TECHNICAL
REPORT

0628

A DIRICHLET PROCESS MIXTURE MODEL FOR THE ANALYSIS OF CORRELATED BINARY RESPONSES

A. JARA, M.J. GARCIA-ZATTERA, AND E. LESAFFRE

IAP STATISTICS
NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

A Dirichlet Process Mixture Model for the Analysis of Correlated Binary Responses

Alejandro Jara, María José García-Zattera and Emmanuel Lesaffre*

Biostatistical Centre, Catholic University of Leuven, Leuven, Belgium

Abstract

The multivariate probit model is a popular choice for modelling correlated binary responses. It assumes an underlying multivariate normal distribution dichotomized to yield a binary response vector. Other choices for the latent distribution have been suggested, but basically all models assume homogeneity in the correlation structure across the subjects. When interest lies in the association structure, relaxing this homogeneity assumption could be useful. Here we propose to replace the latent multivariate normal model by a location and association mixture model defined by a Dirichlet process. Attention is paid to the parametrization of the covariance matrix in order to make the Bayesian computations convenient. Our approach is illustrated on a simulated data set and applied to oral health data from the Signal Tandmobiel® study to examine the hypothesis that caries is mainly a spatially local disease.

Key Words: Multivariate binomial data, Latent variable representation, Probit models, Dirichlet process, Markov chain Monte Carlo.

1 Introduction

A popular tool for analyzing correlated binary data involves the introduction of latent variables. Some examples of this approach include the multivariate probit

*Author for correspondence. Biostatistical Centre, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B-3000 Leuven, Belgium. E-mail : Emmanuel.Lesaffre@med.kuleuven.be

model (Ashford and Sowden 1970; Lesaffre and Molenberghs 1991; Chib and Greenberg 1998), the bivariate lognormal and t-student models (Albert, 1992), the scale mixture of normals (Chen and Dey, 1998), the multivariate logit model (O'Brien and Dunson, 2004) and the multivariate skew-normal model (Chen, 2004). An advantage of this approach is that the dependency structure can be described parsimoniously in terms of correlation coefficients of the latent continuous variables. For the estimation of the parameters a Bayesian approach is popular given the computational complexity involved in these models. Especially the auxiliary variable MCMC algorithms are appealing here since the models are specified in terms of continuous underlying variables satisfying a linear regression model. This simple structure also facilitates generalizations to more complicated data structures, such as clustered data with possibly a mix of continuous and categorical response variables.

An extension of the above models could be to assume a mixture model for the distribution of the latent variables where the mixture is both in location as well as in covariance structure. Mixture models have been around in the literature for a long time (see, e.g., McLachlan and Peel 2000). Finite mixture models can be one option but, paradoxically, rather than handling the very large number of parameters resulting from these models with a large number of mixands, it may be easier to work with an infinite dimensional specification by assuming a random mixing distribution which is not restricted to a specific parametric family. The Dirichlet Process (DP) prior (Ferguson, 1973) is the most widely used in this context (see, e.g., Dey et al., 1998). However, the literature on fully Bayesian inference for the analysis of correlated categorical data is rather limited. We are only aware of Kottas et al. (2005), where a similar approach to the present work is adopted. However, they focussed on multivariate ordinal data and their implementation is not directly applicable to multivariate binary data.

The objective of this paper is to propose a modelling strategy for the analysis of correlated binary responses from a nonparametric Bayesian perspective by incorporating a Dirichlet process mixture of a normal prior as probability model for the latent variables. The mixture is with respect to both location and covariance of the normal kernel and is parameterized such that computations become more tractable. In particular, the model is stated in terms of covariance matrices constrained with respect to the conditional variances, avoiding the difficulties associated with modelling correlation matrices. We show that this provides the required flexibility to accommodate virtually any desired association pattern.

The motivation of this work lies in caries research. In fact, some argue that caries is a disease that is spatially local in the mouth while others believe that it affects the mouth globally or as a mixture of both processes, see *e.g.*, Hujoel et al. (1994) and references therein. In the Signal Tandmobiel® study we found a high association between the spatially remote deciduous molars for the presence/absence of caries (caries experience (CE)). This association was almost as high as the association between adjacent deciduous molars. This triggered us to verify whether this association could be explained by other (unmeasured) confounding factors and/or whether the association structure of CE could differ in certain sub populations of children.

The rest of the article is organized as follows. In Section 2 we state the model and discuss its main features. We also describe our simulation-based model fitting and posterior predictive inference. The methods are illustrated with two examples in Section 3. The first example is based on an artificial data set which sheds light on the performance of the model. The second example is based on the first year's caries experience data of the Signal Tandmobiel® study. We conclude with additional comments and discussion in Section 4.

2 The Semiparametric Bayesian Approach

2.1 Modelling via latent variables

Assume that for each of n experimental units the values of k binary variables $Y_{i1},...,Y_{ik}$ are recorded and let $\boldsymbol{Y}_i = (Y_{i1},...,Y_{ik})^T$, $i = 1,...,n$. A possible representation consists of viewing the binary variables as a discretized version of underlying continuous data, i.e. to introduce a k-dimensional latent variable vector $\boldsymbol{Z}_i = (Z_{i1},...,Z_{ik})^T$ such that,

$$Y_{il} = I_{\{Z_{il}>0\}}, 1 \leq l \leq k, \tag{1}$$

and

$$\boldsymbol{Z}_i \mid \boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\Sigma} \overset{ind}{\sim} f(\boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{2}$$

where $\boldsymbol{\beta}_0$ is a k-dimensional vector of intercepts, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_k^T)^T$, with $\boldsymbol{\beta}_l$, $l = 1, \ldots, k$, a p-dimensional column vector of regression coefficients associated to a p-dimensional vector of covariates of the l^{th} response, $\boldsymbol{x}_{il}$, $\boldsymbol{X}_i$ is a $k \times (k \times p)$-dimensional design matrix given by $\boldsymbol{X}_i = \boldsymbol{I}_{kk} \otimes \boldsymbol{x}_i^T$, $\boldsymbol{I}_{kk}$ is the $k \times k$ identity matrix,

$\otimes$ is the Kronecker product, and $\mathbf{\Sigma}$ is a scale matrix. The density f is determined by the location $\boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}$ and the scale matrix $\mathbf{\Sigma}$.

Modelling then proceeds with the k-dimensional latent vectors $\boldsymbol{Z}_i$. The distribution of $\boldsymbol{Z}_i$ determines the joint distribution of $\boldsymbol{Y}_i$ through (1) and (2), and their scale matrix, $\mathbf{\Sigma}$, captures the association among the observed variables. This modelling perspective is both flexible and general. In contrast, attempts to model the correlation of the binary responses directly may lead to difficulties (see, e.g., García-Zattera et al., 2005). A common distributional assumption is the normality of the k-dimensional latent variable vector, $\boldsymbol{Z}_i \overset{ind}{\sim} N_k\left(\boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}, \mathbf{\Sigma}\right)$, leading to the multivariate probit model. Although other choices for the latent distribution have been suggested in the literature, the fundamental assumption in all models is the homogeneity in the association structure across the experimental units.

2.2 The Proposal

We generalize the traditional multivariate probit model by assuming a mixture of normal distributions model for the latent variable representation. The normal distributions in the mixture vary with respect to both, location $\boldsymbol{\beta_0}$ and covariance matrix $\mathbf{\Sigma}$. Because it is not immediately clear what mixing distribution $G\left(\boldsymbol{\beta}_0, \mathbf{\Sigma}\right)$ is appropriate, we propose a probability model by assuming a prior probability model for G. Specifically, denote by $\mathcal{F}$ the set of all distribution functions on $\mathbb{R}^k \times \mathbb{R}^{k\times(k+1)/2}$, where k and $k \times (k+1)/2$ is the dimensionality of $\boldsymbol{\beta}_0$ and of the vector containing the lower triangular part of the covariance matrix $\mathbf{\Sigma}$, respectively. The Dirichlet Process (DP) (Ferguson, 1973) has become the probability measure most widely used as a prior on this set (see, e.g., Dey et al., 1998).

