed by Erkanli et al. (1993) and Basu and Mukhopadhyay (2000), among others.

The article is organized as follows. Section 2 states our model, discussing its main features. Section 3 discusses simulation-based model fitting and posterior predictive inference. The methods are illustrated with two examples in section 4. We conclude with a summary in section 5.

2 A Bayesian Nonparametric Modeling Approach

2.1 The Model

We define a model for n vectors of ordinal categorical variables $\boldsymbol{V}_i = (V_{i1}, \ldots, V_{ik})$, $i = 1, \ldots, n$. First, as in (1), we introduce latent variables $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ik})$, $i = 1, \ldots, n$, such that

$$V_{ij} = \ell \qquad \text{if} \qquad \gamma_{j,\ell-1} < Z_{ij} \leq \gamma_{j,\ell}, \tag{2}$$

$j = 1, \ldots, k$ and $\ell = 1, \ldots, C_j$. When it simplifies notation we will alternatively write the link (2) between latent variables and ordinal outcome as a (degenerate) probability distribution $p(\boldsymbol{V}|\boldsymbol{Z})$. An important feature of the proposed model is that it allows us to use fixed cutoffs. See the discussion towards the end of this subsection.

Modeling proceeds now with the k-dimensional latent vectors $\boldsymbol{Z}_i$. We generalize tradi-

tional multivariate normal models by assuming a mixture of normals model. The mixture is with respect to both location $\boldsymbol{m}$ and covariance matrix $\boldsymbol{S}$ of the normal kernel. We define a probability model for the mixture distribution by assuming a prior probability model for the mixing distribution $G(\boldsymbol{m}, \boldsymbol{S})$. For reasons of technical convenience and interpretability we choose a Dirichlet process (DP) prior (Ferguson, 1973). The discrete nature of the DP simplifies interpretation. Each point mass $(\boldsymbol{m}, \boldsymbol{S})$ in the discrete mixing distribution G corresponds to a different set of polychoric correlations, implicit in $\boldsymbol{S}$. The location $\boldsymbol{m}$ specifies the factor levels of the contingency table where $\boldsymbol{S}$ defines the polychoric correlation. Details of the construction are discussed below.

We assume $\boldsymbol{Z}_i \overset{iid}{\sim} f$, with $f(\cdot|G) = \int p_{N_k}(\cdot|\boldsymbol{m}, \boldsymbol{S})\, dG(\boldsymbol{m}, \boldsymbol{S})$. Here, $p_{N_k}(\cdot|\boldsymbol{m}, \boldsymbol{S})$ denotes the density of a $N_k(\boldsymbol{m}, \boldsymbol{S})$ distribution. The mixture model f can be equivalently written as a hierarchical model by introducing latent variables $\boldsymbol{\theta}_i = (\boldsymbol{m}_i, \boldsymbol{S}_i)$ and breaking the mixture as

$$\boldsymbol{Z}_i|\boldsymbol{\theta}_i \overset{ind}{\sim} N_k(\boldsymbol{m}_i, \boldsymbol{S}_i), \quad i = 1, \ldots, n, \tag{3}$$

where, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ are an i.i.d. sample of latent variables from the mixing distribution G,

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|G \overset{iid}{\sim} G. \tag{4}$$

Let ρ_i denote the correlation matrix implied by $\boldsymbol{S}_i$. Conditional on $\boldsymbol{\theta}_i$, the correlation matrix ρ_i defines the local dependence structure in the neighborhood of $\boldsymbol{m}_i$. The elements of ρ_i can be interpreted as local polychoric correlation coefficients.

The model is completed with a prior probability model for the random distribution G. We assume

$$G|M, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D} \sim \mathcal{D}(MG_0), \tag{5}$$

a DP prior with total mass parameter M and baseline distribution G_0. For the baseline distribution G_0 we assume a joint distribution of a k-dimensional normal and an independent inverse Wishart. Specifically, we take $G_0(\boldsymbol{m}, \boldsymbol{S}) = N_k(\boldsymbol{m}|\boldsymbol{\lambda}, \boldsymbol{\Sigma})$ $\text{IWish}_k(\boldsymbol{S}|\nu, \boldsymbol{D})$, where $N_k(\boldsymbol{x}|\ldots)$ and $\text{IWish}_k(\boldsymbol{A}|\ldots)$ indicate, respectively, a k-dimensional normal distribution for the random vector $\boldsymbol{x}$ and an inverse Wishart distribution for the $k \times k$ (symmetric and positive definite) random matrix $\boldsymbol{A}$. One of the attractive features of the DP prior is that it allows straightforward posterior inference with MCMC simulation. The computational effort is, in principle, independent of the dimensionality of $\boldsymbol{\theta}_i$. Because of its computational simplicity, the DP is by far the most commonly used prior probability model for random probability measures. The DP generates almost surely discrete measures (e.g., Blackwell and MacQueen, 1973, Sethuraman, 1994). In some applications this discrete nature of the DP is awkward. However, in our setting, the discreteness is an asset as it simplifies interpretation. Let δ_x denote a point mass at x. The DP generates a discrete measure

$$G = \sum_{h=1}^{\infty} w_h \delta_{\boldsymbol{\theta}_h} \tag{6}$$

with stochastically ordered weights w_h. See Sethuraman (1994) for details. *A priori*, the first few weights cover most of the probability mass. *A posteriori*, weights are adjusted as required by the data. Although model (6) includes infinitely many point masses $\boldsymbol{\theta}_h$, only finitely many appear in the hierarchical model (3) and (4). Typically only very few distinct values for $\boldsymbol{\theta}_i$ are imputed. See the discussion in section 3 for details.

To complete the model specification, we assume independent hyperpriors

$$M \sim \text{Gamma}(a_0, b_0), \quad \boldsymbol{\lambda} \sim N_k(\boldsymbol{q}, \boldsymbol{Q}), \quad \boldsymbol{\Sigma} \sim \text{IWish}_k(b, \boldsymbol{B}), \quad \boldsymbol{D} \sim \text{Wish}_k(c, \boldsymbol{C}), \tag{7}$$

where $\text{Gamma}(\ldots)$ and $\text{Wish}_k(\ldots)$ denote a Gamma and a Wishart distribution, respec-

tively, with fixed scalar hyperparameters ν, a_0, b_0, b, c, a k-dimensional vector $\boldsymbol{q}$, and $k \times k$ positive definite matrices $\boldsymbol{Q}$, $\boldsymbol{B}$ and $\boldsymbol{C}$.

The model is stated in terms of covariance matrices instead of correlation matrices. A critical advantage of using covariance matrices is to avoid the difficulties associated with the modeling of correlation matrices. See, for example, Chib and Greenberg (1998), Daniels and Kass (1999) or McCulloch et al. (2000). Also, the variances on the diagonal of $\boldsymbol{S}_i$ play an important role in defining the mixture. Smaller variances imply that the corresponding local polychoric correlation matrices ρ_i are only locally valid, for ordinal scores close to $\boldsymbol{m}_i$.

