Nonparametric spatial models for clustered ordered periodontal data

Dipankar Bandyopadhyay[1,*], Antonio Canale[2]

[1]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA
[2]Department of Economics and Statistics, University of Turin and Collegio Carlo Alberto, Turin, Italy

Abstract

Clinical attachment level (CAL) is regarded as the most popular measure to assess periodontal disease (PD). These probed tooth-site level measures are usually rounded and recorded as whole numbers (in mm) producing clustered (site measures within a mouth) error-prone ordinal responses representing some ordering of the underlying PD progression. In addition, it is hypothesized that PD progression can be spatially-referenced, i.e., proximal tooth-sites share similar PD status in comparison to sites that are distantly located. In this paper, we develop a Bayesian multivariate probit framework for these ordinal responses where the cut-point parameters linking the observed ordinal CAL levels to the latent underlying disease process can be fixed in advance. The latent spatial association characterizing conditional independence under Gaussian graphs is introduced via a nonparametric Bayesian approach motivated by the probit stick-breaking process, where the components of the stick-breaking weights follows a multivariate Gaussian density with the precision matrix distributed as G-Wishart. This yields a computationally simple, yet robust and flexible framework to capture the latent disease status leading to a natural clustering of tooth-sites and subjects with similar PD status (beyond spatial clustering), and improved parameter estimation through sharing of information. Both simulation studies and application to a motivating PD dataset reveal the advantages of considering this flexible nonparametric ordinal framework over other alternatives.

Key words: G-Wishart; Multivariate ordinal; Nonparametric Bayes; Probit stick-breaking; Periodontal disease; Spatial association

1 Introduction

Periodontal disease (PD) is the primary cause of adult tooth loss. It is a collection of inflammatory diseases affecting the periodontium, the aggregate of tissues consisting of the gingiva (or gums), periodontal ligament, cementum and alveolar bone. In Western countries, 35% of the adult population are estimated to develop PD, whereas 10-15% will develop severe periodontitis (Hugoson et al., 1989; Brown et al., 1990). When left untreated, PD causes progressive bone loss around the tooth leading to loosening from the maxillary (upper), or mandibular (lower) jaws, with eventual loss. Progression of PD is usually assessed by dental hygienists using a calibrated periodontal probe via clinical attachment level (or, CAL), the most popular biomarker of PD. CAL is defined as the distance down a tooth's root that is no longer attached to the

[*]*Address of Correspondence*: Department of Biostatistics, Virginia Commonwealth University, 830 East Main Street, One Capitol Square, 7th Floor, PO Box 980032, Richmond, VA 23298-0032, USA. Tel: +1 804 827 2058; Fax: +1 804 828 8900; E-mail: dbandyop@vcu.edu

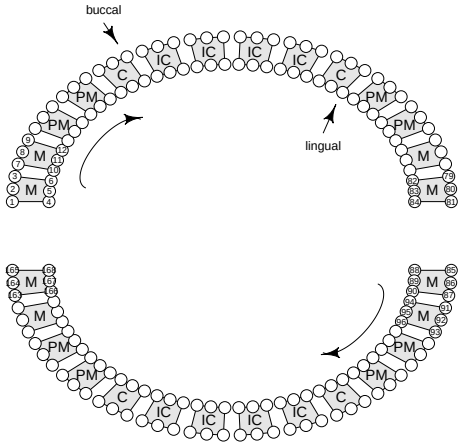
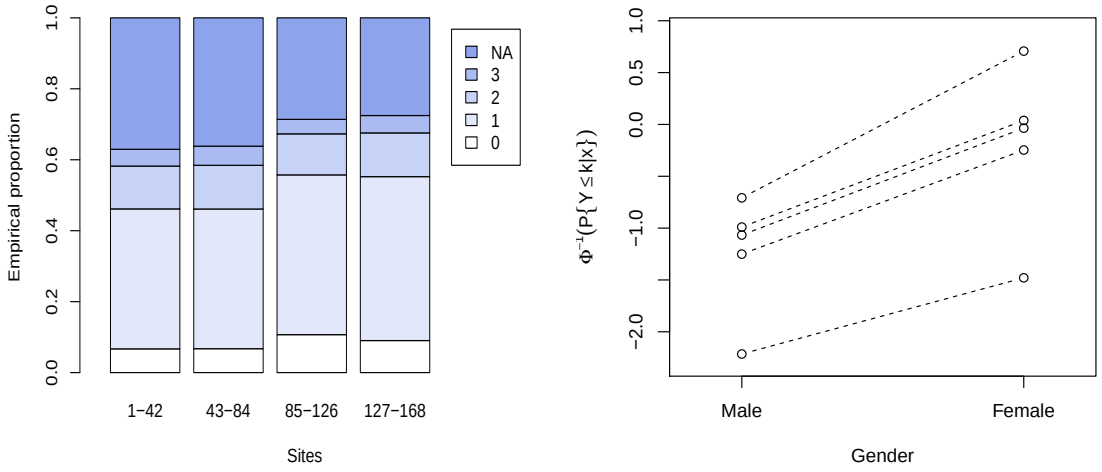


Figure 1: Tooth-types (molars, pre-molars, canines and incisors, abbreviated as M, PM, C and IC, respectively), site-locations (buccal, or cheek side vs lingual, or tongue side), and site numbering (such as 1-6 for the second molar on the upper left quadrant), for a hypothetical subject with no missing teeth.

surrounding bone by the periodontal ligament (Reich and Bandyopadhyay, 2010). During a full periodontal exam, the CAL is measured at six pre-specified tooth-sites for each of the 28 teeth for a subject, excluding the four third-molars (wisdom teeth). Under no missing data, this leads to a multivariate framework with 168 observed CAL observations per subject. Figure 1 depicts the various tooth-types (molars, pre-molars, canines and incisors), site locations (buccal vs lingual), and numbering (such as 1-6, 7-12, etc), from a hypothetical study subject without any missing teeth. Sites 1-84 are located in the maxilla (upper jaw), while sites 85-168 are in the lower jaw (mandible). Lingual refer to those sites that are in direct contact with the tongue, while the ones located opposite are the buccal sites.

A typical periodontal probe used to assess the periodontium health is marked in millimeter increments. Thus, the CAL responses are usually rounded to the nearest millimeter, and might resemble more closely a natural ordering based on the severity of the PD status. The traditional technique for ordinal data regression is to introduce an underlying latent (continuous) variable corresponding to the ordinal outcome via a set of cut-points, and then connecting the data covariates to cumulative probabilities associated with the latent variable and cut-points through a link function (probit, or logit), using both classical and Bayesian techniques (Johnson and Albert, 1999). In the context of ordinal probit regression, the Albert and Chib data augmentation method (Albert and Chib, 1993) induces notoriously high auto-correlations between the cut-points and other parameters, which was improved via the hybrid Gibbs/Metropolis-Hastings (MH) sampling scheme of Cowles (1996). However, analyzing ordinal CAL responses present an array of statistical challenges.