The DP generates a discrete probability measure,

$$G = \sum_{j=1}^{\infty} \omega_j \delta_{\boldsymbol{\theta}_j^C},$$

where ω_j are stochastic weights with decreasing expectations, δ_x is a point mass at x, and $\boldsymbol{\theta}_j^C = \left(\boldsymbol{\beta}_{0j}^C, \mathbf{\Sigma}_j^C\right)$ are *iid* from a baseline probability distribution, G_0. See, Rolin (1992) and Sethuraman (1994), for details. The discrete nature of the DP provides a useful tool for modelling. Indeed, given a particular set of the covariates, each point mass $\left(\boldsymbol{\beta_0}_j^C, \mathbf{\Sigma}_j^C\right)$ in the discrete mixing distribution G corresponds to a different set of correlations of the binary variables.

In summary, we assume $\boldsymbol{Z}_i \mid \boldsymbol{\beta} \overset{ind}{\sim} f_{\boldsymbol{X}_i}$, with

$$f_{\boldsymbol{X}_i}\left(\cdot|\boldsymbol{\beta}, G\right) = \int \phi_k\left(\cdot \mid \boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) dG\left(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}\right),$$

where $\phi_k\left(\cdot|\boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)$ denotes the k-variate normal density. The model can be thought of as a hierarchical model by introducing, subject-specific, latent variables $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_{0i}, \boldsymbol{\Sigma}_i)$, $i = 1, \ldots, n$, and breaking the mixture as

$$\boldsymbol{Z}_i \mid \boldsymbol{\theta}_i, \boldsymbol{\beta} \overset{ind}{\sim} N_k\left(\boldsymbol{\beta}_{0i} + \boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i\right), \tag{3}$$

where $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n$ is an *iid* sample of latent variables from the mixing distribution G, i.e.,

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

and

$$G \mid \alpha, G_0 \sim DP\left(\alpha G_0\right), \tag{5}$$

where $G \sim DP\left(\alpha G_0\right)$ refers to G being a random distribution generated by a Dirichlet process with baseline distribution G_0 and total mass parameter α. Observe that G_0 represents a baseline distribution and the parameter α controls the deviation of G from G_0 in a stochastic manner. In Section 2.4 we specify the choice of the baseline distribution, but first we need to determine the dimension of the latent parameters $\boldsymbol{\theta}_i$.

2.3 Identification Constrains for the Covariance Matrix

An important issue in the specification of the model relates to the choice of $\boldsymbol{\Sigma}_i$. To address this issue, we consider the standard multivariate probit model in which case the intercept vector and the covariance matrix is the same for all i, i.e. $\boldsymbol{\theta}_i \equiv \boldsymbol{\theta}$. The generalization of the arguments to our mixture model then follows readily. In this context, it should be noted that due to the threshold specification in (1) the scale of the latent variable is not likelihood identified (see, e.g., Chib and Greenberg, 1998). As a normalization, restrictions are usually placed on the covariance matrix. Most common is to fix the marginal variance of the latent variables to one implying that the matrix $\boldsymbol{\Sigma}$ becomes a correlation matrix. However, the positive definiteness constraint seriously complicates the choice of the prior probability distribution (see, e.g., Liechty et al., 2004) and the Bayesian computations. Moreover, typically the parameter vector corresponding to the correlation matrix is high-dimensional aggravating the problems.

An alternative normalization consists in constraining the conditional variances as follows. Let $\boldsymbol{\Sigma} = \boldsymbol{TDT}^t$ and $\boldsymbol{\Delta} = \boldsymbol{T}^{-1} = \{-\delta_{jl}\}_{k\times k}$ be a lower triangular matrix with 1's as diagonal entries and $\boldsymbol{D} = \{d_{jj}\}_{k\times k}$ is a diagonal matrix with positive entries. In this representation, the elements of $\boldsymbol{\Delta}$ and $\boldsymbol{D}$ have a statistical interpretation as was shown by Pourahmadi (1999). The reasoning is repeated here for completeness. Assume that $\boldsymbol{Z}_i = \boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{e}_i$. The below-diagonal entries of $\boldsymbol{\Delta}$ are the negative of the regression coefficients of e_{ij} on $e_{i1}, ..., e_{i(j-1)}$, and the jth diagonal entry of $\boldsymbol{D}$ corresponds to the conditional variance of e_{ij} given $e_{i1}, ..., e_{i(j-1)}$, i.e., $d_{jj} = Var\left(e_{ij} \mid e_{i1}, ..., e_{i(j-1)}\right)$. It follows that the vector of errors, $\boldsymbol{e}_i$, can be seen as a transformation of a random vector $\boldsymbol{\epsilon}_i$ following a standard k-variate normal distribution. That is $\boldsymbol{e}_i = \boldsymbol{TD}^{1/2}\boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_i \sim N_k\left(\boldsymbol{0}, \boldsymbol{I}_k\right)$. The distribution of the binary vector can be rewritten as,

$$\begin{aligned} P_{\boldsymbol{X}_i}\left(Y_{i1} = y_1, \ldots, Y_{ik} = y_k\right) = \; & P(\boldsymbol{\beta}_{01} + \boldsymbol{X}_{i1}\boldsymbol{\beta}_1 + \sqrt{d_{11}}\epsilon_{i1} \in A_{y_1}, \\ & \boldsymbol{\beta}_{02} + \boldsymbol{X}_{i2}\boldsymbol{\beta}_2 + t_{21}\sqrt{d_{11}}\epsilon_{i1} + \sqrt{d_{22}}\epsilon_{i2} \in A_{y_2}, \\ & \vdots \\ & \boldsymbol{\beta}_{0k} + \boldsymbol{X}_{ik}\boldsymbol{\beta}_k + \sum_{j=1}^{k-1} t_{kj}\sqrt{d_{jj}}\epsilon_{ij} + \sqrt{d_{kk}}\epsilon_{ik} \in A_{y_k}), \end{aligned} \tag{6}$$

where $A_{y_j} = (0, +\infty)$ for $y_j = 1$ and $A_{y_j} = (-\infty, 0]$ for $y_j = 0$. From expression (6), it is clear to see that fixing the conditional variances equal to one, $d_{jj} = 1$, identifies the model and do not impose unnecessary restrictions upon the parameters. Also, this normalization permits to sample $\boldsymbol{\Sigma}$ directly from its conditional distribution using a proper, for instance, normal prior on the parameters of the vector $\boldsymbol{\delta}$, which is the stacked vector of the negatives of the below-diagonal entries of $\boldsymbol{\Delta}$.

2.4 Prior Distributions

The random distribution G depends, besides on α, also on the hyperparameters associated to the baseline or centering distribution, i.e., G_0 becomes $G_{\boldsymbol{\vartheta}}$, where $\boldsymbol{\vartheta}$ is the set of hyperparameters. Based on the above reasoning, we assumed for the parametrized baseline distribution $G_{\boldsymbol{\vartheta}}$ a joint distribution of a k-dimensional ($\boldsymbol{\beta}_0$ parameters) and a $k \times (k-1)/2$-dimensional ($\boldsymbol{\delta}$ parameters) normal distribution. Specifically, we take

$$G_{\boldsymbol{\vartheta}}\left(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}\right) = N_k\left(\boldsymbol{\beta}_0 \mid \boldsymbol{\mu}_{\boldsymbol{\beta}_0}, \boldsymbol{S}_{\boldsymbol{\beta}_0}\right) N_{k\times(k-1)/2}\left(\boldsymbol{\delta} \mid \boldsymbol{\mu}_{\boldsymbol{\delta}}, \boldsymbol{S}_{\boldsymbol{\delta}}\right), \tag{7}$$

where, $\boldsymbol{\vartheta} = \{\boldsymbol{\mu}_{\boldsymbol{\beta}_0}, \boldsymbol{S}_{\boldsymbol{\beta}_0}, \boldsymbol{\mu}_{\boldsymbol{\delta}}, \boldsymbol{S}_{\boldsymbol{\delta}}\}$, $N_p\left(\boldsymbol{x} \mid ...\right)$ indicates a p-dimensional normal distribution for the vector $\boldsymbol{x}$, $\boldsymbol{\mu}_{\boldsymbol{\beta}_0}$ and $\boldsymbol{S}_{\boldsymbol{\beta}_0}$ is the mean and covariance matrix of the latent vector $\boldsymbol{\beta}_0$, respectively, and, $\boldsymbol{\mu}_{\boldsymbol{\delta}}$ and $\boldsymbol{S}_{\boldsymbol{\delta}}$ is the mean and covariance matrix of the latent vector $\boldsymbol{\delta}$. More formally, the DP prior is

$$G \mid \alpha, \boldsymbol{\vartheta} \sim DP\left(\alpha G_{\boldsymbol{\vartheta}}\right). \tag{8}$$

To complete the model specification, the model could be extended by assuming independent hyperpriors

$$\alpha \sim \Gamma\left(a_0, b_0\right), \ \boldsymbol{\beta} \sim N_p\left(\boldsymbol{b}, \boldsymbol{B}\right), \tag{9}$$

$$\boldsymbol{\mu}_{\boldsymbol{\beta}_0} \sim N_k\left(\boldsymbol{m}, \boldsymbol{\Upsilon}\right), \ \boldsymbol{S}_{\boldsymbol{\beta}_0} \sim IW_k\left(\gamma, \boldsymbol{\Gamma}\right), \tag{10}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{\delta}} \sim N_{k\times(k-1)/2}\left(\boldsymbol{\eta}, \boldsymbol{\Phi}\right), \ \boldsymbol{S}_{\boldsymbol{\delta}} \sim IW_{k\times(k-1)/2}\left(\lambda, \boldsymbol{\Omega}\right), \tag{11}$$

where Γ and IW refers to the Gamma and inverted Wishart distributions, respectively.