Most importantly, with the nonparametric mixture $f(\cdot|G)$ for the latent variables we can model essentially any probability distribution for contingency tables. We give a constructive proof. Using (6) we can write the random mixture $f(\cdot|G)$ as a countable mixture of normal kernels, $\sum_{h=1}^{\infty} w_h \; p_{N_k}(\cdot|\boldsymbol{m}_h, \boldsymbol{S}_h)$. Consider any set of probabilities $\{p_{\ell_1 \cdots \ell_k}\}$ for the contingency table. Let $\{\tilde{p}_{\ell_1 \cdots \ell_k}\}$ be the corresponding set of probabilities under the mixture model $f(\cdot|G)$. Hence

$$\tilde{p}_{\ell_1 \cdots \ell_k} = P\left(\bigcap_{j=1}^{k} \left\{ \gamma_{j,\ell_j - 1} < Z_j \leq \gamma_{j,\ell_j} \right\} | G \right),$$

where $\boldsymbol{Z} = (Z_1, ..., Z_k) \sim f(\cdot|G)$ and the cutoffs, for each j, have fixed (say unit spaced) values. For each cell $(\ell_1, \ldots, \ell_k)$ with $p_{\ell_1 \cdots \ell_k} > 0$ center one term of the mixture $f(\cdot \mid G)$ within the rectangle defined by the corresponding cutoffs. For example, we could, for some h, set $m_{hj} = \gamma_{j,\ell_j} - 0.5$, j = 1,...,k, and choose $\boldsymbol{S}_h$ such that $\sqrt{1-\epsilon}$ of the mass of the $N_k(\boldsymbol{m}_h, \boldsymbol{S}_h)$ kernel is within the rectangle $(\gamma_{1,\ell_1 - 1}, \gamma_{1,\ell_1}) \times \ldots \times (\gamma_{k,\ell_k - 1}, \gamma_{k,\ell_k})$. Finally, set the corresponding weights w_h equal to $\sqrt{1-\epsilon}\, p_{\ell_1 \cdots \ell_k}$ to obtain $|p_{\ell_1 \cdots \ell_k} - \tilde{p}_{\ell_1 \cdots \ell_k}| < \epsilon$, for all cells $(\ell_1, \ldots, \ell_k)$, i.e., an arbitrarily accurate approximation of the probability distribution

for the contingency table by means of the mixture model for the latent variables.

This feature of the model has two important implications. First, the argument above shows that the mixture model can accommodate any given set of contingency table probabilities, including "irregular patterns" that can not be explained by a single multivariate normal probit model. The model provides a flexible modeling framework for contingency tables. In particular, it allows local dependence structure to vary across the contingency table, a feature that can be revealing of underlying patterns in applications.

Secondly, our model also provides a simple way to deal with the cutoffs. The random number of components and the distinct locations and covariance matrices in the mixture yield the result above without the need to consider random cutoffs. Hence, in the implementation of the model, there is no loss of generality in assuming fixed cutoffs. An important practical advantage of this approach is that the typically complex updating mechanisms for cutoffs (see, e.g., Cowles, 1996) are not required. In addition to the argument above, in section 4.1 we provide empirical evidence that the model is robust to the choice of cutoffs.

We conclude with a remark regarding two parametric models that result as limiting cases of the DP mixture model (2) – (7) when $M \to 0^+$ and $M \to \infty$. The former case yields the multivariate probit model, i.e., $\boldsymbol{\theta}_1 = ... \ \boldsymbol{\theta}_n = \boldsymbol{\theta}$ with $\boldsymbol{\theta} \mid \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D} \sim G_0$. The latter case results in a parametric exchangeable mixture model, i.e., the $\boldsymbol{\theta}_i$, conditionally on $\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D}$, become i.i.d. G_0. Given the discreteness of G, we expect the DP mixture model to outperform these parametric models, in terms of posterior predictive inference, when clusters are anticipated in the latent variables associated with the contingency table. In section 4 we offer a comparison of the nonparametric model with both of the parametric models above.

2.2 Prior Specification

The practical use of model (2) – (7) requires adopting specific values for ν, a_0, b_0, b, c, $\boldsymbol{q}$, $\boldsymbol{Q}$, $\boldsymbol{B}$ and $\boldsymbol{C}$.

The discrete nature of the DP realizations leads to a clustering structure defined by the grouping together of subjects with identical $\boldsymbol{\theta}_i$. Denote by n^* the number of resulting clusters. Then, $\mathrm{E}(n^*|M) \approx M \log((M+n)/M)$ and $\mathrm{Var}(n^*|M) \approx M\{\log((M+n)/M) - 1\}$ (see, e.g., Liu, 1996). Using the fact that *a priori* $\mathrm{E}(M) = a_0/b_0$ and $\mathrm{Var}(M) = a_0/b_0^2$ and an additional approximation based on Taylor series expansions we get

$$\begin{aligned} \mathrm{E}(n^*) &\approx \frac{a_0}{b_0}\log\left(1+\frac{nb_0}{a_0}\right) \\ \mathrm{Var}(n^*) &\approx \frac{a_0}{b_0}\log\left(1+\frac{nb_0}{a_0}\right) - \frac{a_0}{b_0} + \left\{\log\left(1+\frac{nb_0}{a_0}\right) - \frac{nb_0}{a_0+nb_0}\right\}^2 \frac{a_0}{b_0^2}. \end{aligned}$$

Equating these expressions with prior judgement for the mean and variance of n^* we obtain two equations that we can numerically solve for a_0 and b_0.

To provide a default specification for the remaining hyperparameters, we consider model (2) – (7) with $M \to 0^+$, i.e., the probit model, $\boldsymbol{Z}_i \overset{iid}{\sim} N_k(\boldsymbol{m}, \boldsymbol{S})$, with prior $(\boldsymbol{m}, \boldsymbol{S})|\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D} \sim G_0$ and the hyperpriors for $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{D}$ given in (7). For each dimension of the contingency table, $j = 1, \ldots, k$, we fix a rough approximation of the center and range of $\boldsymbol{Z}_j$ by computing approximate center e_j and range r_j of the cutoffs $\{\gamma_{j,1}, \ldots, \gamma_{j,C_j-1}\}$. (For instance, for the data set of section 4.2 we use cutoffs $-1, 0, 1, 2$ and take $e_j = 0$ and $r_j = 10$, a value about 3 times the actual range, $\gamma_{j,C_j-1} - \gamma_{j,1}$, of the cutoffs.) Let $\boldsymbol{H} = \mathrm{diag}((r_1/4)^2, \ldots, (r_k/4)^2)$. Matching $(e_1, \ldots, e_k)$ and $\boldsymbol{H}$ with the prior moments for $\boldsymbol{m}$ we find $\boldsymbol{q} = (e_1, \ldots, e_k)$ and $(b-k-1)^{-1}\boldsymbol{B} + \boldsymbol{Q} = 2\,\boldsymbol{H}$, where the left-hand sides of these two expressions arise as marginal prior moments for $\boldsymbol{m}$. Additionally, in the latter equality we used an extra variance inflation

factor 2. Splitting $2\boldsymbol{H}$ equally between the two summands we set $(b-k-1)^{-1}\boldsymbol{B} = \boldsymbol{Q} = \boldsymbol{H}$. We used $b = k+2$ to fix the largest possible dispersion in the prior for $\boldsymbol{\Sigma}$ (subject to finite $\mathrm{E}(\boldsymbol{\Sigma})$). To fix ν, c and $\boldsymbol{C}$, note that $\mathrm{E}(\boldsymbol{S}) = (\nu-k-1)^{-1}c\,\boldsymbol{C}$, and smaller values of $(\nu-k-1)^{-1}c$ yield more dispersed priors for $\boldsymbol{D}$. This observation can be used to specify ν and c (with $c \geq k$ that ensures a proper prior for $\boldsymbol{D}$). Finally, $\boldsymbol{C}$ is specified by setting $\mathrm{E}(\boldsymbol{S}) = \boldsymbol{H}$.