Our motivating example in this paper comes from a clinical study (Fernandes et al., 2009) conducted at the Medical University of South Carolina (MUSC) to determine the PD status of Type-2 diabetic Gullah-speaking African-Americans (henceforth, GAAD study). These data are multidimensional in nature, i.e., each dimension of the multivariate outcome (tooth-sites within a subject) corresponds to a level of a categorical variable. Routine univariate analysis by summing the outcomes for a subject may not be optimal because it ignores both the underlying clustered and ordinal structure. Univariate ordinal models can also be constructed, but that would ignore the joint evolution of clustered responses within a subject (Laffont et al., 2014). Hence, a variety of models were developed to tackle multivariate ordinal data for cross-sectional and repeated measures outcomes, mostly under a latent multivariate Gaussian proposition (Agresti and Natarajan, 2001). Web Figure 1 (see supplementary material associated with this paper) presents the summary



(a) (b)

Figure 2: CAL data. Panel (a) plots the (unadjusted) empirical proportions of site-level ordinal responses, aggregated over 4 quadrants. Numbers 0-3 and NA corresponds to an increasing gradient (starting with white) representing the 5 ordinal categories 'Normal', 'Slight', 'Moderate', 'Severe', and 'Missing', respectively. Panel (b) plots the empirical proportion approximations (straight lines) of the function $f_k(x) = \Phi^{-1}(P\{Y \leq k|x\})$, corresponding to the ordinal category k, where $k = 0, \ldots, K-1$ for the covariate Gender.

quartile plots of the site-level ordinal responses for the whole mouth, across all subjects, from the GAAD study, while Figure 2 (panel-a) summarizes the unadjusted empirical proportions of five ordinal PD categories, aggregated over four quadrants (i.e., half-jaw). We use numbers 0-3 and NA corresponding to an increasing blue gradient (starting with white) that represents the 'Normal (disease free)', 'Slight', 'Moderate', 'Severe' and finally, 'Missing' categories, respectively. We observe that the empirical proportions corresponding to these categories are not homogenous; in particular, the mild category dominates in each of the bars, the proportion of the severe category the least, and the missing category higher for the maxilla (upper jaw, sites 1-84) than the mandible (lower jaw, sites 85-168). These irregular patterns of latent disease status may not be explained by a usual (symmetric) multivariate Gaussian density. Furthermore, the inferential framework for the above models assumes parallelism of the regression lines (i.e., equal slopes) characterizing the ordinal outcomes. This is restrictive, and often violated under our multivariate framework. For example, Figure 2 (panel-b) that plots the value of the function $f_k(x) = \Phi^{-1}(P\{Y \leq k|x\})$, where Y denotes the ordinal random variable representing the PD categories, with $P\{Y \leq k|x\}$ estimated via its empirical proportion for $k = 0, \ldots, K-1$, for the covariate x (here Gender), presents evidence against this assumption. In addition, this ordinal data can exhibit spatial-referencing (Reich and Bandyopadhyay, 2010) i.e., higher category sites tend to be proximally located (see Figure 1 in the Supplementary Material), and failure to account for this clustering might lead to biased estimates (Boehm et al., 2013). Note that, for our example, the spatial lattice we want to incorporate involves replication (i.e., separate spatial lattices for each subject), and is fundamentally different from the traditional disease mapping setting where multiple subjects are observed at each spatial location.

To alleviate the above shortcomings, we propose an ordinal probit regression framework for clustered data where the underlying (continuous) flexible latent process can capture the true non-homogenous (irregular) spatially-referenced disease status beyond the capabilities of a standard spatial model. Our model

development is in the spirit of the nonparametric Bayesian framework of Kottas et al. (2005) and Leon-Novelo et al. (2010), with the ordinal cut-points fixed in advance leading to fine-tuned and straightforward Markov chain Monte Carlo (MCMC) algorithms. Next, the latent probit score is regressed on covariates controlled for spatial random effects that follows the probit stick-breaking process (PSBP) of Rodriguez and Dunson (2011), motivated by applications to areal lattice data. The PSBP follows similar stick-breaking construction popularly used in defining the Dirichlet Process (Ferguson, 1973, 1974) but provides increased flexibility for spatial modeling by replacing the characteristic beta-distributed sticks with probit-transformed elements that are jointly distributed as zero-mean multivariate normal with a G-Wishart hyperprior (Roverato, 2002) for the precision matrix. Under the umbrella of Gaussian graphical models, this G-Wishart prior can be associated with both decomposable and non-decomposable graphs embedding tremendous flexibility in estimating spatial associations for lattice data (Dobra et al., 2011), over the restricted but popular conditionally-autoregressive (CAR) (Banerjee et al., 2014) priors. Our MCMC sampling algorithms utilizes a slice sampler approach for the stick-breaking priors (Kalli et al., 2011) and the efficient propositions of Wang and Li (2012) to sample from the G-Wishart density. This computationally elegant Bayesian nonparametric framework has not been explored earlier for modeling clustered ordinal data, in particular to assess PD status. Furthermore, it can provide new knowledge to periodontists beyond the usual risk assessments in discovering both groups of sites inside mouth and groups of subjects with similar disease status by exploiting the natural clustering induced by the nonparametric priors.

The rest of the article proceeds as follows. Section 2 develops the hierarchical ordinal modeling framework. Section 3 outlines the joint probability model and related MCMC implementation routines. Section 4 applies the model to the motivating GAAD dataset, while Section 5 conducts simulation studies to evaluate the robustness of our methods under alternative specifications of the spatial random effects. Finally, Section 6 concludes this study.

2 Hierarchical Spatial Ordinal Model

Suppose $y_i(s_j)$ be the ordinal CAL measure at site s_j for the ith subject, $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where n and m denote, respectively, the total number of subjects and sites within a subject, and $y_i(\boldsymbol{s})$ represent the response vector for the ith subject where $\boldsymbol{s} = (s_1, s_2, \ldots, s_m)$. The CAL response $y_i(s_j)$ takes the value $k = 0, 1, \ldots, K-1$, where 0 represents the baseline (disease-free) state. Let X be the $n \times p$ design matrix of subject-level covariates, with the ith row x_i representing a vector of p covariates for subject i and Z be the $m \times q$ design matrix of site-level covariates, with the jth row z_j representing a vector of q covariates at location s_j. Here, our objective is robust quantification of covariate effects on the cell probability $P\{y_i(s_j) = k\}$, accounting for the effect of between-site spatial referencing. A standard ordered probit model (ignoring clustering) assumes a multinomial selection process for $y_i(\boldsymbol{s})$, such that $y_i(s_j) \overset{ind}{\sim}$ Multinomial $(1, (\pi_{0ij}, \ldots, \pi_{Kij}))$, with $\sum_{k=0}^{K-1} \pi_{kij} = 1$. However, this independence assumption is violated in our model due to the clustering. Next, assuming a continuous latent variable $y_i^*(s_j)$, $P\{y_i(s_j) = k\}$ is expressed as the probability that $y_i^*(s_j)$ lies in the interval $(a_k, a_{k+1}]$. Thus, we have

$$y_i(s_j) = k \text{ iff } y_i^*(s_j) \in (a_k, a_{k+1}],\ k = 0, \ldots, K-1 \quad (1)$$

$$y_i^*(s_j) = x_i\boldsymbol{\beta} + z_j\boldsymbol{\gamma} + u_i(s_j), \quad (2)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)'$ are $p \times 1$ and $q \times 1$ parameter vectors respectively, and $u_i(s_j)$ is the random-effect term to account for the subject and spatial clustering. We assume $u_i(s_j) \sim P$, where P is a random probability measure.