2.5 Prior Specification

The practical implementation of DP mixture model (1), (3), (4), and (8)- (11) requires adopting values for the hyperparameters a_0, b_0, $\boldsymbol{b}$, $\boldsymbol{B}$, $\boldsymbol{m}$, $\boldsymbol{\Upsilon}$, γ, $\boldsymbol{\Gamma}$, $\boldsymbol{\eta}$, $\boldsymbol{\Phi}$, λ and $\boldsymbol{\Omega}$. The discrete nature of the DP realizations leads to their well-known clustering properties. The choice of a_0 and b_0 needs some careful thought, as the parameter α directly controls the number of distinct components. When $\alpha \to 0^+$ and $\alpha \to \infty$ parametric models arise as limiting cases of the DP mixture model (1), (3), (4), and (8)- (11). The former case yields the multivariate probit model, i.e., $\boldsymbol{Z}_i \overset{ind}{\sim} N_k(\boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\boldsymbol{\theta}_1 = ... = \boldsymbol{\theta}_n = \boldsymbol{\theta} \equiv (\boldsymbol{\beta}_0, \boldsymbol{\Sigma})$ with prior distributions $(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}) \mid \boldsymbol{\vartheta} \sim G_{\boldsymbol{\vartheta}}$ and $\boldsymbol{\beta} \mid \boldsymbol{b}, \boldsymbol{B} \sim N_p(\boldsymbol{b}, \boldsymbol{B})$. The latter case results in a parametric exchangeable mixture model, i.e., $\boldsymbol{Z}_i \overset{ind}{\sim} N_k(\boldsymbol{\beta}_{0i} + \boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$ and $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n \mid \boldsymbol{\vartheta} \overset{iid}{\sim} G_{\boldsymbol{\vartheta}}$, with $\boldsymbol{\theta}_i \equiv (\boldsymbol{\beta}_{0i}, \boldsymbol{\Sigma}_i)$.

For any other choice of α the result is a DP process that produces a discrete G. In other words, $\boldsymbol{Z}_i \overset{ind}{\sim} N_k(\boldsymbol{\beta}_{0i} + \boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$ and the $\boldsymbol{\theta}_i$'s will be allowed to cluster and the clustering depends on α. Therefore, the values of a_0 and b_0 in the prior for α affects the number of expected mixtures. Strategies for the specification of these hyperparameters are often based on approximations of the conditional mean and conditional variance of the number of clusters, given the precision parameter α (see, e.g., Kottas et al., 2005). Specifically, denoting by n^* the number of resulting clusters, this approach relies on

$$E(n^*|\alpha) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} \approx \alpha \log\left(\frac{\alpha + n}{\alpha}\right) \tag{12}$$

and

$$Var(n^*|\alpha) = \sum_{i=1}^{n} \frac{\alpha(i-1)}{(\alpha + i - 1)^2} \approx \alpha \left\{\log\left(\frac{\alpha + n}{\alpha}\right) - 1\right\}. \tag{13}$$

As noted by Florens et al. (1992), however, the approximation in (12) may be dangerous when α is considered a function of n. For instance, (12) gives 0 instead of 1 with $\alpha = \frac{1}{n}$. Better approximations may be obtained by noting that $E(n^*|\alpha) = \sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1} = \alpha\left\{\psi_0(\alpha+n) - \psi_0(\alpha)\right\}$ (Florens et al., 1992) and $Var(n^*|\alpha) = \sum_{i=1}^{n} \frac{\alpha(i-1)}{(\alpha+i-1)^2} = \alpha\left\{\psi_0(\alpha+n) - \psi_0(\alpha)\right\} + \alpha^2\left\{\psi_1(\alpha+n) - \psi_1(\alpha)\right\}$, where $\psi_0(.)$ and $\psi_1(.)$ represents the digamma and trigamma function, respectively. Using these results, an approximation based on a first-order Taylor series expansion, and the fact that a priori $E(\alpha \mid a_0, b_0) = \frac{a_0}{b_0}$ and $Var(\alpha \mid a_0, b_0) = \frac{a_0}{b_0^2}$ we get

$$E(n^*) \approx \frac{a_0}{b_0}\left\{\psi_0\left(\frac{a_0 + nb_0}{b_0}\right) - \psi_0\left(\frac{a_0}{b_0}\right)\right\} \tag{14}$$

and

$$\begin{aligned} Var(n^*) \approx\ & \frac{a_0}{b_0}\left\{\psi_0\left(\frac{a_0 + nb_0}{b_0}\right) - \psi_0\left(\frac{a_0}{b_0}\right)\right\} + \frac{a_0^2}{b_0^2}\left\{\psi_1\left(\frac{a_0 + nb_0}{b_0}\right) - \psi_1\left(\frac{a_0}{b_0}\right)\right\} + \\ & \left\{\frac{a_0}{b_0}\left[\psi_1\left(\frac{a_0 + nb_0}{b_0}\right) - \psi_1\left(\frac{a_0}{b_0}\right)\right]\right. \\ & \left. + \psi_0\left(\frac{a_0 + nb_0}{b_0}\right) - \psi_0\left(\frac{a_0}{b_0}\right)\right\}^2 \frac{a_0}{b_0^2}. \end{aligned} \tag{15}$$

Equating these expressions with prior judgement at the mean and variance of n^* it is possible to obtain the corresponding values for a_0 and b_0. These expressions could be used in order to evaluate the robustness of the model to the specification of prior distribution for the precision parameter.

2.6 Posterior Inference

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. To explore the posterior distribution $p\left(\boldsymbol{Z}, \boldsymbol{\theta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\vartheta} | \boldsymbol{y}\right)$ we used a Gibbs sampling approach based on sampling from the appropriate full conditional distributions. These are obtained by considering the finite dimensional posterior that arises after integrating out the random measure G,

$$\begin{aligned} p\left(\boldsymbol{Z}, \boldsymbol{\theta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\vartheta} | \boldsymbol{y}\right) \quad \propto & \quad \prod_{i=1}^{n} p\left(\boldsymbol{Y}_i | \boldsymbol{Z}_i\right) \prod_{i=1}^{n} p\left(\boldsymbol{Z}_i \mid \boldsymbol{\theta}_i, \boldsymbol{\beta}\right) p\left(\boldsymbol{\theta} \mid \alpha, \boldsymbol{\vartheta}\right) \\ & \quad p\left(\alpha\right) p\left(\boldsymbol{\vartheta}\right) p\left(\boldsymbol{\beta}\right), \end{aligned} \tag{16}$$

where $p\left(\boldsymbol{\theta} \mid \alpha, \boldsymbol{\vartheta}\right)$ arises by exploiting the Polya urn representation of DP (Blackwell and MacQueen, 1973) and the other factors are defined by expressions (1), (3), (4), and (8) - (11). Blackwell and MacQueen (1973) discovered a fundamental connection between the DP and the sampling of balls from an urn. Their result shows that if G is a DP with base measure $\alpha G_{\boldsymbol{\vartheta}}$, then a sample $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n$ generated from the following conditional distributions

$$\boldsymbol{\theta}_i \mid \boldsymbol{\theta}^{(-i)} \sim \frac{\alpha}{\alpha + n - 1} G_{\boldsymbol{\vartheta}} + \frac{1}{\alpha + n - 1} \sum_{j=1, j \neq i}^{n} \delta_{\boldsymbol{\theta}_j}, \quad i = 1, \ldots, n, \tag{17}$$

is a random sample from G. Naturally, the importance of this result is that for computation, reference can be made to a space of finite, rather than infinite, dimensions. Essentially, the random G has been integrated out.