3 Computational Approach to Posterior Inference

Let $\boldsymbol{\theta}_i = (\boldsymbol{m}_i, \boldsymbol{S}_i)$, $\underline{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$, $\underline{\boldsymbol{Z}} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, and data $= \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n\}$. We use Gibbs sampling to explore the posterior distribution $p(\underline{\boldsymbol{Z}}, \underline{\boldsymbol{\theta}}, M, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D} \mid \text{data})$. The required full conditionals are obtained by considering the finite dimensional posterior that emerges after integrating out the random measure G (e.g., Blackwell and MacQueen, 1973),

$$p(\underline{\boldsymbol{Z}}, \underline{\boldsymbol{\theta}}, M, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D} \mid \text{data}) \propto \prod_{i=1}^{n} p(\boldsymbol{V}_i|\boldsymbol{Z}_i) \prod_{i=1}^{n} p_{N_k}(\boldsymbol{Z}_i|\boldsymbol{\theta}_i) p(\underline{\boldsymbol{\theta}}|M, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D}) p(M) p(\boldsymbol{\lambda}) p(\boldsymbol{\Sigma}) p(\boldsymbol{D}), \quad (8)$$

where $p(\underline{\boldsymbol{\theta}} \mid M, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D})$ arises by exploiting the Polya urn characterization of the DP (Blackwell and MacQueen, 1973) and the other factors are defined by (2), (3) and (7).

To sample latent variables $\underline{\boldsymbol{Z}}$, note that the full conditional posterior distribution of $\boldsymbol{Z}_i$ depends only on $\boldsymbol{V}_i$, $\boldsymbol{m}_i$ and $\boldsymbol{S}_i$ and is proportional to

$$\exp\left\{-(\boldsymbol{Z}_i - \boldsymbol{m}_i)^T \boldsymbol{S}_i^{-1} (\boldsymbol{Z}_i - \boldsymbol{m}_i)/2\right\} \prod_{j=1}^{k} I\{\gamma_{j,\ell_j-1} < Z_{ij} \leq \gamma_{j,\ell_j}\},$$

where $(\ell_1, \ldots, \ell_k)$ is defined by the value of $\boldsymbol{V}_i$. To update $\boldsymbol{Z}_i$ we draw from each of its coordinates, conditional on the rest. These conditional distributions are obtained by

considering first the univariate normal that results after the conditioning is done in the non-truncated multivariate version, and truncating this to the interval $(\gamma_{j,\ell_j-1}, \gamma_{j,\ell_j}]$.

Updating the latent mixture parameters $\underline{\boldsymbol{\theta}}$ and the hyperparameters $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$, $\boldsymbol{D}$ and M proceeds with standard posterior simulation methods for DP mixtures. See, for example, MacEachern and Müller (1998).

For the following discussion it is convenient to reparameterize the latent variables $\underline{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$. The discrete nature of the DP implies positive probabilities for ties among the $\boldsymbol{\theta}_i = (\boldsymbol{m}_i, \boldsymbol{S}_i)$. Let $n^* \leq n$ be the number of unique values among the $\boldsymbol{\theta}_i$. Denote the set of unique values (clusters) by $\underline{\boldsymbol{\theta}}^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{n^*}^*)$, where $\boldsymbol{\theta}_r^* = (\boldsymbol{m}_r^*, \boldsymbol{S}_r^*)$. Let $\boldsymbol{w} = (w_1, \ldots, w_n)$ be a vector of configuration indicators with $w_i = r$ if and only if $\boldsymbol{\theta}_i = \boldsymbol{\theta}_r^*$, and let n_r be the size of the rth cluster. Then $(\underline{\boldsymbol{\theta}}^*, \boldsymbol{w})$ is an equivalent representation of $\underline{\boldsymbol{\theta}}$, with $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{w_i}^*$.

Implementing L iterations of the earlier described MCMC algorithm we obtain posterior samples $\underline{\boldsymbol{\theta}}_l^*$, $\boldsymbol{w}_l$, n_l^*, $\underline{\boldsymbol{Z}}_l$, M_l, $\boldsymbol{\lambda}_l$, $\boldsymbol{\Sigma}_l$, $\boldsymbol{D}_l$, l = 1,...,L. Posterior draws n_l^* and $(\boldsymbol{m}_{rl}^*, \boldsymbol{S}_{rl}^*)$, r = 1,...,n_l^*, indicate the number of clusters, and their associated locations and covariance matrices, suggested by the data. The marginal posterior distribution of ρ_i corresponding to a data point $\boldsymbol{V}_i$ is implicit in the posterior of $\boldsymbol{S}_i$, i = 1,...,n.

We next turn to the posterior predictive distribution for a future observation $\boldsymbol{V}_0$. Denote by $\boldsymbol{Z}_0$ the associated latent vector. The assumptions of model (2) – (7) yield $p(\boldsymbol{V}_0, \boldsymbol{Z}_0 \mid \text{data}) = p(\boldsymbol{V}_0 \mid \boldsymbol{Z}_0)\, p(\boldsymbol{Z}_0 \mid \text{data})$, where $p(\boldsymbol{Z}_0 \mid \text{data})$ is the posterior predictive distribution of $\boldsymbol{Z}_0$ that can be developed using the structure induced by the DP prior. Denoting by $\boldsymbol{\phi} = (\underline{\boldsymbol{\theta}}^*, \boldsymbol{w}, M, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D})$ the entire parameter vector,

$$p(\boldsymbol{Z}_0 \mid \text{data}) = \int \int p_{N_k}(\boldsymbol{Z}_0 \mid \boldsymbol{m}_0, \boldsymbol{S}_0)\; \mathrm{d}p(\boldsymbol{m}_0, \boldsymbol{S}_0 \mid \boldsymbol{\phi})\; \mathrm{d}p(\boldsymbol{\phi} \mid \text{data}) \tag{9}$$

where

$$p(\boldsymbol{m}_0, \boldsymbol{S}_0 \mid \boldsymbol{\phi}) = \frac{M}{M+n} G_0(\boldsymbol{m}_0, \boldsymbol{S}_0) + \frac{1}{M+n} \sum_{r=1}^{n^*} n_r \delta_{(\boldsymbol{m}_r^*, \boldsymbol{S}_r^*)}(\boldsymbol{m}_0, \boldsymbol{S}_0). \tag{10}$$

Note that the association in the posterior predictive distribution for the ordinal variables, $p(\boldsymbol{V}_0 \mid \text{data})$, is driven by the dependence structure in the posterior predictive distribution for the latent variables, $p(\boldsymbol{Z}_0 \mid \text{data})$, and this, in turn, is parametrized by $(\boldsymbol{m}_0, \boldsymbol{S}_0)$.

Moreover, expressions (9) and (10) readily provide draws from $p(\boldsymbol{Z}_0 \mid \text{data})$ and Monte Carlo approximations to $p(\boldsymbol{z}_0 \mid \text{data})$ for any grid of values $\boldsymbol{z}_0$. They also clarify structure and the nature and amount of learning implied by the model. Note that $p(\boldsymbol{Z}_0 \mid \text{data})$ emerges by averaging, with respect to the posterior $p(\boldsymbol{\phi} \mid \text{data})$, the distribution

$$p(\boldsymbol{Z}_0 \mid \boldsymbol{\phi}) = \frac{M}{M+n} \int p_{N_k}(\boldsymbol{Z}_0 \mid \boldsymbol{m_0}, \boldsymbol{S_0}) dG_0(\boldsymbol{m_0}, \boldsymbol{S_0}) + \frac{1}{M+n} \sum_{r=1}^{n^*} n_r p_{N_k}(\boldsymbol{Z}_0 | \boldsymbol{m}_r^*, \boldsymbol{S}_r^*). \tag{11}$$

This is a mixture of multivariate normals, specified by the distinct locations, $\boldsymbol{m}_r^*$, and covariance matrices, $\boldsymbol{S}_r^*$, with an additional term that allows for a new cluster. The weight for this additional term decreases with increasing sample size, arguably an appealing feature of the model. As we observe more and more data, the chance of new patterns emerging in future observations decreases.