With a nonparametric approach, we model $P \sim \Pi$, where Π is a prior over the space of density functions. In particular, we assume Π to be a (dependent) probit stick-breaking process (PSBP) (Rodriguez and

Dunson, 2011), i.e.

$$\begin{aligned} P &= \sum_{h=1}^{\infty} w_h(s_j) N(\theta_h, \sigma^2),\ \theta_h \overset{iid}{\sim} N(0,1) \\ w_h(s_j) &= \Phi(\alpha_{hj}) \prod_{l<h} (1 - \Phi(\alpha_{lj})), \end{aligned} \tag{3}$$

for all $j \in \{1, \ldots, m\}$, where $\Phi(\cdot)$ is the cumulative distribution function of standard normal $\mathrm{N}(0,1)$. Note that none of the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in (2) include the intercept, because it is implicitly implied via $\theta_h \overset{iid}{\sim} N(0,1)$. Thus, the stick-breaking weights $w_h(s_j)$ in the PSBP framework are generated from probit transformations of Gaussian distributed α_{hj}'s as opposed to the traditional Sethuraman construction that specifies a Beta$(1, \lambda)$ prior (Sethuraman, 1994; Ishwaran and James, 2001). Also, these weights lead to straightforward construction of predictor-dependencies and facilitates MCMC implementation by simplifying the relevant data augmentation steps (Chung and Dunson, 2009). The summation in the representation of P can be finite (known, or unknown) or infinite, and it can be shown that $\sum_{h=1}^{\infty} w_h(s_j) = 1$, almost surely, via arguments available in Rodriguez and Dunson (2011).

Next, we complete the PSBP specification by incorporating spatial dependencies. Specifically, for each h, we assume

$$\alpha_h = (\alpha_{h1}, \ldots, \alpha_{hm})' \sim \mathrm{MVN}(0, \Sigma) \tag{4}$$

where MVN denotes the multivariate normal density, with Σ the corresponding positive-definite (p.d.) $m \times m$ variance-covariance matrix. Let $\Omega = \Sigma^{-1}$ be the corresponding precision matrix. For multivariate lattice data, a standard way to model spatial association is to assign a conditionally-autoregressive (CAR) prior (Besag, 1974) to Σ, i.e., $\Sigma = \sigma^2_{sp} Q(\rho)^{-1}$, where σ^2_{sp} is the spatial variance parameter, $Q(\rho) = D - \rho W$, D is a diagonal matrix with the jth diagonal entry representing the number of neighbors at location s_j, W is the *adjacency matrix* with entries $w_{jj'} = 1$ if s_j is a neighbor of $s_{j'}$, and $= 0$ otherwise, and ρ is the smoothing parameter controlling the degree of spatial dependence. However, this is restrictive given that the standard CAR model do not allow for non-stationarity, spatially-varying autocorrelation and smoothing parameters. Conditional independence among the elements of α_h can be imposed by assigning $\Omega \sim \text{G-Wish}_W(\kappa, S)$, where $\text{G-Wish}_W(\kappa, S)$ denotes the G-Wishart prior (Roverato, 2002) with symmetric p.d. scale matrix S and degrees of freedom κ, constrained to have null entries for each zero in W. Given the normal distribution for α_h, the G-Wishart is the conjugate prior for Ω constrained under W and is defined as:

$$p(\Omega|W) = I_W(\kappa, S)^{-1} |\Omega|^{\frac{\kappa-2}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(S\Omega)\right\} \mathbb{I}\{\Omega \in M_W\} \tag{5}$$

where $\kappa > 2$ is the degrees of freedom parameter, M_W is the set of symmetric p.d. matrices with zero off-diagonal elements whenever the jj'th element of W is null and $I_W(\kappa, S)$ is the normalizing constant given by $I_W(\kappa, S) = \int |\Omega|^{\frac{\kappa-2}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(S\Omega)\right\} \mathbb{I}\{\Omega \in M_W\} d\Omega$. The G-Wishart specification provides flexibility by removing the stationarity assumption, and enhances computational efficiency via conjugacy and exact marginal likelihood calculations (Carvalho et al., 2007) in our Bayesian framework.

The implication of the PSB prior on the marginal setup is immediate. Marginalizing out the random measure P, the distribution of the latent $y_i^*(s_j)$ is an infinite mixture of Gaussians with locations $x_i\boldsymbol{\beta} + z_j\boldsymbol{\gamma} + \theta_h$ and scale σ^2. Introducing latent indicators $\xi_i(s_j)$ such that $\xi_i(s_j) = h$ if the response for site s_j of subject i comes from the hth mixture component with probability $w_h(s_j)$, the latent $y_i^*(s_j)$ is normally distributed (conditionally), i.e., $(y_i^*(s_j)|\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta_h, \xi_i(s_j) = h) \sim N(x_i\boldsymbol{\beta} + z_j\boldsymbol{\gamma} + \theta_h, \sigma^2)$. Thus, marginalizing with respect to the latent $y_i^*(s_j)$, the probability that site s_j of subject i takes the kth categorical value is given by

$$\mathrm{pr}(y_i(s_j) = k|\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta_h, \xi_i(s_j) = h) = \Phi\left(\frac{a_{k+1} - x_i\boldsymbol{\beta} - z_j\boldsymbol{\gamma} - \theta_h}{\sigma}\right) - \Phi\left(\frac{a_k - x_i\boldsymbol{\beta} - z_j\boldsymbol{\gamma} - \theta_h}{\sigma}\right)$$

In our setup, we fix the cut-point parameters a_{kj} without loss of generality instead of estimating those using the class of data augmentation Gibbs algorithms for probit models (Albert and Chib, 1993). Note that the two extreme cut-points are always fixed at $-\infty$ and ∞ in any setup. Fixing thresholds during estimation has earlier been considered in the context of nonparametric modeling of ordinal (Kottas et al., 2005) and count (Canale and Dunson, 2011) outcomes. Our framework bypasses assessing the parallelism assumptions of regression lines characterizing the ordered categories, which are mostly violated in a typical multivariate setup. For concerns with identifiability, we fix σ^2 to 1 which leads to the choice of the cutpoints as $\{-\infty, -4, 0, 4, 8, \infty\}$, using the arguments in Leon-Novelo et al. (2010). More details on the practicality of this choice appear in Section 6. Note that this choice, a priori, provides wide support $(-\infty, -4]$ and $[8, \infty)$ to the extreme events, i.e. 'normal' and 'missing', respectively, as a majority of PD datasets are expected to have some proportion of both healthy and missing sites (see Figure 2). Then, we rely on the flexibility of our nonparametric proposition to estimate the cell frequencies of the intermediate categories. Further sensitivity studies (see Section 4) reveal that derived parameter inference is not sensitive to this choice of fixed cutoffs.

3 Bayesian Inference

Incorporating pertinent background (prior) information, our Bayesian approach utilizes the Gibbs sampler and MCMC algorithms to draw samples from the posterior quantities of interest. In the next subsections, we present the joint probability model that incorporates prior choices and the data likelihood, and outline the relevant MCMC implementation schemes.