In the rest of this section, we provide details on some of the resulting conditional distributions and the implementation of the Gibbs sampler.

2.6.1 Updating $\boldsymbol{Z}$

To update the latent data vector $\boldsymbol{Z}$, note that the full conditional distribution of $\boldsymbol{Z}_i$ depends only on $\boldsymbol{Y}_i$, $\boldsymbol{\beta}_{0i}$, $\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}_i$, and corresponds to a truncated multivariate normal distribution,

$$\boldsymbol{Z}_i \mid \boldsymbol{y}_i, \boldsymbol{\beta}_{0i}, \boldsymbol{\Sigma}_i, \boldsymbol{\beta} \sim N_k\left(\boldsymbol{\beta}_{0i} + \boldsymbol{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}_i\right) \prod_{j=1}^{k} I\left\{Z_{ij} \in A_{y_{ij}}\right\}. \tag{18}$$

These conditional distributions are obtained by considering each of its coordinates, conditional on the rest (Geweke, 1991). The sampling scheme consists of a cycle of Gibbs steps through the components of $\boldsymbol{Z}_i$, which have truncated univariate normal distributions.

2.6.2 Updating $\boldsymbol{\theta}$ and the hyperparameters

Updating the latent mixture parameters $\boldsymbol{\theta}$ and the hyperparameters α and $\boldsymbol{\vartheta}$ proceeds with standard posterior simulation methods for DP mixtures (see, e.g., MacEachern and Müller, 1998). The discrete nature of the DP implies positive probabilities for ties among the $\boldsymbol{\theta}_i$. Let $n^* \leq n$ be the number of different values or clusters among the $\boldsymbol{\theta}_i$. Denote the set of different values by $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_{n^*}^*)$, let $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)$ be a vector of configuration indicators with $\xi_i = j$ if and only if $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j^*$, and let n_r be the size of the r^{th} cluster (the number of $\xi_i = r$). Then $(\boldsymbol{\theta}^*, \boldsymbol{\xi})$ is an equivalent representation of $\boldsymbol{\theta}$, with $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{\xi_i}^*$. Note that under this alternative parametrization $\boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_{n^*}^*$ are independently drawn from $G_{\boldsymbol{\vartheta}}$. Then the conditional posterior distribution of $\boldsymbol{\theta}_j^*$ is obtained by combining the baseline prior $G_{\boldsymbol{\vartheta}}$ with the likelihood $\prod_{i:\xi_i=j} p(\boldsymbol{Z}_i \mid \boldsymbol{\theta}_j^*, \boldsymbol{\beta})$. That is, it is the posterior based on a random sample of the latent variables which are drawn from the same $\boldsymbol{\theta}_j^*$. All of the conditional distributions are straightforward to derive and sample from. More details can be found in MacEachern and Müller (1998).

2.6.3 The posterior predictive distribution for future discrete responses

We next turn to the posterior predictive distribution for a future observation vector $\boldsymbol{Y}_0$ given $\boldsymbol{X}_0$. Denote by $\boldsymbol{Z}_0$ the associated latent vector. The assumptions of model (1), (3), (4), and (8)- (11) yield $p(\boldsymbol{Y}_0, \boldsymbol{Z}_0 \mid \boldsymbol{y}) = p(\boldsymbol{Y}_0 \mid \boldsymbol{Z}_0) p(\boldsymbol{Z}_0 \mid \boldsymbol{y})$, where $p(\boldsymbol{Z}_0 \mid \boldsymbol{y})$ is the posterior predictive distribution of $\boldsymbol{Z}_0$ that can be developed using the structure induced by the DP prior. Let $\boldsymbol{\phi} = (\boldsymbol{\theta^*}, \boldsymbol{\xi}, \alpha, \boldsymbol{\beta}, \boldsymbol{\vartheta})$ be the entire set of parameters, then

$$p(\boldsymbol{Z}_0 \mid \boldsymbol{y}) = \int\int p(\boldsymbol{Z}_0 \mid \boldsymbol{\theta}_0, \boldsymbol{\beta})\, p(\boldsymbol{\theta}_0 \mid \boldsymbol{\phi})\, p(\boldsymbol{\phi} \mid \boldsymbol{y})\, d\boldsymbol{\theta}_0 d\boldsymbol{\phi}, \tag{19}$$

where $p(\boldsymbol{Z}_0 \mid \boldsymbol{\theta}_0, \boldsymbol{\beta})$ is a k-variate normal distribution and

$$\boldsymbol{\theta}_0 \mid \boldsymbol{\phi} \sim \frac{\alpha}{\alpha + n} G_{\boldsymbol{\vartheta}} + \frac{1}{\alpha + n} \sum_{r=1}^{n^*} n_r \delta_{\boldsymbol{\theta}_r^*}. \tag{20}$$

Expressions (19) and (20) readily provide draws from $p\left(\boldsymbol{Z}_0 \mid \boldsymbol{y}\right)$ and Monte Carlo approximations to $p\left(\boldsymbol{z}_0 \mid \boldsymbol{y}\right)$ for any grid of values $\boldsymbol{z}_0$. If more general inferences on the distribution of the latent variables, $F_{\boldsymbol{X}_i}$, are needed, alternatives approaches can be used. See, e.g., Guglielmi and Tweedie (2001), Gelfand and Kottas (2002), Guglielmi et al. (2002), and Regazzini et al. (2002). Note also that expressions (19) and (20) help to clarify the nature and the amount of learning implied by the model. The predictive distribution for the latent variable emerges by averaging, with respect to the posterior distribution of the parameters $\boldsymbol{\phi}$, the distribution

$$p\left(\boldsymbol{Z}_0 \mid \boldsymbol{\phi}\right) \quad = \quad \frac{\alpha}{\alpha+n}\int p\left(\boldsymbol{Z}_0 \mid \boldsymbol{\theta}_0, \boldsymbol{\beta}\right) dG_{\boldsymbol{\vartheta}}\left(\boldsymbol{\theta_0}\right) + \frac{1}{\alpha+n}\sum_{r=1}^{n^*} n_r p\left(\boldsymbol{Z}_0 \mid \boldsymbol{\theta_r^*}, \boldsymbol{\beta}\right).$$

Expression (21) corresponds to a mixture of multivariate normal distributions, specified by the different locations, $\boldsymbol{\beta}_0^*$, and covariance structure, $\boldsymbol{\Sigma}^*$, with an additional term that allows for a new component. The weight for this additional term, $\frac{\alpha}{\alpha+n}$, decreases as the sample size increases. This corresponds to an appealing feature of the model because it is expected that the chance of discovering a new pattern in a future observation decreases as the amount of observed data increases.

2.6.4 Model Choice

Finally, regarding formal model determination, Basu and Chib (2003) discuss the use of Bayes factors for the DP mixture model. Alternatively, a cross validation model comparison criteria could be used. Indeed in the present work we have adopted the pseudo-Bayes factor (PsBF) (see, e.g., Geisser and Eddy 1979; Gelfand and Dey 1994) for model comparison. The PsBF for model M_1 versus model M_2 is defined as $PsBF_{M_1,M_2} = \prod_{i=1}^{n} \frac{p_{M_1}(\boldsymbol{Y}_i|\boldsymbol{Y}_{-i})}{p_{M_2}(\boldsymbol{Y}_i|\boldsymbol{Y}_{-i})}$, where $p_{M_r}\left(. \mid \boldsymbol{Y}_{-i}\right)$ is the posterior predictive distribution under model M_r based on the data vector $\boldsymbol{Y}_{-i}$ that results after excluding the ith observation $\boldsymbol{Y}_i$. The individual ratio of cross-validation predictive densities known as conditional predictive ordinates (CPO) have also been used. The CPOs measure the influence of individual observations and are often used as predictive model checking tools. The evaluation of these expressions involves the computation of multivariate normal probabilities, which was carried out here by using the methodology described in Genz (1993).