Finally, of interest is also inference for the table cell probabilities. For any cell $(\ell_1, \ldots, \ell_k)$ let $A_{\ell_1,\ldots,\ell_k} = \bigcap_{j=1}^{k} \left\{\gamma_{j,\ell_j-1} < Z_j \leq \gamma_{j,\ell_j}\right\}$ denote the corresponding range for the latent variables. We find

$$P(V_1 = \ell_1, \ldots, V_k = \ell_k \mid G) = P(A_{\ell_1,\ldots,\ell_k} \mid G) = \int \int_{A_{\ell_1,\ldots,\ell_k}} dp_{N_k}(\boldsymbol{Z} \mid \boldsymbol{m}, \boldsymbol{S})\, dG(\boldsymbol{m}, \boldsymbol{S})$$

i.e., $P(V_1 = \ell_1, \ldots, V_k = \ell_k | G)$ is a linear functional of $f(\cdot|G)$. Its posterior can be obtained using the approach of Gelfand and Kottas (2002). We omit details here simply noting

that the approach uses posterior draws from $p(\boldsymbol{\phi} \mid \text{data})$ and involves approximation of DP realizations, using the constructive definition in Sethuraman (1994), and evaluation of k-dimensional normal probabilities.

4 Data Illustrations

We present results from the analysis of two data sets in sections 4.1 and 4.2, after discussing below model comparison and a simple diagnostic that, under $k = 2$, can indicate when a mixture of normal probit model might be preferred over a single bivariate normal model.

To define such a diagnostic, we consider the table *(adjacent) log odds ratios*, $\psi_{ij} = \log p_{i,j} + \log p_{i+1,j+1} - \log p_{i,j+1} - \log p_{i+1,j}$, $i = 1, ..., C_1 - 1$, $j = 1, ..., C_2 - 1$. Denoting by $f(z_1, z_2)$ the density of the underlying latent variables (Z_1, Z_2), the log odds ratios ψ_{ij}, for each table cell (i, j), can be viewed as a discrete second difference approximation to $\partial^2 \log f(z_1, z_2)/\partial z_1 \partial z_2$, which is a constant (that depends on the correlation) for a normal density $f(z_1, z_2)$. Hence a large range, or, even more emphatically, different signs, in the observed log odds ratios point to the potential limitations of a probit model. Moreover, as we illustrate in sections 4.1 and 4.2, model-based posterior inference for the ψ_{ij} can be used to compare the probit model with the nonparametric model.

Regarding formal model comparison, Basu and Chib (2003) discuss the use of Bayes factors for DP mixture models. Alternatively, one could consider cross validation model comparison criteria. We illustrate the use of a criterion of this type in Section 4.1.

4.1 A Simulated Data Set

We test the performance of the proposed approach using simulated data from a non-standard distribution for the underlying latent variables. Specifically, setting $k = 2$, we generated $n = 100$ latent observations from a mixture (with equal weights) of two bivariate normals with means $(-1.5, -1.5)$ and $(0.5, 0.5)$, variances $(0.25, 0.25)$ for both components, and covariances -0.175 and 0.0875, respectively. The latent data are plotted in Figure 1. Using cutoffs -2.5, -1.5, -0.5, 0.5, 1.5 for both latent variables, a contingency table (see Table 1) is generated by grouping the latent data. For this table, empirical log odds ratios can be computed for only two cells, specifically, $\psi_{22} = -3.0$ and $\psi_{44} = 1.2$. These values suggest different dependence structure in two different parts of the table and hence indicate that a single bivariate normal model for the latent variables might not suffice.

We used the MCMC algorithm of section 3 to fit the model. Posterior inference is quite robust to different values of prior hyperparameters. For an illustration, Figure 1 plots the posterior predictive density $p(\boldsymbol{z}_0|\text{data})$ under four alternative priors for M, specifically, Gamma(a_0, b_0) distributions with $(a_0, b_0) = (2, 12)$, $(2, 5.5)$, $(2, 1.8)$, and $(2, 0.41)$, yielding $\text{E}(n^*) \approx 1$, 2, 5, and 15, respectively. The respective (0.05,0.25,0.5,0.75,0.95) posterior percentiles for n^* are (2,2,2,2,3), (2,2,2,3,4), (2,2,3,4,5), and (2,2,3,4,6). Posterior inference on n^* is highly informative and consistent across alternative priors. In all cases, the posterior for n^* indicates the need for at least two components in the mixture model. For the other hyperparameters, following section 2.2 and based on $e_j = 0$ and $r_j = 10$, $j = 1, 2$, we take $\boldsymbol{q} = (0, 0)^T$, $\boldsymbol{H} = \text{diag}(6.25, 6.25)$, $\boldsymbol{Q} = \boldsymbol{B} = \boldsymbol{H}$ and $b = 4$ for the priors for $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}$. Moreover, we set $\nu = 10$, $c = 5$ and $\boldsymbol{C} = \text{diag}(8.75, 8.75)$ yielding a rather dispersed prior

for $\boldsymbol{D}$. The posteriors of $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{D}$ (not shown), being very concentrated compared with their priors, indicate that the prior choices are vague compared to the likelihood, as well as that, at least with this sample size, the data enable learning for the model hyperparameters.

The DP mixture model successfully captures the clustering in the cells of the table driven by the bimodal underlying distribution for latent variables. Clustering in terms of locations is depicted in Figure 1. Clustering with regard to the dependence structure is illustrated in Table 1 where we provide posterior medians for the log odds ratios (for all cells (i, j) for which the ergodic averages of posterior draws for ψ_{ij} are numerically stable). In addition, 95% posterior interval estimates for ψ_{22} and ψ_{44} (with respective observed values -2.996 and 1.204) are given by $(-3.360, -1.034)$ and $(-0.305, 1.682)$, respectively. Association between the ordinal variables can also be assessed through inference for the ρ_i. The posterior means $\mathrm{E}(\rho_i \mid \text{data})$, i = 1,...,100, range from -0.712 to -0.686, for 52 pairs (Z_{i1}, Z_{i2}) corresponding to the upper-left part of Table 1, and from 0.082 to 0.177, for the remaining 48 pairs (Z_{i1}, Z_{i2}) corresponding to the lower-right part of Table 1. The posterior predictive distribution of $\rho_0 = \boldsymbol{S}_{0,12}/(\boldsymbol{S}_{0,11}\boldsymbol{S}_{0,22})^{1/2}$ is bimodal with modes at -0.725 and 0.255. See expression (10) for the posterior predictive distribution of $\boldsymbol{S}_0$. This posterior predictive distribution includes averaging over the future $\boldsymbol{V}_0$.