3.1 Joint probability model

Let $\boldsymbol{y} = \{y_i(s_j)\}, i = 1, \ldots, n; j = 1, \ldots, m$ denote the observed data and $\boldsymbol{y}^*$, the vector of latent variables as defined in Section 2. Let $\boldsymbol{w} = (w_h(s_j))$ be the matrix of PSB weights, and $\boldsymbol{\alpha} = (\alpha_1', \ldots, \alpha_\infty')'$. Finally let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_\infty)'$ be the vector of the mixture locations. Then, the joint probability model of $\boldsymbol{y}$ and the prior densities is given by:

$$
\begin{aligned}
&\text{pr}(\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{w}, \boldsymbol{\alpha}, \Sigma, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \\
&\prod_{i=1}^{n}\prod_{j=1}^{m} \mathbb{I}\{a_{y_i(s_j)} \leq y_i^*(s_j) \leq a_{y_i(s_j)+1}\} \times \\
&\prod_{i=1}^{n}\prod_{j=1}^{m} \phi\left(y_i^*(s_j) - x_i\boldsymbol{\beta} - z_j\boldsymbol{\gamma} - \theta_h\right) \times \\
&\prod_{j=1}^{m}\prod_{h=1}^{\infty} \mathbb{I}\left\{w_h(s_j) = \Phi(\alpha_{hj}) \prod_{l<h}(1 - \Phi(\alpha_{lj}))\right\} \times \\
&\prod_{h=1}^{\infty} \phi_m(\alpha_h; 0, \Sigma) \times \pi_\Sigma(\Sigma^{-1}) \times \prod_{h=1}^{\infty} \phi(\theta_h), \times \pi_\beta(\boldsymbol{\beta}) \times \pi_\gamma(\boldsymbol{\gamma})
\end{aligned}
$$

where $\mathbb{I}\{\cdot\}$ is an indicator function, $\phi_d(\cdot; \mu, \Lambda)$ denotes the probability density function of a d-variate Gaussian density with mean vector μ and covariance matrix Λ, with the subscript d suppressed when $d = 1$, π_β and π_γ are the prior distributions for the regression coefficients where each individual components follows non-informative zero-mean normal with variance 400, and π_Σ is the G-Wishart prior for the precision matrix Σ^{-1} defined in (5), with $S = (D - \rho W)$ corresponding to the proper CAR specification, $\rho = 0.9$, $\kappa = m$ and M_W is the set of symmetric p.d. matrices with zero off-diagonal entries for each 0 in the adjacency

matrix W. Construction of the spatial adjacency matrix W is illustrated via the 2-tooth periodontal grid presented as Web Figure 2 (see supplementary material). Each of the mid-buccal/lingual sites (red dots) have 2 Type-1 neighbors, while the mesio-/disto- buccal/lingual sites (black dots) have 3 Type I-III neighbors.

3.2 MCMC Implementation

In this section, we sketch the computational framework of our nonparametric ordinal model. All full conditional posterior distributions have tractable closed forms, allowing straightforward implementation of MCMC posterior simulation via Gibbs sampling. At the onset, we introduce latent cluster indicators $\xi_i(s_j)$ such that $\xi_i(s_j) = h$ iff $y_i(s_j)$ is sampled from the hth mixture component. The Gibbs sampler iterates through the following steps:

1. The first step involves data augmentation in the spirit of Canale and Dunson (2011) to simulate the latent variable $y_i^*(s_j)$ from the corresponding truncated normal distribution, conditional on the model parameters and the cluster indicators. Assuming the random effects $u_i(s_j)$ drawn from $N(\theta_h, 1)$, we have:
$$y_i^*(s_j) \sim \mathrm{N}(x_i\boldsymbol{\beta} + z_j\boldsymbol{\gamma} + \theta_h, 1)\mathbb{I}\{(a_{y_i(s_j)}, a_{y_i(s_j)+1}]\},$$
where $N(\mu, \sigma^2)\mathbb{I}_\Delta$ denotes the normal density with mean μ and variance σ^2, truncated to the set Δ.

2. Next, conditional on the latent continuous variables, the cluster indicators $\boldsymbol{\xi} = (\xi_i(s_j))'$, and the atoms $\boldsymbol{\theta}$, we sample the regression parameters. Their posterior distributions are again normal, namely
$$[\boldsymbol{\beta}|\boldsymbol{y}^*, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\xi}] \sim \mathrm{MVN}(\hat{\boldsymbol{\beta}}, V_\beta), \quad [\boldsymbol{\gamma}|\boldsymbol{y}^*, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\xi}] \sim \mathrm{MVN}(\hat{\boldsymbol{\gamma}}, V_\gamma),$$
where
$$V_\beta = (mX'X + 1/400I_p)^{-1}, \quad \hat{\boldsymbol{\beta}} = (\sum_{j=1}^{m} X'\boldsymbol{y}_j^*)V_\beta,$$
$$\boldsymbol{y}_j^* = \{y_i^*(s_j) - \theta_{\xi_i(s_j)} - z_j\boldsymbol{\gamma}\} \text{ for } i = 1, \ldots, n$$
and similarly
$$V_\gamma = (nZ'Z + 1/400I_q)^{-1}, \quad \hat{\boldsymbol{\gamma}} = (\sum_{i=1}^{n} Z'\boldsymbol{y}_i^*)V_\gamma$$
$$\boldsymbol{y}_i^* = \{y_i^*(s_j) - \theta_{\xi_i(s_j)} - x_i\boldsymbol{\beta}\} \text{ for } j = 1, \ldots, m.$$

3. The latent cluster indicators $\xi_i(s_j)$ are updated using a modification of the slice sampling approach of Kalli et al. (2011). The Kalli *et al.* approach falls under the class of 'conditional' methods (Walker, 2007) developed for Dirichlet processes mixtures that samples a sufficient but finite number of variables at each Markov chain iteration with the correct stationary distribution as an alternative to the more traditional 'marginal' approach (MacEachern and Müller, 1998) of integrating out the random distribution function. Here, for each i and j, we simulate the slice variables $v_{ij} \sim U(0, w_{\xi_i(s_j)})$, and then allocate the (latent) cluster indicators as arising from a multinomial sampling scheme:
$$\Pr(\xi_i(s_j) = h) \propto \mathbb{I}\{w_h(s_j) > v_{ij}\}N(x_i\boldsymbol{\beta} + z_j\boldsymbol{\gamma} + \theta_h, 1).$$
This approach avoids the truncation of the infinite sum available in the blocked Gibbs sampler of Ishwaran and James (2001).

4. Conditionally on the cluster memberships $\xi_i(s_j)$, we can update each cluster specific location parameter using $N\left(\frac{n_h\bar{y}_h^*}{1+n_h}, \frac{1}{1+n_h}\right)$, where n_h is the cluster size, i.e. the number of all sites across all subjects allocated to cluster h, and $\bar{y}_h^*$ the sample mean of $\{y_i^*(s_j) - \theta_{\xi_i(s_j)} - x_i\boldsymbol{\beta} - z_j\boldsymbol{\gamma}\}$ for cluster h.