3 Applications

In order to evaluate the performance of the semiparametric model (model M_1) developed here, we present results from two data sets in Sections 3.1 and 3.2. We use the MCMC algorithm of Section 2.6 to fit the models. In addition, we considered the two parametric models that result as limiting cases of model M_1. As discussed in Section 2.5, these are the multivariate probit model (model M_2) and the exchangeable mixture model (model M_3), under the normalization of the conditional variances.

As pointed out by Berger and Guglielmi (2001), for appropriate model comparison it is desirable, if possible, to match the prior specifications in the two models, at least for similar parameters. Here we are going to compare the DP mixture model (model M_1) with alternatives parametric models (models M_2 and M_3) and the DP mixture model should be a generalization of the parametric ones. Because the relevant parametric alternatives (e.g., probit or logit models) do not consider prior distributions given in expressions (10) and (11), we will consider the DPM model defined by (1), (3), (4), and (8)- (9).

3.1 A Simulated Data Set

We tested our DP mixture model using a simulated data for the underlying latent variables. We set $k = 2$ and generated $n = 200$ latent observations from a mixture of two bivariate normals, with equal weights. The mean vectors are $(-0.75, -0.75)$ and $(0.5, 0.5)$ and the covariance matrices are

$$\begin{pmatrix} 1.000 & -0.375 \\ -0.375 & 1.000 \end{pmatrix} \text{ and } \begin{pmatrix} 0.500 & 0.175 \\ 0.175 & 0.500 \end{pmatrix}.$$

Additionally, we included a discrete covariate uniformly distributed between -0.5 to 0.5 with intervals of 0.1. The values for the associated regression coefficients were $\beta_{11} = \beta_{12} = 1.5$. For each of the models, the respective MCMC scheme was run with four independent chains, with randomly chosen starting points and a burn-in period of length 20,000. Samples were saved every 20 iterations until completing a Monte Carlo sample of size 2,500 in all cases. Convergence was assessed using standard criteria (Cowles and Carlin, 1996) as implemented in the BOA package (Smith, 2005).

Posterior inference was quite robust to different values of prior hyperparameters. Figure 1 shows the posterior predictive density $p(\boldsymbol{z}_0 \mid \boldsymbol{y}, \boldsymbol{x}_0)$, for a subject with average covariates, under four alternative Gamma priors for the precision parameter α. Specifically, with $(a_0, b_0) = (2.0, 1.8)$, $(2.0, 3.5)$, $(5.0, 5.5)$ and $(15, 7.5)$. The predictive $p(\boldsymbol{z}_0 \mid \boldsymbol{y}, \boldsymbol{x}_0)$ is estimated as an average over conditional predictives,

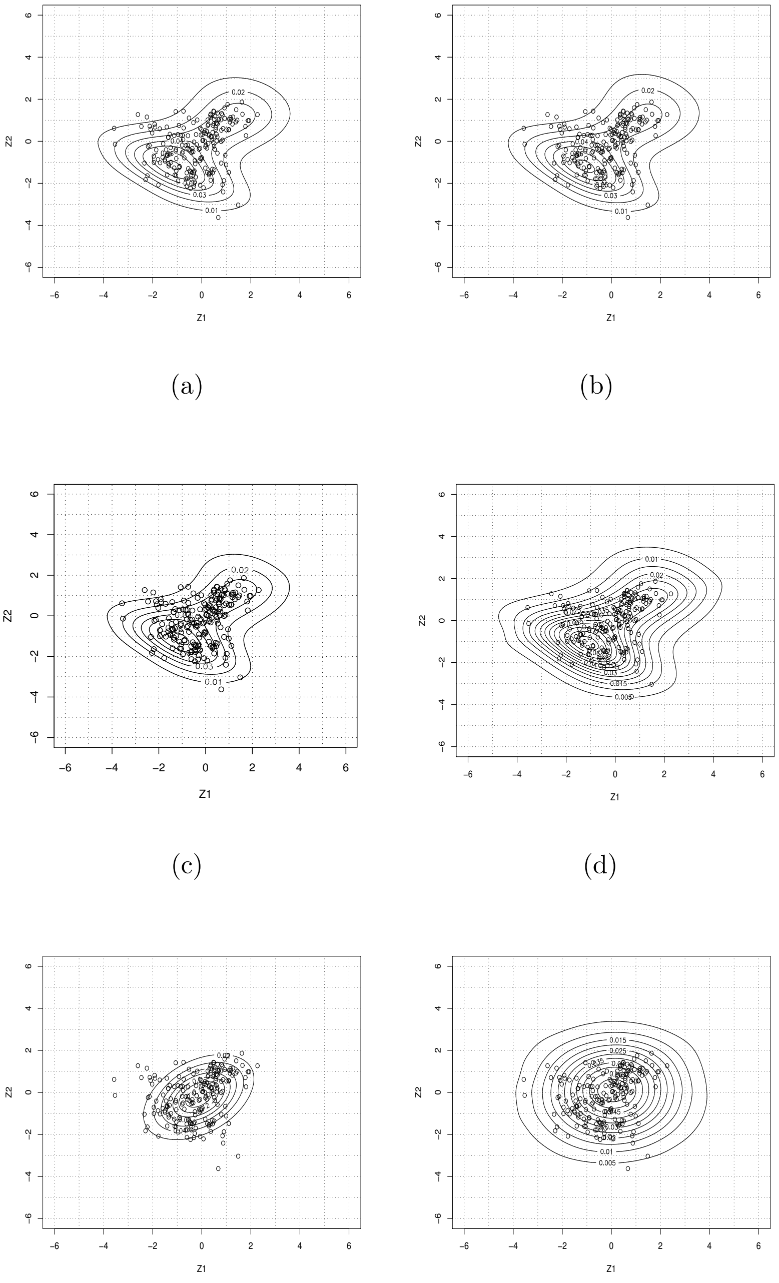
$$\begin{aligned} p(\boldsymbol{z}_0 \mid \boldsymbol{y}, \boldsymbol{x}_0) &= \int\int p(\boldsymbol{z}_0 \mid \boldsymbol{\theta}_0, \boldsymbol{\beta}, \boldsymbol{x}_0)\, p(\boldsymbol{\theta}_0, \boldsymbol{\beta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}_0 d\boldsymbol{\beta} \\ &\approx \frac{1}{T}\sum_{t=1}^{T} p\left(\boldsymbol{z}_0 \mid \boldsymbol{\theta}_0^t, \boldsymbol{\beta}^t, \boldsymbol{x}_0, \boldsymbol{y}\right), \end{aligned}$$

where $(\boldsymbol{\theta}_0^t, \boldsymbol{\beta}^t)$ are the imputed values after t scans of the Gibbs sampler scheme. The four prior settings yielded an expected (sd) prior number of clusters of 6.3 (3.9), 4.0 (2.5), 5.5 (2.7), and 9.8 (3.3), respectively. The corresponding $(0.05, 0.50, 0.95)$ posterior percentiles for n^* were $(3, 6, 13)$, $(2, 5, 11)$, $(3, 6, 10)$, and $(5, 9, 14)$, showing that the posterior inference on n^* is 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, we took $\boldsymbol{b} = \boldsymbol{\mu}_{\boldsymbol{\beta}_0} = (0, 0)^T$, $\boldsymbol{\mu}_{\boldsymbol{\delta}} = 0$, $\boldsymbol{S}_{\boldsymbol{\beta}_0} = cI_2$, $\boldsymbol{S}_{\boldsymbol{\delta}} = c$, with $c = 1$, and $\boldsymbol{B} = \text{diag}(10.0, 10.0)$. A sensitivity analysis for the choice of the hyperparameters revealed robustness of the posterior results. Specifically, we took $\boldsymbol{B} = \text{diag}(1000000, 1000000)$ and the results were basically the same. The posterior variance of the hyperparameters (not shown) indicated that the prior choices were indeed vague (the ratio posterior over prior variance was lower than 0.05). The DP mixture model successfully captures the bimodal underlying distribution of the latent variables as can be seen in Figure 1. Clustering in terms of the dependence structure is illustrated in Figure 2, where the association between the binary variables is assessed via the correlation coefficient of the latent variables. The posterior predictive distribution of ρ_0 is bimodal with modes at -0.739 and 0.755. The posterior means of the individual correlation coefficients range from -0.46 to -0.12 for 140 pairs, from -0.04 to 0.00 for 7 pairs, and from 0.15 to 0.26 for 53 cases.