We next turn to the comparison of the nonparametric model (model M_1) with the two parametric competitors discussed in section 2.1, i.e., the probit model (model M_2) and the exchangeable mixture model (model M_3). We used the same priors on $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{D}$, given above, for all three models and a Gamma(2,1.8) prior on M. Figure 1 provides evidence in favor of the nonparametric mixture model. Moreover, as Table 1 indicates, the probit model fails to capture the clustering in the log odds ratios suggested

by the data, whereas the nonparametric model is more successful in this regard. Lastly, we consider a formal model comparison criterion based on cross validation. We evaluate $Q(M_r) = n^{-1} \sum_{i=1}^{n} \log p_{M_r}(\boldsymbol{V}_i \mid \boldsymbol{V}_{(-i)})$, where $p_{M_r}(\cdot \mid \boldsymbol{V}_{(-i)})$ is the posterior predictive distribution, under model M_r, based on the data vector $\boldsymbol{V}_{(-i)}$ that results after excluding the ith observation $\boldsymbol{V}_i$ (see, e.g., Bernardo and Smith, 2000, p. 403). We find $Q(M_1) = -2.327$, $Q(M_2) = -2.793$ and $Q(M_3) = -2.888$. Not surprisingly for the simulated data, the cross-validation criterion favors M_1.

Finally, to address sensitivity of posterior inference with respect to the choice of the cutoff points, we compute marginal posterior distributions for the cell probabilities $p_{\ell_1 \ell_2} = P(V_1 = \ell_1, V_2 = \ell_2 | G)$, ℓ_1, $\ell_2 = 1, \ldots, 6$, under different choices for the cutoffs and a wide range of priors for M. As anticipated, results were robust to both specifications. We plotted the marginal posterior distributions for the unknown table cell probabilities, using two sets of cutoffs and two alternative priors for M. Specifically we used the two sets of cutoffs $\{-2.5, -1.5, -0.5, 0.5, 1.5\}$ and $\{-25, -15, -5, 5, 15\}$. As alternative priors for M we used a Gamma(2,0.9) and a Gamma(2,12) distribution. Marginal posterior density plots (not shown) were practically indistinguishable under all four combinations of cutoffs and priors.

4.2 A Data Set of Interrater Agreement

We consider a data set from Melia and Diener-West (1994) reporting extent of scleral extension (extent to which a tumor has invaded the sclera or "white of the eye") as coded by two raters, A and B, for each of $n = 885$ eyes. The coding scheme uses five categories: 1 for "none or innermost layers", 2 for "within sclera, but does not extend to scleral surface", 3 for "extends to scleral surface", 4 for "extrascleral extension without transection" and 5 for

"extrascleral extension with presumed residual tumor in the orbit". The data set is available from the StatLib data sets archive at http://lib.stat.cmu.edu/datasets/csb/ch16a.dat. We provide the observed cell relative frequencies in Table 2. The 10 empirical log odds ratios that can be computed for this table range from -0.843 to 1.689, hence indicating that a single bivariate normal will likely not be a sufficiently flexible model for the latent variables.

To fit the model to these data, we use cutoffs -1, 0, 1, 2 for both variables. Again, following section 2.2, we use $e_j = 0$ and $r_j = 10$, $j = 1, 2$. Hence the priors for $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{D}$ are the same as in section 4.1. In addition, we take a Gamma(2,2.3) prior for M, which yields $\mathrm{E}(n^*) \approx 6$ and $\sqrt{\mathrm{Var}(n^*)} \approx 4.29$. As expected, based on the results of section 4.1 and the larger sample size available here, experimentation with other prior choices for $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$, $\boldsymbol{D}$ and M revealed robustness of posterior results.

An aspect of the inference that is interesting for many applications is the association between the ordinal variables. Such inference is provided in ρ_0, the correlation coefficient implied in $\boldsymbol{S}_0$. Panel (c) in Figure 2 shows the posterior for ρ_0 under the DP mixture model. For comparison, panel (d) shows the posterior under a bivariate probit model. Note how the probit model underestimates the association of the ordinal variables, as reported by ρ_0. This happens because the probit model is forced to fit a single covariance matrix for the latent data, thus failing to recognize clusters that might be present in the data. This can be seen by comparing panels (a) and (b) of Figure 2, which include draws from the posterior predictive distribution $p(\boldsymbol{Z}_0|\text{data})$ under the nonparametric and the probit model, respectively.

The $(0.25, 0.5, 0.75)$ posterior percentiles for n^* are given by $(6, 7, 8)$. In fact, we obtain $\mathrm{P}(n^* \geq 4 \mid \text{data}) = 1$. Note that, although $\mathrm{P}(n^* > 4 \mid \text{data}) = 0.943$, we find that the four

largest clusters always account for almost all of the probability mass. In general, we caution against overinterpreting inference on n^*. To mitigate concerns related to the identifiability of the mixture we recommend to use a prior on M, and thus indirectly on n^*, that strongly favors small numbers of clusters n^*.

Figure 3 plots the posterior means of Z_{i1} against the posterior means of Z_{i2}, arranged by the posterior means $\mathrm{E}(\rho_i \mid \text{data})$. Posterior summaries (means and 95% interval estimates) for the table cell probabilities $P(V_1 = \ell_1, V_2 = \ell_2|G)$, $(\ell_1, \ell_2) = 1, ..., 5$, are given in Table 2. Finally, Figure 4 provides the posteriors for four log odds ratios, specifically, ψ_{11}, ψ_{13}, ψ_{41} and ψ_{43}, under the DP mixture model and the probit model.

All the results indicate the utility of mixture modeling for this data set. Although one of the clusters clearly dominates the others, identifying the other three is important. One of them corresponds to agreement for large values (4 and 5) in the coding scheme, whereas the other two indicate regions of the table where the two raters tend to agree less strongly.

5 Summary

We have proposed a nonparametric Bayesian approach to model multivariate ordinal data. We have introduced a DP mixture model for latent variables defining classification probabilities in the corresponding contingency table. Two features of the model were extensively discussed. First, the flexibility provided by the probability model on latent variables allows us to handle virtually any data structure. Second, this flexibility can be achieved with fixed cutoffs, thus avoiding the most difficult computational challenge arising in posterior simulation for related models. The two examples illustrate these points.

References

Albert, J.H. (1992), "Bayesian Estimation of the Polychoric Correlation Coefficient," *Journal of Statistical Computation and Simulation*, 44, 47-61.

Albert, J.H., and Chib S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669-679.

Basu, S., and Chib, S. (2003), "Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models," *Journal of the American Statistical Association*, 98, 224-235.

Basu, S., and Mukhopadhyay, S. (2000), "Bayesian analysis of binary regression using symmetric and asymmetric links," *Sankhya, Series B, Indian Journal of Statistics*, 62, 372-387.

Bernardo, J.M., and Smith, A.F.M. (2000), *Bayesian Theory*, Chichester: Wiley.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: M.I.T. Press.

Blackwell, D., and MacQueen, J.B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353-355.

Bradlow, E.T., and Zaslavsky, A.M. (1999), "Hierarchical Latent Variable Model for Ordinal Data from a Customer Satisfaction Survey with "No Answer" Responses," *Journal of the American Statistical Association*, 94, 43-52.

Chen, M.-H., and Dey, D.K. (2000), "Bayesian Analysis for Correlated Ordinal Data Models," in Generalized Linear Models: A Bayesian Perspective, eds. D.K. Dey, S. Ghosh

and B.K. Mallick, pp. 135-162, New York: Marcel Dekker.

Chib, S. (2000), "Bayesian Methods for Correlated Binary Data," in Generalized Linear Models: A Bayesian Perspective, eds. D.K. Dey, S. Ghosh and B.K. Mallick, pp. 113-131, New York: Marcel Dekker.