5. The PSB weights $w_h(s_j)$ are updated by introducing a collection of conditionally independent latent variables $\eta_{ijh} \sim N(\alpha_{hj}, 1)$, $h = 1, \ldots, \infty$, for each i, j. Defining $\xi_i(s_j) = h$ if and only if $\eta_{ijh} > 0$ and $\eta_{ijl} < 0$ for $l < h$, we have

$$\begin{aligned} \text{pr}(\xi_i(s_j) = h) &= \text{pr}(\eta_{ijh} > 0, \eta_{ijl} < 0 \text{ for } l < h) \\ &= \Phi(\alpha_{hj}) \prod_{l<h} \{1 - \Phi(\alpha_{lj})\} = w_h(s_j) \end{aligned}$$

This data augmentation step leads to a convenient Gibbs sampling routine, where we can sample the augmented variables η_{ijh} conditionally on the latent variables and indicators via

$$\eta_{ijh} \,|\, \cdots \sim \begin{cases} N(\alpha_{hj}, 1)\mathbb{1}_{\mathbb{R}^-} \text{ if } h < \xi_i(s_j), \\ N(\alpha_{hj}, 1)\mathbb{1}_{\mathbb{R}^+} \text{ if } h = \xi_i(s_j). \end{cases}$$

6. Next, we update $\boldsymbol{\alpha}$, whose elements α_h have full conditional posterior distribution given by

$$\alpha_h = (\alpha_{h1}, \ldots, \alpha_{hm})' \sim \text{N}((\Sigma^{-1} + nI)^{-1}\bar{\eta_h}, (\Sigma^{-1} + nI)^{-1})$$

where $\bar{\eta_h} = (\eta_{1h}, \ldots, \eta_{mh})'$, $\quad \eta_{jh} = \frac{1}{n}\sum_{i=1}^{n} \eta_{ijh}$, for $h = 1, \ldots, h^* = \max\{h_1^*, \ldots, h_m^*\}$, where h_j^* is the minimum integer satisfying $\sum_{h=1}^{h_j^*} w_h(s_j) > 1 - \min\{v_{ij}\}$. Conditionally on the new values of $\boldsymbol{\alpha}$, one can compute the new PSB weights following their definition in equation (3).

7. The final step of the Gibbs sampler updates the spatial dependence parameter Σ. Given the G-Whisart hyperprior for Σ and normality assumptions for $\boldsymbol{\alpha}$, we attain conjugacy. Indeed, the full conditional posterior distribution for Σ^{-1} is again G-Wishart with mean $(D - \rho W) + \boldsymbol{\alpha}\boldsymbol{\alpha}'$ and degrees of freedom $\kappa + h^*$, constrained to have a zero entry for each zero in W. Although a number of proposals (Wang and Carvalho, 2009; Mitsakakis et al., 2011; Dobra et al., 2011) are available to sample from the G-Wishart, we follow the efficient block Gibbs sampler of Wang and Li (2012) which exploits a combination of the partial analytic structure of the G-Wishart and the double Metropolis-Hastings algorithm (Liang, 2010) to bypass matrix completion, within-graph proposal tuning, and evaluation of prior normalizing constants.

4 Application: GAAD data

The motivating GAAD data was collected at MUSC primarily to explore the relationship between PD and diabetes, as determined by the popular marker HbA1c, or glycosylated hemoglobin. Our dataset has 288 Type-2 diabetic subjects with complete covariate information and responses recorded for at least one tooth per subject. Severity of PD recorded at each tooth-site were classified into the categories, (i) No PD (CAL = 0 mm), (ii) Slight PD (CAL $\in \{1, 2\}$ mm), (iii) Moderate PD (CAL $\in \{3, 4\}$ mm), (iv) Severe PD (CAL ≥ 5 mm) and (v) Missing (if the tooth is missing, with all sites missing), following the American Association of Periodontology 1999 classification (Armitage, 1999). In addition, several subject- and site-level covariates were recorded. Subject level covariates include Age (in years), gender (1 = Female, 0 = Male), body mass index, or BMI (in kg/m^2), smoking status (1= a smoker either current or past, 0 = never) and HbA1c (1= high/uncontrolled with numerical level > 7, 0=controlled, if ≤ 7). We also included some site-level spatial covariates, such as site in gap (i.e., 1 if it is a mesio-/disto- buccal/lingual site, 0 otherwise, see Figure 2 in the Supplementary Material) and jaw indicator (1 = maxilla, or upper jaw, 0 = mandible, or lower jaw). Continuous covariates are standardized to have zero mean with variance 1. The mean age is 55 years with a range from 26 to 87 years. About 75% subjects in the data are females, 67% are obese (BMI ≥ 30), 31% are

smokers, and 59% with uncontrolled HbA1c. Including all the covariates mentioned above, we posit two competing models to fit our data that varies with the random effects specification. These are: (i) Model 1 (GW Model): Ordinal model with a standard multivariate normal structure for $u_i(s_j)$, and G-Wishart prior for the precision matrix; and (ii) Model 2 (PSB Model): Ordinal model with the PSBP structure described in Section 2 for $u_i(s_j)$. Note that we used the same (fixed) cut-points to fit and compare Models 1 and 2.

Although we were successful in deriving relevant Gibbs steps for the MCMC implementation of Model 2, the computational complexity was substantial leading to a longer convergence time of the respective MCMC chains for the parameters. Hence, we refrain from running multiple chains (in parallel), and used one single long chain. Computing codes were written in `R` with calls to `C` subroutines for the most demanding operations. The first 5,000 iterations were discarded as the burn-in, and posterior summaries were computed using 15,000 more iterations. Posterior convergence was assessed using trace plots, autocorrelation function (ACF) plots and Geweke's diagnostics (John, 1992) available in the `R` package `coda` (Plummer et al., 2006). The traceplots for the regression parameters under the PSB model are reported in Web Figures 3–6 of the Supplementary Material. For the Geweke measure, the hypothesis of the equality of the means for the first (10%) and last (50%) portions of the chain (after burn-in) was never rejected for any parameters. Specifically, the values of the Z statistics are -0.845, -1.259, -0.228, 0.180, 1.466, 0.239, and 1.706, with the corresponding p-values 0.398, 0.208, 0.820, 0.857, 0.143, 0.811, and 0.087 for the parameters corresponding to the fixed-effects Age, Female, BMI, Smoker, HbA1c, Maxilla, and GAP, respectively. For model selection, we use the posterior predictive L-measure for categorical data (Chen et al., 2004). Utilizing some user-defined loss-function, the L-measure quantifies model fit by comparing model-based posterior predictive distribution (ppd) to equivalent features from the observed data. Dichotomizing the ordinal responses $y_i(s_j)$ for subject i as $y_{i,k+1}(s_j) = 1$, if $y_i(s_j) = k$, and 0 otherwise, we define the binary vector $\tilde{y}_i(s_j) = (y_{i,1}(s_j), \ldots, y_{i,K}(s_j))'$, with only one component taking the value 1. Next, denote $\tilde{y}_i^{pr}(s_j)$ as the predicted binary vector obtained from the replicates $y_i^{pr}(s_j)$ that are draws from the ppd. Then, the L-measure using the squared-error loss is defined as:

$$\mathrm{L} = \sum_{i,j} \sigma_{ij}^2 + \nu \sum_{i,j} (\mu_{ij} - \tilde{y}_i(s_j))^2 \tag{6}$$

where $\mu_{ij} = E(y_i^{pr}(s_j)|\boldsymbol{y})$, $\sigma_{ij}^2 = \mathrm{Var}(y_i^{pr}(s_j)|\boldsymbol{y})$ and $0 < \nu < 1$. The first term on the right hand side of (6) is the penalty term where model overfitting might result into large predictive variance, whereas the second term (without the ν) represents goodness-of-fit. Choosing $\nu = 0.5$, the values of L for the two models are respectively 51,740 and 38,108. Clearly, our PSB model has better predictive performance than its parametric competitor.