Figures 1 and 2 provide evidence in favor of the DP mixture model. Moreover, we evaluated $\overline{CPO}\,(M_r) = n^{-1}\sum_{i=1}^{n} \log p_{M_r}(\boldsymbol{Y}_i \mid \boldsymbol{Y}_{-i})$. Since $\overline{CPO}\,(M_1) = -1.18$, $\overline{CPO}\,(M_2) = -1.28$ and $\overline{CPO}\,(M_3) = -1.27$ the cross validation criterion favors M_1. The same is true for the $PsBF$, calculated on the base of the posterior predictive distributions. Indeed, the $2\log_{10} PsBF$ for model M_1 versus model M_2 and M_3 was 17.98 and 14.69, respectively.

Finally, we computed the posterior estimate and the 95% highest posterior density (HPD) intervals for the regression coefficients of the continuous covariate. The



(a)

(b)

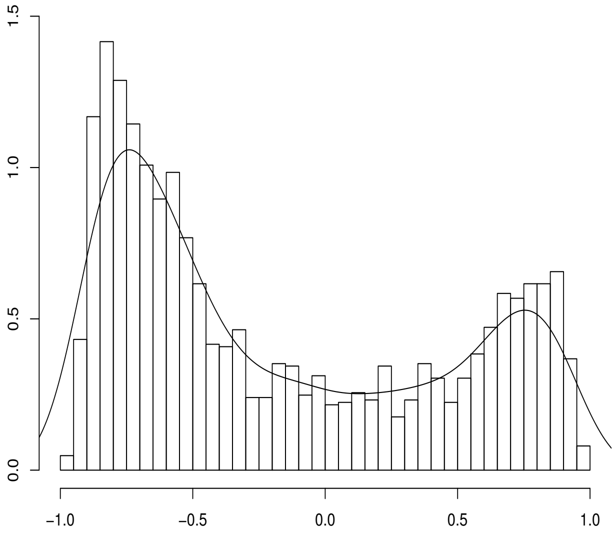
(c)

(d)

(e)

(f)

Figure 1: Simulated data set. Posterior predictive density $p\,(\boldsymbol{z}_0 \mid \boldsymbol{y}, \boldsymbol{x}_0)$ results for the mixture model with a Gamma(2.0,1.8), Gamma(2.0,3.5), Gamma(5.0,5.5), and Gamma(15,7.5) prior for α (panels (a)-(d), respectively), and the parametric limiting cases of the DPM model when $\alpha \to 0^+$ (panel(e)) and $\alpha \to \infty$ (panel(f)). In all cases, $p\,(\boldsymbol{z}_0 \mid \boldsymbol{y}, \boldsymbol{x}_0)$ is overlaid on a plot of the realizations of the latent variables.



(a)

Figure 2: Simulated data set. Posterior predictive distribution for the correlation coefficient under the DP mixture model with a Gamma(2.0,1.8) prior for α.

results are shown in Table 1. While the significance of the regression coefficients is consistent across the models and the 95% HPD intervals overlap, the regression coefficients for the probit model (model M_2) were much lower than the DP mixture model.

3.2 The Signal Tandmobiel® Study

The Signal Tandmobiel® study is a 6-year longitudinal oral health study, involving 4468 children, conducted in Flanders (Belgium) between 1996 and 2001. Dental data were collected on gingival condition, dental trauma, tooth decay, presence of restorations, missing teeth, stage of tooth eruption, orthodontic treatment need, etc. All by using established criteria. Further, information on oral hygiene and dietary behavior was collected from a questionnaire filled-in by the parents. The children were examined annually during their primary school time by one of sixteen trained and half yearly calibrated dental examiners. The average age of the children at entry was 7.1 years (sd=0.4). More details on the Signal Tandmobiel® study can be found in Vanobbergen et al. 2000.

In this study one of the response of interest is caries experience (CE), which

Table 1: Simulated data set. Bayesian posterior estimates for the regression coefficients.

Model	Coefficient	Mean	Std Dev	Median	95% HPD Interval
M_1	β_{11}	1.877	0.647	1.790	(0.685 - 3.120)
	β_{12}	1.286	0.482	1.222	(0.433 - 2.268)
M_2	β_{11}	0.939	0.235	0.945	(0.481 - 1.389)
	β_{12}	0.861	0.226	0.857	(0.418 - 1.297)
M_3	β_{11}	1.459	0.356	1.453	(0.750 - 2.137)
	β_{12}	1.098	0.297	1.097	(0.530 - 1.694)

is defined as a binary variable indicating whether a tooth is decayed, filled, or missing due to caries. This information was recorded for each tooth and for each child, leading to clustered binary data (teeth within mouth). It is of interest to assess the associations in the different teeth for CE because it can help the dentists in optimizing their clinical examination and will direct them in preventive and restorative actions. From a scientific viewpoint, the exploration of CE patterns in the mouth is useful for refining the understanding of the etiology of the disease. Indeed, at present, it is still not yet established whether caries is a spatially local disease in the mouth or not.

Based on the first year's data of the Signal Tandmobiel® study, García-Zattera et al. (2005) examined the associations in the eight deciduous molars for CE. At first sight, the results were puzzling since, besides a high association between adjacent and contra-lateral molars, there was also a high association between vertically opponent and diagonally opponent molars. The first two associations are known and relatively easy to explain from a dental point of view (Psoter et al., 2003). However, the third and especially the fourth association is more difficult to understand. In fact, the association between diagonally opponent teeth was believed to be the result of omitting important child-specific covariates and/or the (assumed) transitivity of the associations.

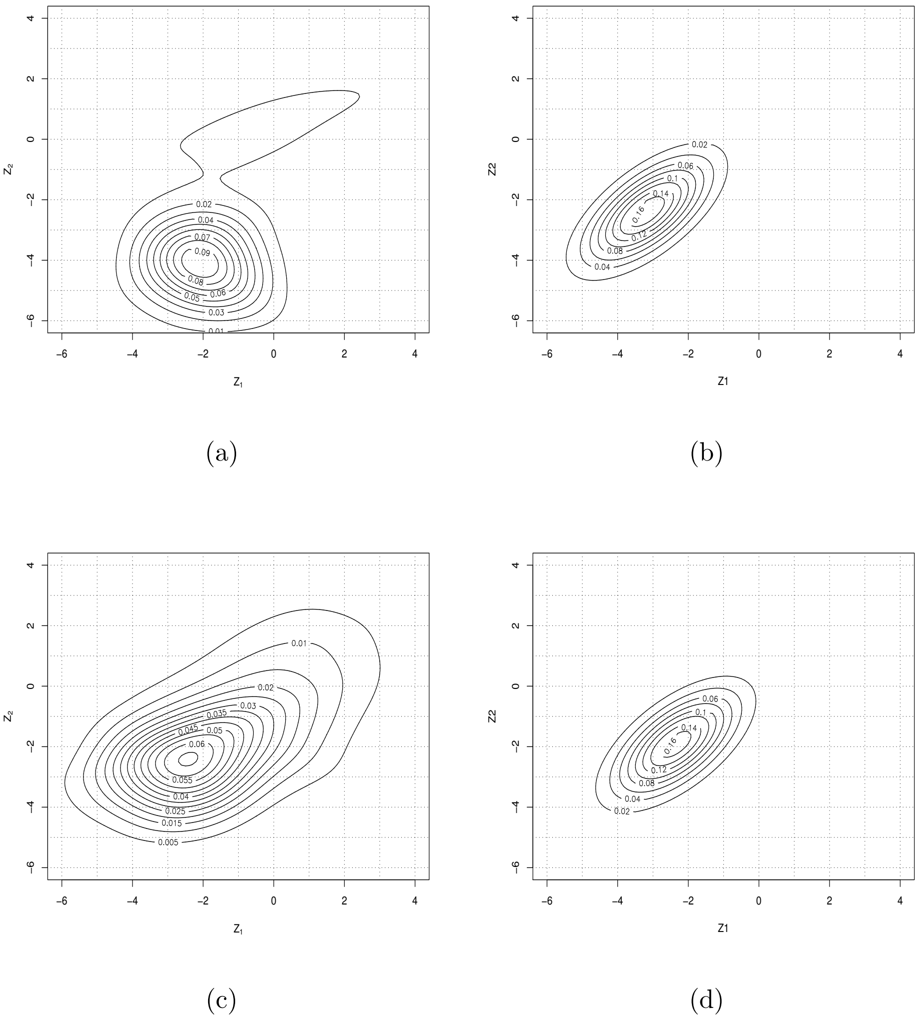
For ease of exposition we have restricted our attention to the CE in the adjacent molars: tooth 54 (fourth molar in upper right corner of mouth) and tooth 55 (fifth molar in upper right corner of mouth), and diagonal opponent second deciduous molars, tooth 55 and tooth 75 (fifth molar in lower left corner of mouth), according to the European notation. While we wish to evaluate the effect of covariates on the probability of developing caries, the emphasis is on the evaluation of the association structure. In particular, we wish to know whether caries of one tooth increases the likelihood of caries on another tooth (not necessarily adjacent), and if there is a unique common association structure for the different teeth combinations. The covariates included in the model are age (in years) (Age), gender (boys versus girls) (Gender), age at start of brushing (in years) (Startbr), regular use of fluoridated supplements (yes versus no) (Sysfl), daily use of sugar containing drinks (no versus yes) (Drinks), number of between-meal snacks (two or less than two a day versus more than two a day) (Meals) and frequency of tooth brushing (once or more a day versus less than once a day) (Freqbrus).