Chib, S., and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, 347-361.

Cowles, M.K. (1996), "Accelerating Monte Carlo Markov Chain Convergence for Cumulative-link Generalized Linear Models," *Statistics and Computing*, 6, 101-111.

Cowles, M.K., Carlin, B.P., and Connett, J.E. (1996), "Bayesian Tobit Modeling of Longitudinal Ordinal Clinical Trial Compliance Data with Nonignorable Missingness," *Journal of the American Statistical Association*, 91, 86-98.

Daniels, M.J., and Kass, R.E. (1999), "Nonconjugate Bayesian Estimation of Covariance Matrices and its Use in Hierarchical Models," *Journal of the American Statistical Association*, 94, 1254-1263.

Erkanli, A., Stangl, D. and Müller, P. (1993), "A Bayesian analysis of ordinal data using mixtures," *ASA Proceedings of the Section on Bayesian Statistical Science*, 51-56, American Statistical Association (Alexandria, VA).

Ferguson, T.S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209-230.

Gelfand, A.E., and Kottas, A. (2002), "A Computational Approach for Full Nonparamet-

ric Bayesian Inference under Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 11, 289-305.

Goodman, L.A. (1985), "The Analysis of Cross-classified Data Having Ordered and/or Unordered Categories: Association Models, Correlation Models, and Asymmetry Models for Contingency Tables with or without Missing Entries," *The Annals of Statistics*, 3, 10-69.

Johnson, V.E., and Albert, J.H. (1999), *Ordinal Data Modeling*, New York: Springer.

Liu, J.S. (1996), "Nonparametric Hierarchical Bayes via Sequential Imputations," *The Annals of Statistics*, 24, 911-930.

MacEachern, S.N., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223-238.

McCulloch, R.E., Polson, N.G., and Rossi, P.E. (2000), "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," *Journal of Econometrics*, 99, 173-193.

Melia, B.M., and Diener-West, M. (1994), "Modeling Interrater Agreement on an Ordered Categorical Scale," in Case Studies in Biometry, eds. N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest and J. Greenhouse, pp. 323-338, New York: John Wiley and Sons.

Newton, M.A., Czado, C., and Chappell, R. (1995), "Bayesian Inference for Semiparametric Binary Regression," *Journal of the American Statistical Association*, 91, 132-141.

Olsson, U. (1979), "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient," *Psychometrika*, 44, 443-460.

Read, T.R.C., and Cressie, N.A.C. (1988), *Goodness-of-fit Statistics for Discrete Multivariate Data*, New York: Springer.

Ronning, G., and Kukuk, M. (1996), "Efficient Estimation of Ordered Probit Models," *Journal of the American Statistical Association*, 91, 1120-1129.

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639-650.

Table 1: For the simulated data, posterior medians for log odds ratios under the DP mixture model (in bold) and under the probit model (in italics). We only report inference for log odds ratios for cells where corresponding MCMC summaries could be computed with reasonable numerical accuracy. The observed cell relative frequencies are also included. Rows correspond to V_1 (latent Z_1) and columns to V_2 (latent Z_2).

	1	2	3	4	5	6
1	0 *1.216*	0 *0.985* **-2.835**	0 *1.003* **-2.554**	0.01 *1.014*	0	0
2	0 *0.996*	0.05 *0.928* **-2.061**	0.2 *0.942* **1.056**	0 *0.986*	0 *1.024*	0
3	0.01 *1.004* **-2.528**	0.2 *0.94* **0.628**	0.04 *0.924* **5.405**	0.01 *0.938* **0.649**	0 *1.003* **0.49**	0
4	0.01 *1.005*	0 *0.986*	0 *0.942* **0.659**	0.14 *0.928* **0.626**	0.09 *0.993* **0.594**	0
5	0	0 *0.999*	0 *1.007*	0.07 *0.988* **0.571**	0.15 *1.209* **0.757**	0.01
6	0	0	0	0	0.01	0

Table 2: For the interrater agreement data, observed cell relative frequencies (in bold) and posterior summaries for table cell probabilities (posterior mean and 95% central posterior intervals). Rows correspond to rater A and columns to rater B.

	1	2	3	4	5
1	**.3288** .3264	**.0836** .0872	**.0011** .0013	**.0011** .0020	**.0011** .0008
	(.2940, .3586)	(.0696, .1062)	(.0002, .0041)	(.0003, .0055)	(.0, .0033)
2	**.2102** .2136	**.2893** .2817	**.0079** 0.0080	**.0079** .0070	**.0034** .0030
	(.1867, .2404)	(.2524, .3112)	(.0033, .0146)	(.0022, .0143)	(.0006, .0074)
3	**.0023** .0021	**.0045** .0060	**.0** .0016	**.0023** .0023	**.0** .0009
	(.0004, .0059)	(.0021, .0118)	(.0004, .0037)	(.0004, .0059)	(.0, .0030)
4	**.0034** .0043	**.0113** .0101	**.0011** .0023	**.0158** .0142	**.0023** .0027
	(.0012, .0094)	(.0041, .0185)	(.0004, .0058)	(.0069, .0238)	(.0006, .0066)
5	**.0011** .0013	**.0079** .0071	**.0011** .0020	**.0090** .0084	**.0034** .0039
	(.0001, .0044)	(.0026, .0140)	(.0003, .0054)	(.0033, .0159)	(.0011, .0090)