Table 1 reports the posterior parameter estimates and their 95% credible intervals (CI). Parameters whose estimated 95% CIs do not include 0 are considered significant. The 95% CIs for Age excluded zero and the posterior estimates were positive for both models, revealing that periodontal health deteriorates with age. Males have a higher level of PD than the females. A positive association between BMI and PD is also revealed. Smoking is observed to be positively influencing PD, with a higher effect for the GW model. Uncontrolled HbA1c is a positive indicator of PD under the PSB specification, but the association lacks significance in the GW model. A tooth-site located in the maxilla is believed to have higher PD levels than a mandibular site. A site in a gap also exhibit higher PD levels than a non-gap site from the PSB model, whereas the association from the GW model is not significant. Subsequent sensitivity analysis by changing the chains initialization and using more non-informative variances for the priors on the regression parameters exerted minimal effect on the posterior estimates. Despite some obvious shifts, the signs of the mean estimates and the length of the corresponding 95% CIs not covering zero remained relatively unchanged.

In an attempt to provide meaningful interpretation to the posterior (point) estimates of covariate effects presented in Table 2, we present Figure 3 which plots the posterior predictive probabilities $\Pr(y_i(s_j)^{new} =$

Table 1: Posterior mean and 95% credible intervals for the regression coefficients under the parametric GWishart (GW) and the nonparametric probit stick-breaking (PSB) mixture models.

Covariate	GW			PSB		
	2.5%	Mean	97.5%	2.5%	Mean	97.5%
Age	0.126	0.136	0.146	0.095	0.110	0.126
Female	-0.122	-0.098	-0.074	-0.192	-0.158	-0.126
BMI	0.001	0.012	0.022	0.003	0.019	0.034
Smoker	0.073	0.095	0.117	0.025	0.057	0.090
Hba1c	-0.006	0.014	0.033	0.050	0.080	0.111
Maxilla	0.139	0.159	0.179	0.117	0.147	0.178
Gap	-0.045	-0.023	0.001	0.123	0.157	0.191

$k|-)$ of a maxillary (upper jaw) in-gap tooth-site $y_i(s_j)$ occupying the kth ordinal category from a random subject with (mean) age 55.27 years and (mean) BMI = 35.33, and under various combinations of gender, smoking, and HbA1c levels. For example, while the predictive probability of a healthy tooth-site from a random female smoker with high HbA1c is 0.073, it is 0.21 (much higher) for a random female smoker with low (controlled) HbA1c. As expected, these probabilities for the extreme 'Missing' category are respectively 0.33 vs 0.245 for the same comparison groups. In general, multiple previous studies have revealed strong evidence of diabetes as an important risk factor for PD, and the level of glycemic control (as determined by HbA1c) to be an important determinant in this relation (Mealey and Oates, 2006; Herring and Shah, 2006). This finding reconfirms the hypothesized link between diabetes and PD. Quite interestingly, this predictive probability flips [0.17 vs 0.09] while comparing female non-smokers in high vs low HbA1c groups, respectively. Note that smoking has long been considered as a major risk-factor for PD (Johnson and Hill, 2004). Another interesting observation is that within the high HbA1c group, female smokers have a lower probability of having healthy tooth-sites as compared to female non-smokers [0.073 vs 0.173]. Similar direction is observed between male smokers vs male non-smokers. Furthermore, within the low HbA1c group, both male and female smokers have a substantially higher probability of having missing tooth-sites (category 4) than their non-smoking counterparts [0.245 vs 0.033 for females and 0.104 vs 0.009 for the males]. However, counter-intuitively, the direction reverses for severe PD (category 3), i.e., both male and female non-smokers have a higher probability than the smokers. Hence, although the (overall) effect of smoking on the 'ordinal' PD responses was positive (see Table 1), it's effect on study sub-groups can be different. In addition, the probabilities corresponding to the extreme (missing) category are higher than the other categories across all 4 covariate combinations for the high HbA1c group, implying that subjects with high HbA1c usually have a higher proportion of missing teeth (missing due to previous occurrence of PD). Using the estimates, computing the predictive probabilities for other covariate combinations are straightforward. These probabilities can in turn be expressed as odds ratios, as desired.

Note that our Bayesian nonparametric propositions induce a natural stochastic clustering/grouping among the subjects and tooth-sites within/across subjects, such that subjects and tooth-sites with identical latent PD states (determined through the random structures) are grouped together. This clustering is different than the standard 'clustering by design' (i.e., tooth-sites in a mouth) induced in this dataset, and should be interpreted differently. Nonetheless, the direct interpretation of this clustering is often difficult due to label switching problems (Jasra et al., 2005). Among the possible solutions, here we utilize the approach of Medvedovic and Sivaganesan (2002). This approach exploits hierarchical clustering based on a dissimilarity matrix D obtained from the proportion of MCMC samples in which two sites were assigned to the same mixture component. Thus, for a given subject i, the (k,l)th element $d_{k,l}$ of D_i is computed as the ratio between the number of MCMC iterations where $\xi_i(s_k) \neq \xi_i(s_l)$ over the total number of MCMC iterations after burn in. Similarly, the dissimilarity matrix D_j between subjects for a given location s_j can be computed as the ratio between the number of MCMC iterations where $\xi_i(s_j) \neq \xi_l(s_j)$ over the total number of MCMC iterations