The priors for the hyperparameters and the MCMC specifications were the same as in Section 3.1. In addition, we have taken a Gamma(15,10) prior for α, which yields a expected (sd) prior number of cluster of 12.1 (4.1). Experimentation with other prior choices for the hyperparameters revealed robustness of the posterior results. The results of fitting the DPM model are shown in Table 2. The regular use of fluoridated supplements and the consumption of sugar containing drinks, were significant for the three teeth. These results imply that the probability of developing caries is higher for children who took sugar drinks in-between meals and who did not use fluoridated supplements. The age at start brushing had a significant impact on CE in teeth 54 and 75.

In Figure 3 the posterior predictive distribution $p(\boldsymbol{z}_0 \mid \boldsymbol{y}, \boldsymbol{x}_0)$ and in Figure 4 the posterior predictive distribution for the correlation coefficient $p(\rho_0 \mid \boldsymbol{y})$ is shown for the two pairs of molars and for two models, i.e. the DP mixture model and the conventional probit model. The conventional bivariate probit models show a similar behavior for the adjacent pair of deciduous molars (54 and 55) and the diagonally opponent deciduous molars (55 and 75). As mentioned before, this is difficult to understand from a dental point of view. In contrast, the DP mixture model identifies two components with respect to the location and with respect to the correlation coefficient for the adjacent pair of teeth (54 and 55). Further, even though two clusters dominate the inference, note that the posterior predictive distribution is concentrated on the positive support. For the diagonally opponent second deciduous molars (55 and 75), the picture was completely different. In this case, one cluster dominates the posterior inference (see Figure 3(c)) and the posterior predictive distribution of the correlation coefficient looks less informative (see Figure 4(c)) with

Table 2: Signal Tandmobiel® study: Posterior mean (95% HPD) of the multiple regression model coefficients fitted to teeth 54, 55 and 75 using the DP mixture model.

Covariate	T54	T55	T75
Age	0.056	-0.030	-0.105
(years)	(-0.051 ; 0.161)	(-0.123 ; 0.060)	(-0.189; 0.024)
Gender	0.022	0.010	0.119
(girls versus boys)	(-0.081 ; 0.124)	(-0.086 ; 0.109)	(0.027 ; 0.205)
Regular use of fluoridated	0.164	0.183	0.271
supplements (no versus yes)	(0.052 ; 0.265)	(0.079 ; 0.282)	(0.171 ; 0.365)
Daily consumption of sugar	0.257	0.240	0.171
containing drinks (yes versus no)	(0.145 ; 0.367)	(0.133 ; 0.347)	(0.073 ; 0.266)
Intake of in-between-meals	0.114	0.081	0.040
($>$ 2 versus $\leq$ 2 a day)	(0.000 ; 0.222)	(-0.029 ; 0.184)	(-0.054 ; 0.136)
Frequency of brushing	0.003	0.056	0.069
($<$ 1 versus $\geq$ 1 a day)	(-0.149 ; 0.147)	(-0.089 ; 0.190)	(-0.058 ; 0.192)
Age at start brushing	0.061	0.031	0.119
(years)	(0.009 ; 0.106)	(-0.016 ; 0.075)	(0.074 ; 0.161)



(a)

(b)

(c)

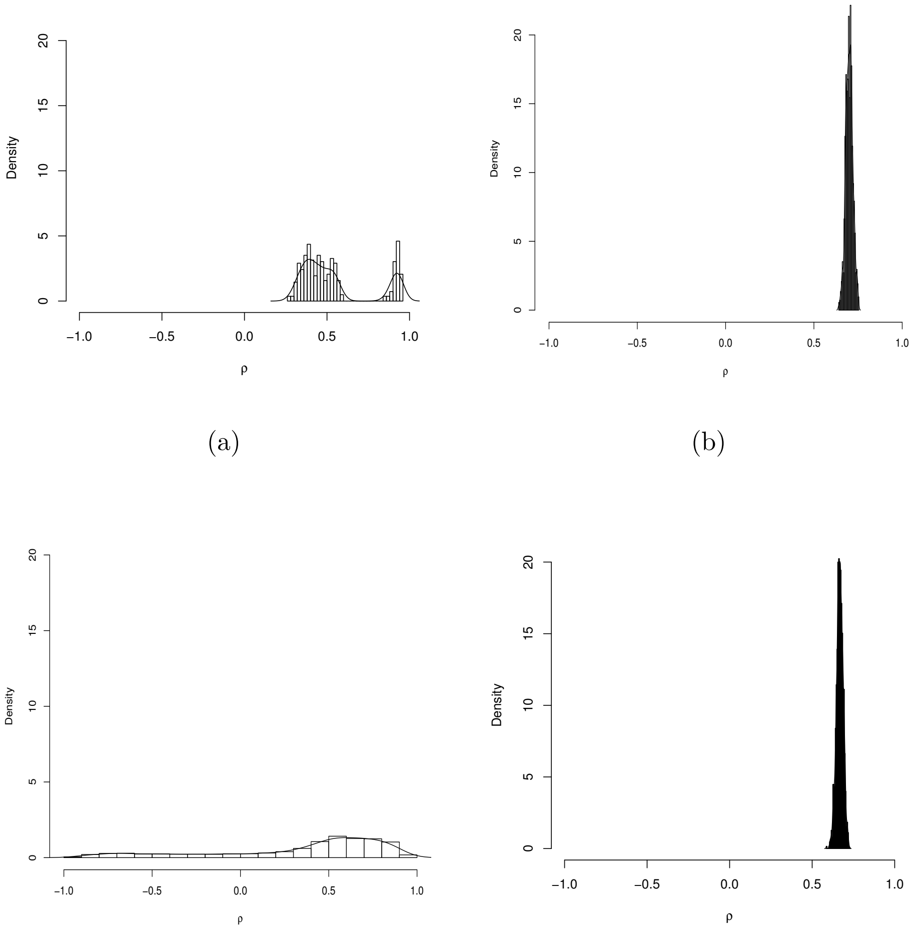
(d)

Figure 3: The Signal Tandmobiel® study. Posterior predictive density $p\left(\boldsymbol{z}_0 \mid \boldsymbol{y}, \boldsymbol{x}_0\right)$ results for the DP mixture model for adjacent and diagonally opponent molars (panels (a) and (c), respectively), the conventional probit model (panels (b) and (d), respectively)

a positive probability of being negative. Indeed, the posterior predictive probability $P(\rho_0 < 0 \mid \boldsymbol{y}) = 0.224$.

In both pairs of teeth, the results supported the departure of the normality assumption of the latent variable distribution. The cross validation model comparison criteria preferred the DP mixture model. The PsBF calculated on the base of the posterior predictive distributions, strongly confirmed this. The $2\log_{10} PsBF$ for the DP mixture model versus the bivariate probit model was 153.09 and 93.60 for the adjacent and diagonally opponent pair, respectively.

Note also that in both pairs the posterior predictive distribution of the correlation coefficient under the bivariate probit model overestimates the association structure (see Figures 4(b) and 4(d)), suggesting that the high association in the



(a)

(b)

(c)

(d)

Figure 4: The Signal Tandmobiel® study. Posterior predictive density $p(\rho_0 \mid \boldsymbol{y})$ results for the DP mixture model for adjacent and diagonally opponent molars (panels (a) and (c), respectively), the conventional probit model (panels (b) and (d), respectively)

bivariate probit model is partly due to ignoring unobserved confounders and effect modifiers defining subpopulations of individuals with different spatial patterns in CE, even after adjusting for the effects of the known covariates. Although the existence of groups could be due to the multifactorial etiology of caries, note that this is just a convenient explanation to justify the use of a mixture model. In any case, the ability to identify individual CE patterns raises issues for further investigation. Moreover, the results suggest that inference under the usual normality assumption of latent variable could be misleading.