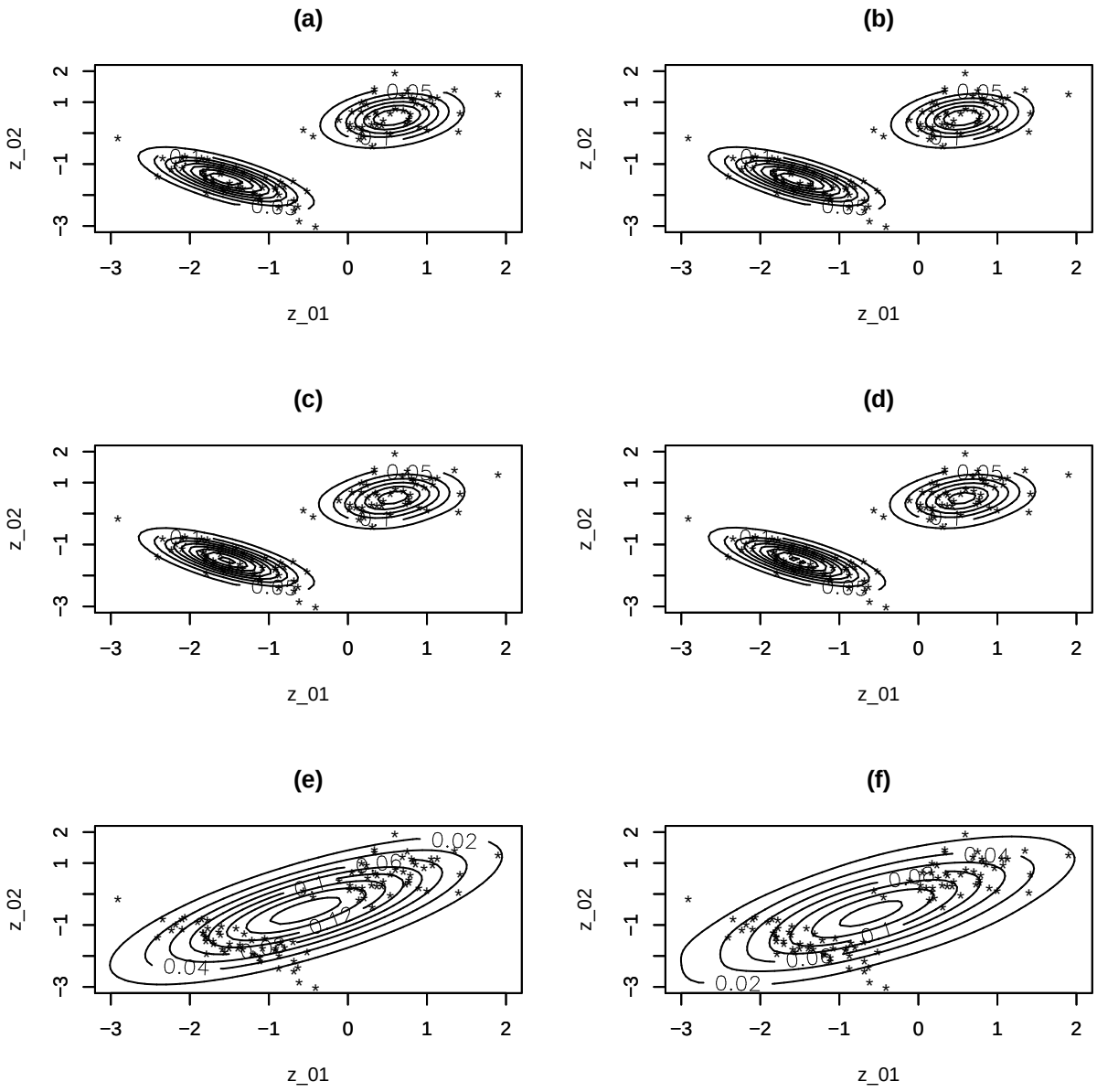

Figure 1: Simulated data. Posterior predictive density $p(\boldsymbol{z}_0|\text{data})$ under the DP mixture model with a Gamma(2,12), Gamma(2,5.5), Gamma(2,1.8) and Gamma(2,0.41) prior for M (panels (a) - (d), respectively), the probit model (panel (e)), and the parametric exchangeable model (panel (f)). In all cases, $p(\boldsymbol{z}_0|\text{data})$ is overlaid on a plot of latent data.

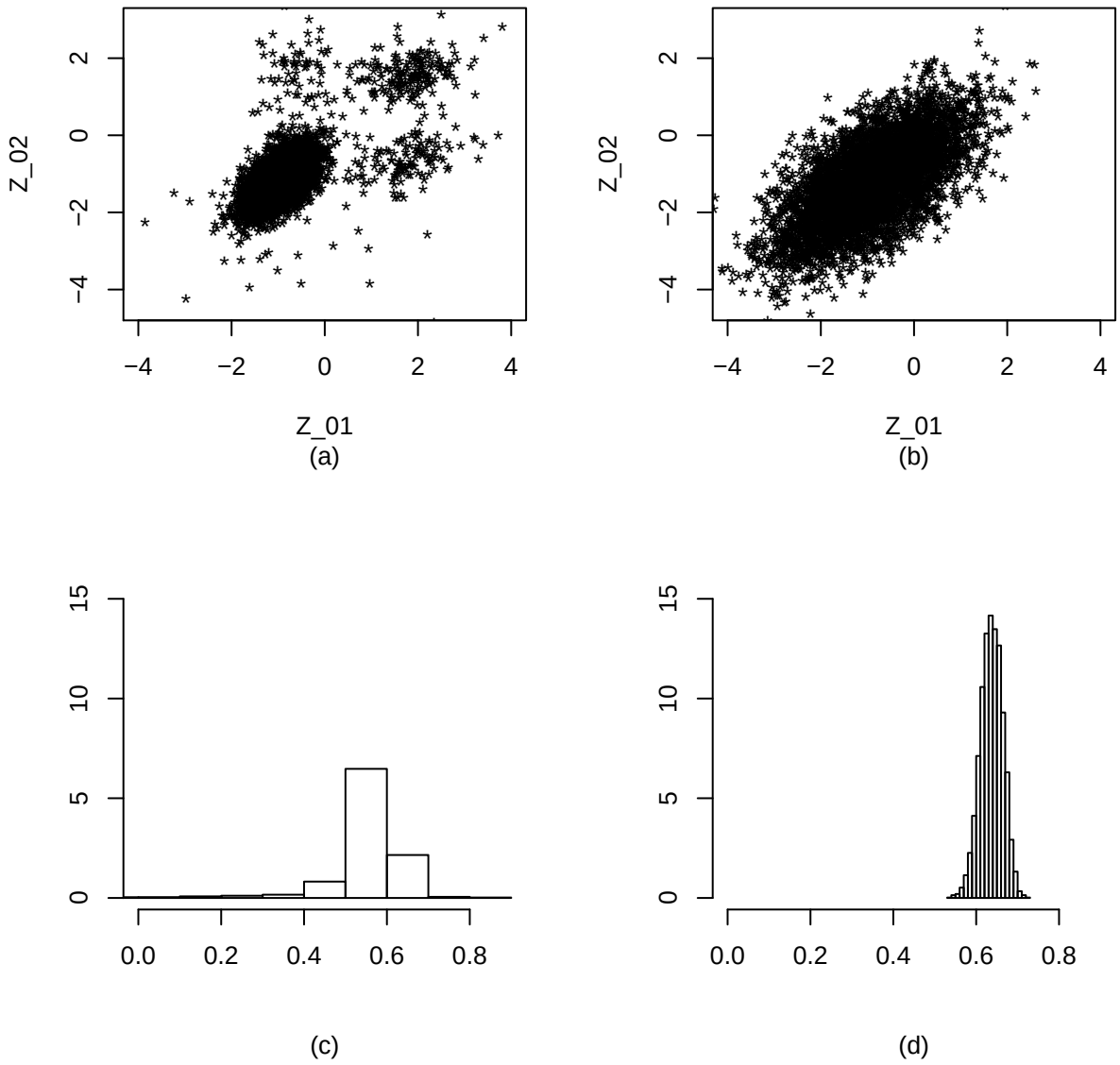

Figure 2: Interrater agreement data. Draws from $p(\boldsymbol{Z}_0|\text{data})$ and $p(\rho_0|\text{data})$ under the DP mixture model (panels (a) and (c), respectively) and the probit model (panels (b) and (d), respectively).