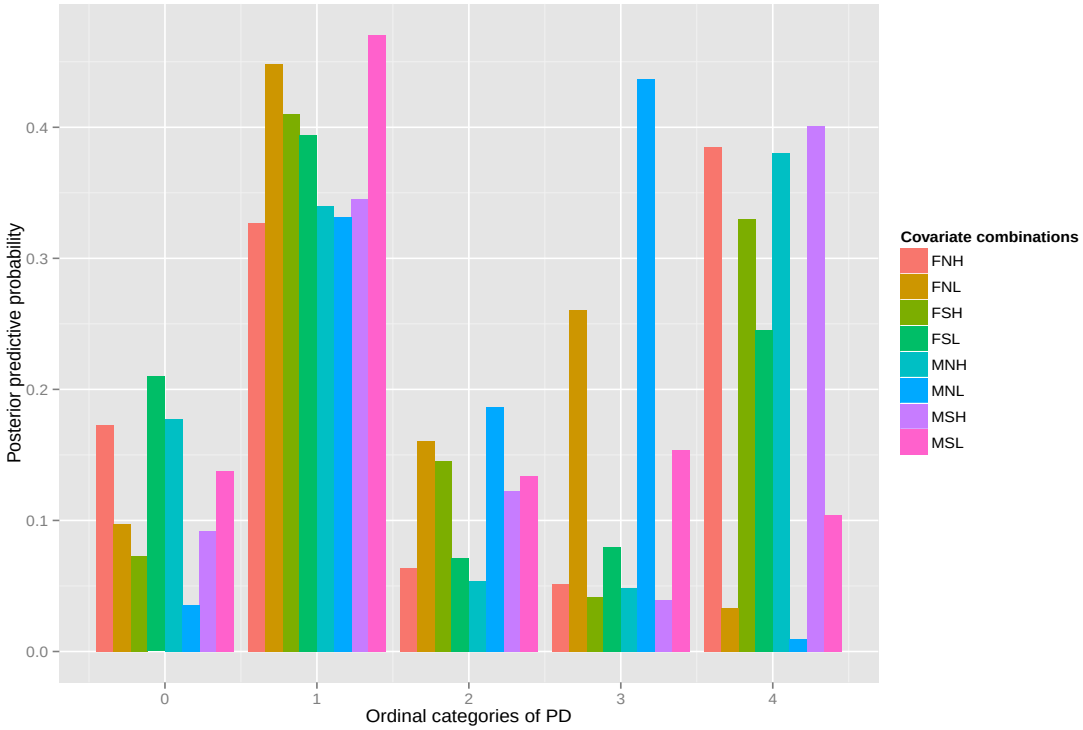
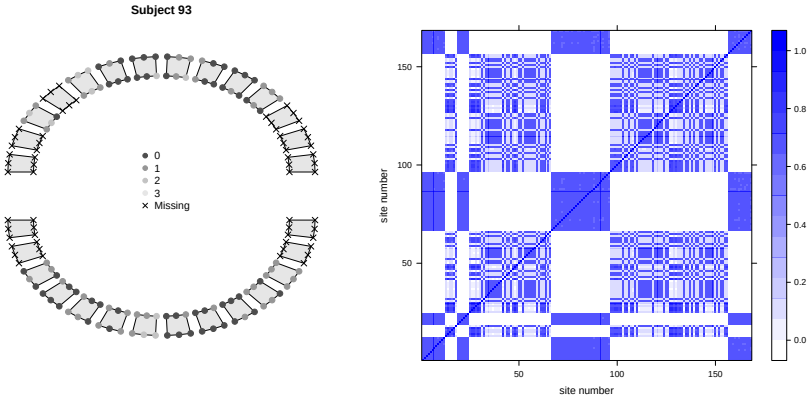


Figure 3: Posterior predictive probabilities $\Pr(y_i(s_j)^{new} = k|-)$ of a maxillary (upper jaw) in-gap tooth-site $y_i(s_j)$ from a random subject with (mean) age 55.27 years and (mean) BMI = 35.33, occupying the kth ordinal category, under various combinations of gender, smoking, and HbA1c levels. For these combinations, we have the following legends: F = Female, M = Male, N = Non-smoker, S = Smoker, H = high HbA1c, and L = low (controlled) HbA1c.

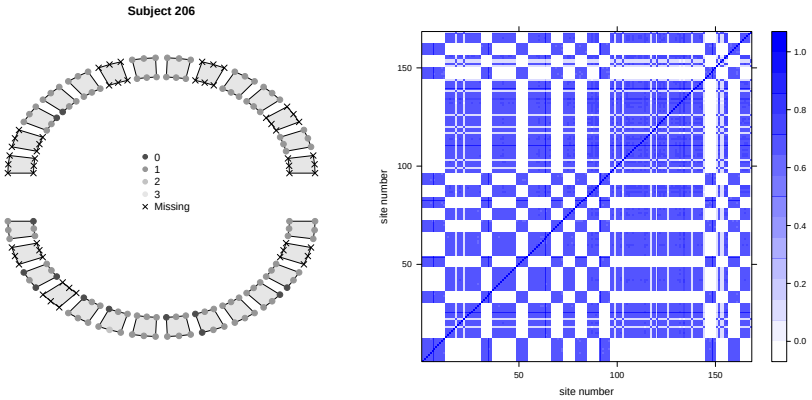
after burn in.

Figure 4 reports the heatmap of the posterior probabilities of site-clustering for two arbitrary subjects 93 and 206. This is calculated as 1 minus the elements of the matrix D_i, with the colors white and blue respectively representing no clustering and maximum clustering probabilities. Our flexible PSBP prior presents distinct clustering features among tooth-sites located in similar positions on the upper and lower jaw. Specifically, for subject # 93, the sites 1-12, 67-84, 85-96, etc corresponding to molars are missing. From the deep blue squares at the 4 corners and the center of the heatmap, it is evident that these (missing) molars are clustered. Furthermore, from the (rectangular) blue patch at the bottom of the heatmap centering the x-axis, sites 1-12 and 85-96 (corresponding to diagonally located molars attached to the left mandible and right maxilla) appear clustered. Subject # 206 produces a checkerboard, and reconstructs the missingness pattern by efficiently clustering missing teeth at various locations, for example, lateral incisors on upper jaw are clustered with missing molars on both jaws. Next, Panel (e) in Figure 4 presents averaged posterior probabilities of site-clustering across the subjects. Similar patterns as above are observed, with significant clustering among the anterior teeth sites (incisors and canines) in the upper and lower jaws. Clustering is also evident among molars located in similar positions on both jaws, and among maxillary molars on the left and right quadrants. Note, our spatial prior that utilizes the CAR adjacency structure (see Web Figure 2 in the supplementary material) does not consider mandibular and maxillary sites as neighbors. Regardless, our PSB proposition utilizing the G-Wishart precision could detect spatial patterns not known a priori. In the same vein, we also calculate the posterior probabilities of possible subject-level clustering utilizing the matrix D_j defined above. Figure 5 plots the heatmap of this subject dissimilarity matrix whose elements are the mean $d_{i,l}$'s across all locations, along with the dendogram obtained using the complete linkage. Note that the subject order in the row and column of this levelplot is not the original one. Using our PSB model, we computed the posterior cluster size over the MCMC replicates counting at each MCMC iteration the number of sites allocated to each cluster. It turns out that, a posteriori, 97% of the sites are allocated to the first 5 clusters. Hence, we cut the dendrogram in Figure 5 into five branches with cluster numbers $1, 2, \ldots, 5$ having cluster sizes $84, 18, 53, 102$, and 31, respectively, such that the elements (subjects) within each cluster have similar levels of PD.