4 Concluding remarks

We have proposed a semiparametric Bayesian approach to model multivariate binary data. The core of the nonparametric component is the introduction of a Dirichlet process mixture model for latent variables defining classification groups with respect to the location and the correlation coefficients.

The approach has been successfully applied to the examples of this paper, favoring the DP mixture model over the conventional multivariate probit (and logit, but not shown) specification. By specifying the mixture with respect to both the location and the association of the normal kernel, our approach can also be a useful tool for handling and detection of outliers in multivariate binomial data. In this context, Aitkin and Wilson (1980) first suggested using a finite mixture model as a way of handling data with multiple outliers, especially when some of the outliers group into clumps. Unlike to this approach and other model-based clustering applications (see, e.g., Dasgupta and Raftery, 1998), our approach does not require choosing a number of groups in advance. As the proposed model combines both, regression parameter estimation and the construction of clusters, its robustness against outliers and model misspecification is expected. This is subject of ongoing research.

The model can be extended in several ways. For example, to analyse multivariate binary data jointly with ordinal and/or continuous variables. Alternatively, the DP mixture specification could also include the regression coefficient vector associated to covariates. This would allow that the effect of the covariables could be different across the clusters. These and other extensions are the subject of the current work.

Acknowledgements

The authors thank Fernando Quintana for his input in this work. The authors are grateful to the two referees for many detailed suggestions and helpful comments that have led to a considerable improvement in the manuscript.

The first two authors have been partly supported by the Research Grant OT/00/35, Catholic University Leuven. The first author has also been supported by the KUL-PUC bilateral (Belgium-Chile) grant BIL05/03 and the second and third author by the Research Grant OT/05/60, Catholic University Leuven. The authors also acknowledge the partial support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs. Data collection was supported by Unilever, Belgium. The Signal Tandmobiel® study comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

References

Aitkin, M. and Wilson, G. T.: 1980, Mixture model, outliers, and the EM algorithm, *Technometrics* **22**, 325–331.

Albert, J. H.: 1992, Bayesian estimation of the polychoric correlation coefficient, *Journal of Statistical Computation and Simulation* **44**, 47–61.

Ashford, J. R. and Sowden, R. R.: 1970, Multi-variate probit analysis, *Biometrics* **26**, 535–546.

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.

Berger, J. O. and Guglielmi, A.: 2001, Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives, *Journal of the American Statistical Association* **96**, 174–184.

Blackwell, D. and MacQueen, J.: 1973, Ferguson distributions via Pólya urn schemes, *The Annals of Statistics* **1**, 353–355.

Chen, M.-H.: 2004, Skewed link modeels for categorical response data, *in* M. G. Genton (ed.), *Skew-elliptical Distributions and Their Applications: A Journey Beyond Normality*, Chapman and Hall/CRC, pp. 131–151.

Chen, M.-H. and Dey, D.: 1998, Bayesian modelling of correlated binary responses via scale mixture of multivariate normal link functions, *Sankhyā* **60**, 322–343.

Chib, S. and Greenberg, E.: 1998, Analysis of multivariate probit models, *Biometrika* **85**, 347–361.

Cowles, M. K. and Carlin, B. P.: 1996, Markov chain Monte Carlo convergence diagnostics: a comparative study, *Journal of the American Statistical Association* **91**, 883–904.

Dasgupta, A. and Raftery, A. E.: 1998, Detecting features in spatial point process with clutter via model-base clustering, *Journal of the American Statistical Association* **93**, 294–302.

Dey, D., Müller, P. and Sinha, D.: 1998, *Practical Nonparametric and Semiparametric Bayesian Statistics*, Springer.

Ferguson, T. S.: 1973, A Bayesian analysis of some nonparametric problems, *The Annals of Statistics* **1**, 209–230.

Florens, J.-P., Mouchart, M. and Rolin, J.-M.: 1992, Bayesian analysis of mixtures: Some results on exact estimability and identification, *in* J. M. Bernardo, J. O. Berger and S. A. F. M (eds), *Proceedings of the Fourth Valencia International Meeting*, Clarendon Press, pp. 127–145.

García-Zattera, M. J., Jara, A., Lesaffre, E. and Declerk, D.: 2005, Conditional independence of multivariate binary data with an application in caries research, *Technical report*, Catholic University of Leuven, Biostatistical Centre.

Geisser, S. and Eddy, W.: 1979, A predictive approach to model selection, *Journal of the American Statistical Association* **74**, 153–160.

Gelfand, A. E. and Dey, D.: 1994, Bayesian model choice: asymptotics and exact calculations, *Journal of the Royal Statistical Society, Series B* **56**, 501–514.

Gelfand, A. E. and Kottas, A.: 2002, A computational approach for full nonparametric Bayesian inference under Dirichlet Process Mixture models, *Journal of Computational and Graphical Statistics* **11**, 289–304.

Genz, A.: 1993, Comparison of methods for the computation of multivariate normal probabilities, *Computing Science and Statistics* **25**, 400–405.

Geweke, J.: 1991, Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints, *in* E. Keramidas (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Interface Foundation of North American, Inc., Fairfax.

Guglielmi, A., Holmes, C. and Walker, S.: 2002, Perfect simulation involving functions of a Dirichlet process, *Journal of Computational and Graphical Statistics* **11**, 306–310.

Guglielmi, A. and Tweedie, R.: 2001, Markov chain Monte Carlo estimation of the law of the mean of a Dirichlet process, *Bernoulli* **7**, 573–592.

Hujoel, P. P., Lamont, R. J., DeRouen, T. A., Davis, S. and Leroux, B. G.: 1994, Within-subject coronal caries distribution patterns: An evaluation of randomness with respect to the midline, *Journal of Dental Research* **73 (9)**, 1575–1580.

Kottas, A., Müller, P. and Quintana, F.: 2005, Nonparametric Bayesian modeling for multivariate ordinal data, *Journal of Computational and Graphical Statistics* **14**, 610–625.

Lesaffre, E. and Molenberghs, G.: 1991, Multivariate probit analysis: A neglected procedure in medical statistics, *Statistics in Medicine* **10**, 1391–1403.

Liechty, J., Liechty, M. and Müller, P.: 2004, Bayesian correlation estimation, *Biometrika* **91**, 1–14.

MacEachern, S. N. and Müller, P.: 1998, Estimating mixture of Dirichlet Process Models, *Journal of Computational and Graphical Statistics* **7 (2)**, 223–338.

McLachlan, G. J. and Peel, D.: 2000, *Finite Mixture Models*, John Wiley and Sons, New York, USA.

O'Brien, S. and Dunson, D.: 2004, Bayesian multivariate logistic regression, *Biometrics* **60**, 739–746.

Pourahmadi, M.: 1999, Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation, *Biometrika* **86 (3)**, 677–690.

Psoter, W. J., Zhang, H., Pendrys, D. G., Morse, D. E. and Mayne, S. T.: 2003, Classification of dental caries patterns in the primary dentition: a multidimensional scaling analysis, *Community Dent Oral Epidemiol* **31**, 231–238.

Regazzini, E., Guglielmi, A. and Di Nunno, G.: 2002, Theory and numerical analysis for exact distributions of functionals of a Dirichlet Process, *The Annals of Statistics* **30**, 1376–1411.

Rolin, J.-M.: 1992, Some useful properties of the Dirichlet Process, *Technical report*, Core Discussion Paper **9207**, Université Catholique de Louvain, Belgium.

Sethuraman, J.: 1994, A constructive definition of Dirichlet prior, *Statistica Sinica* **2**, 639–650.

Smith, B. J.: 2005, *Bayesian Output Analysis Program (BOA) for MCMC*, College of Public Health, University of Iowa, Iowa, USA.
URL: *http://www.public-health.uiowa.edu/boa*

Vanobbergen, J., Martens, L., Lesaffre, E. and Declerck, D.: 2000, The Signal Tandmobiel project, a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results, *European Journal of Paediatric Dentistry* **2**, 87–96.