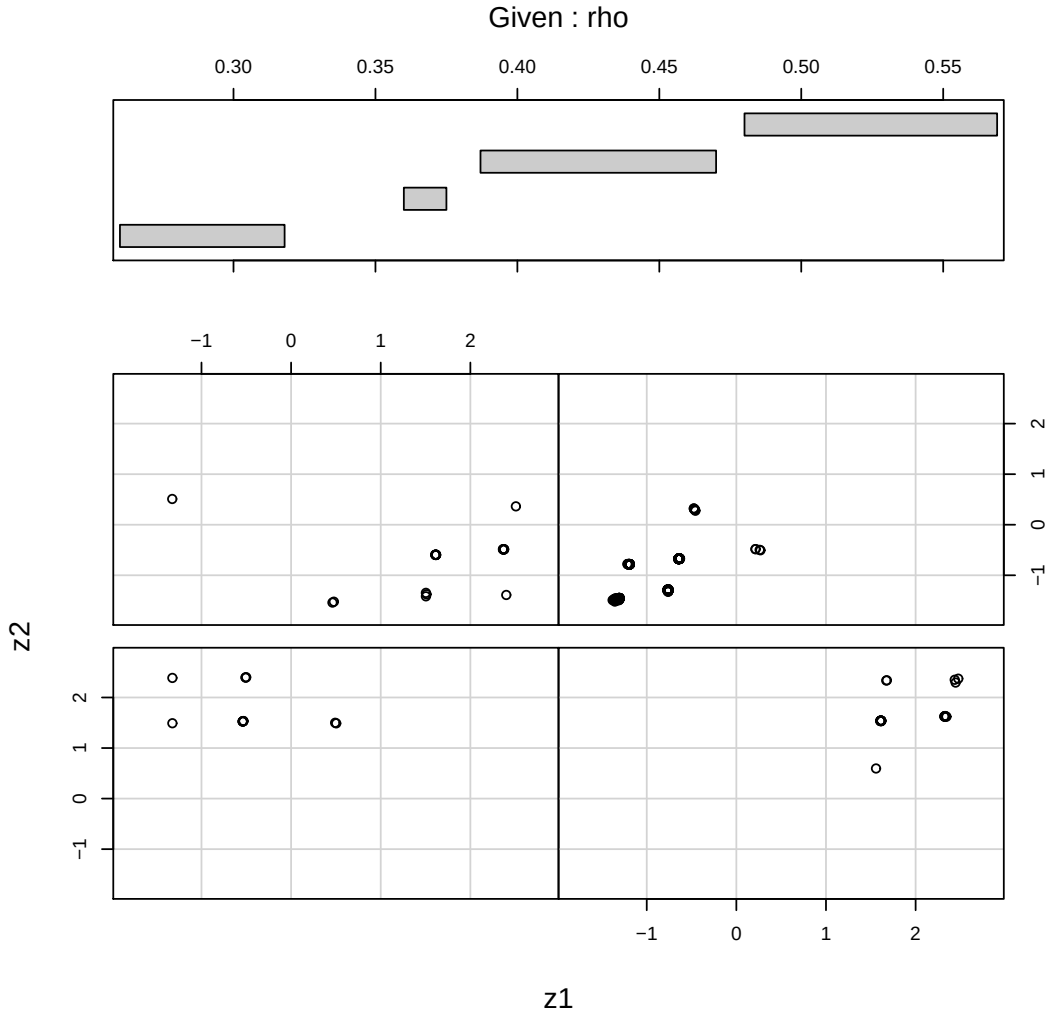

Figure 3: Interrater agreement data. Plots of pairs of posterior means $(\overline{Z}_{i1}, \overline{Z}_{i2}) = (\mathrm{E}(Z_{i1} \mid \text{data}), \mathrm{E}(Z_{i2} \mid \text{data}))$, arranged by $\overline{\rho}_i = \mathrm{E}(\rho_i \mid \text{data})$. The top panel indicates four subsets for $\overline{\rho}_i$, $(0.26, 0.32)$, $(0.36, 0.38)$, $(0.39, 0.47)$ and $(0.48, 0.57)$ with 14, 28, 25 and 818 associated pairs, respectively. The bottom four panels show $(\overline{Z}_{i1}, \overline{Z}_{i2})$ with the left bottom panel corresponding to the $(0.26, 0.32)$ subset for $\overline{\rho}_i$, the right bottom panel to $(0.36, 0.38)$, the left top panel to $(0.39, 0.47)$ and the right top panel to $(0.48, 0.57)$.

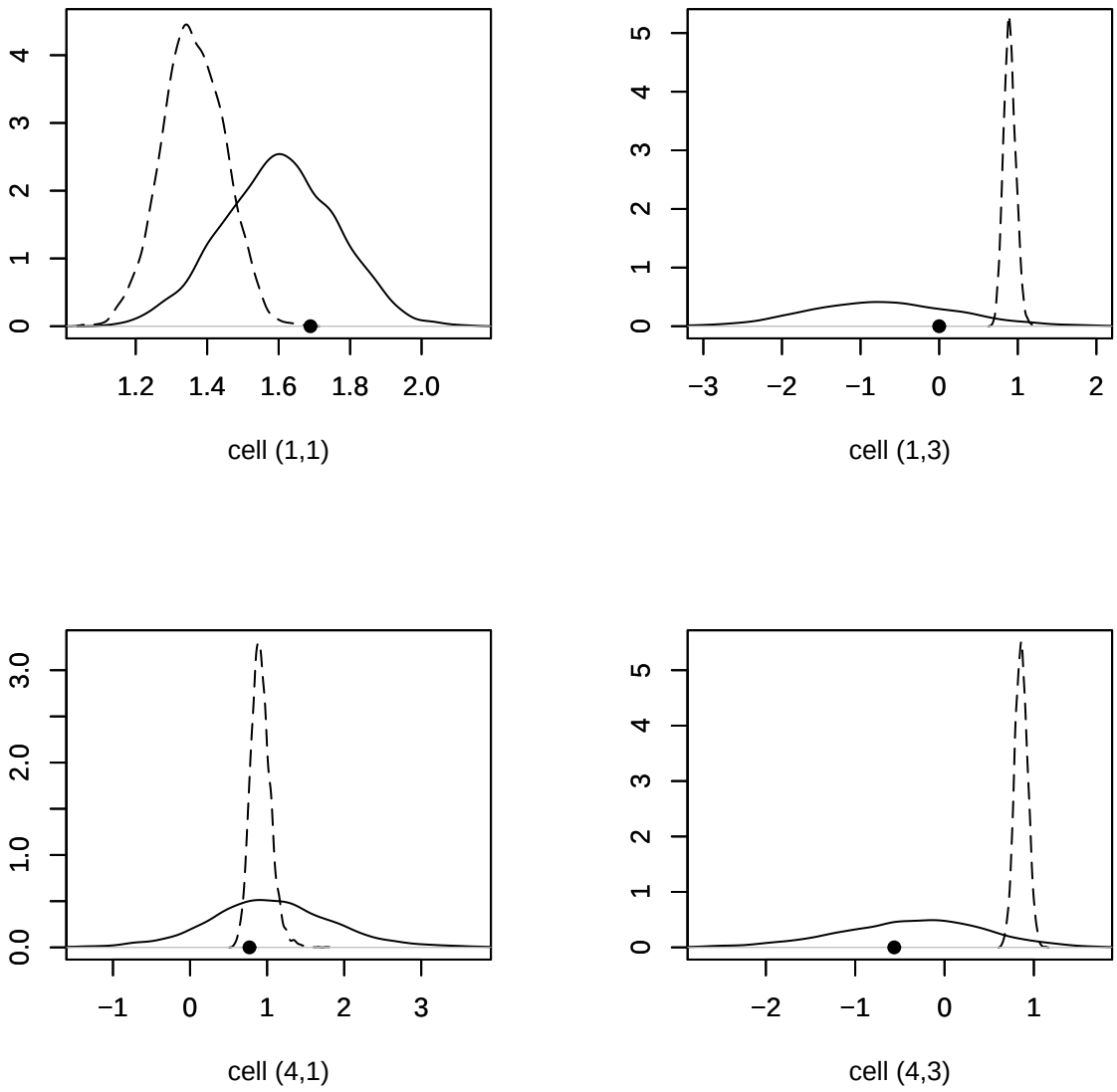

Figure 4: Interrater agreement data. Posteriors for four log odds ratios under the non-parametric model (solid lines) and the probit model (dashed lines). The circles denote the corresponding empirical log odds ratios.