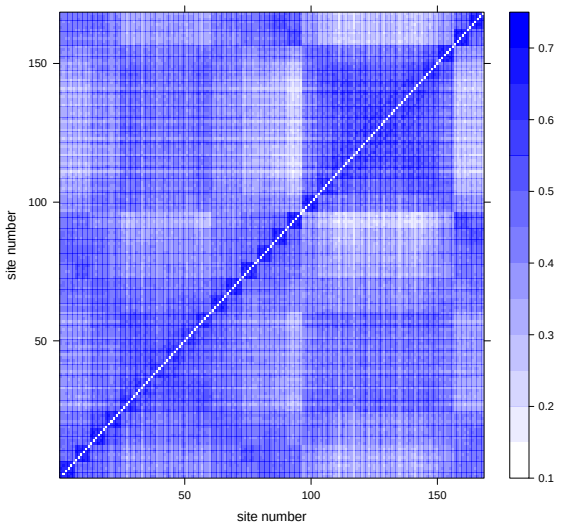
Figure 6 presents the empirical proportions of CAL categories and distributions of subject-level covariates (boxplots for the continuous and bar plots for the categorical covariates), across these 5 clusters. Note that, ideally, the differences among these clusters are statistically significant only with respect to the CAL categories, and may not be significant with respect to one or more covariates. This is not a problem, given that the clusters are a byproduct of the nonparametric inference, and are related to the response variable rather than to the covariates. In Panel (a), white represents the baseline 'Normal/No PD' category with the ordered PD categories represented by increasingly deeper shades of blue. For the bar plots in Panels (c), (e) and (f), blue represents the females, smoker and uncontrolled HbA1c categories, respectively, and white represents the compliments (i.e., males, non-smokers, and controlled HbA1c). These plots convey interesting patterns, coherent with the sign of the posterior estimates in Table 1. For example, cluster 5 subjects exhibit the highest proportion of the extreme category (missing sites) and the lowest proportion of the healthy category, compared to the other clusters. These subjects also represent the group with the highest (mean) age, proportion of males, and smokers, compared to the other clusters. On the contrary, subjects in cluster 2 subjects having the lowest proportion of missing sites are mostly females with the lowest mean age and proportion of smokers. Counter-intuitively, the boxplots in Panel (d) reveal that mean BMI values between clusters 2 and 5 do not differ substantially. This can be partially explained by the fact that BMI appeared borderline significant in Table 1. Furthermore, in tune to the established positive relationships of smoking (Pihlstrom et al., 2005) and age with PD, the proportion of smokers (from Panel e) and the mean age (from Panel b) roughly represent a similar ordering to the proportion with missing sites across the 5 clusters.



(a) (b)



(c) (d)



(e)

Figure 4: Tooth-site clustering. The upper panels (a) and (b) plots, respectively, the raw data and the mean posterior probability heatmap of clustering under our PSB model for subject # 93. The mid panels (c) and (d) plots the same, respectively, for subject # 206. The lower panel (e) is a heatmap plot of these posterior probabilities for the average mouth.

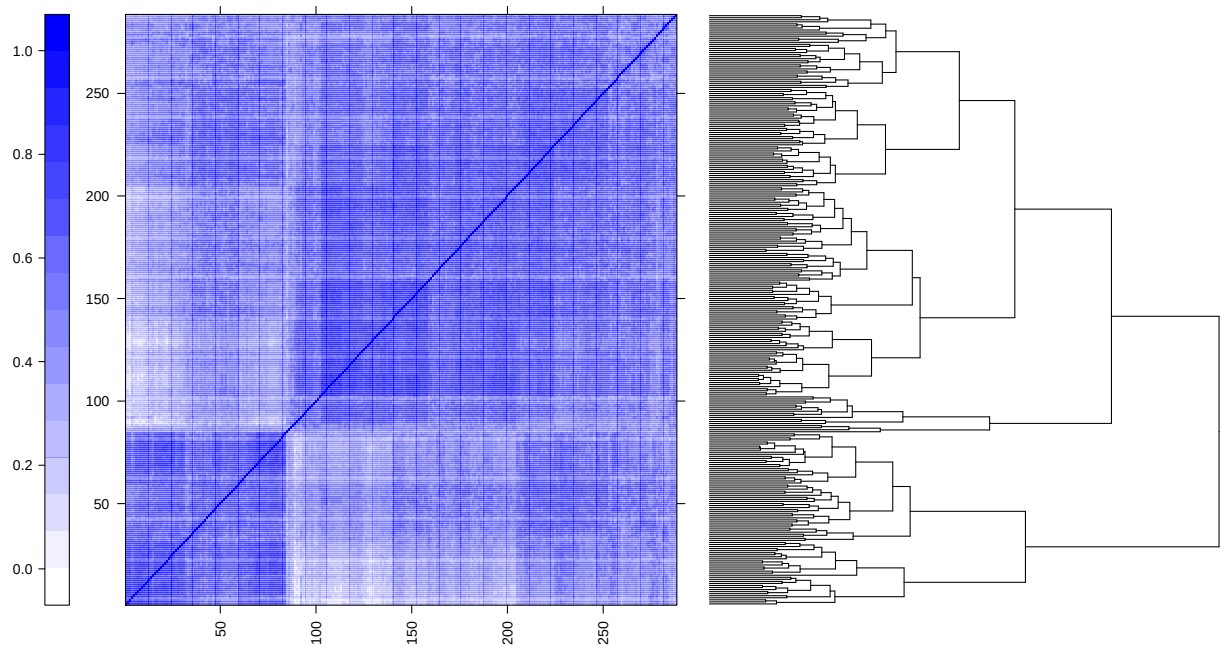


Figure 5: Clustering of subjects under the PSB specification. The colored plot is the mean-across-site dissimilarity matrix. Both the x-axis and y-axis plots the tooth-site numbers. Complete linkage dendogram is also included.

5 Simulation Studies

In this section, we conduct a small simulation study to investigate the finite sample properties of our non-parametric PSB model, and to understand the effect of spatial referencing and model misspecification on fixed-effects regression parameters. For computational reasons, we assume only 84 sites (representing half-mouth) for each subject and no site-level covariates. We consider sample sizes of $n = 50$ and $n = 100$. The datasets are generated as:

$$\begin{aligned} y_i(s_j) &= k \text{ iff } y_i^*(s_j) \in (a_k, a_{k+1}], \\ y_i^*(s_j) &= x_{i1}\beta_1 + x_{i2}\beta_2 + u_i(s_j), \end{aligned}$$

where the two subject-level covariates x_{i1} and x_{i2} are generated independently from the $\mathrm{N}(0, 1)$ density with parameters $\beta_1 = -3$ and $\beta_2 = 2$. The cutoffs are fixed at $a_0 = -\infty$, $a_1 = -4$, $a_2 = 0$, $a_3 = 4$, $a_4 = 8$, and $a_5 = \infty$, such that $K = 4$.

For the random effects vector $u_i = (u_i(s_1), \ldots, u_i(s_m))'$, we have three different specifications.

a. **Design 1**: $u_i \overset{ind}{\sim} \mathrm{MVN}(0, (D - 0.9W)^{-1})$, i.e., the random effects follow a parametric CAR density, with D and W appropriately chosen (as in Web Figure 2) to represent 84 sites,

b. **Design 2**: $u_i(s_j) \sim \sum_{h=1}^{\infty} w_h(s_j) N(\theta_h, 1), w_h(s_j) = \Phi(\alpha_{hj}) \prod_{l<h} (1 - \Phi(\alpha_{lj}))$, $\theta_h \overset{iid}{\sim} N(0, 1), \boldsymbol{\alpha}_h = (\alpha_{h1}, \ldots, \alpha_{hm})' \overset{ind}{\sim} \mathrm{MVN}(0, (D - 0.9W)^{-1})$, i.e., the random effects follow a PSB structure with a CAR proposal for α_h,

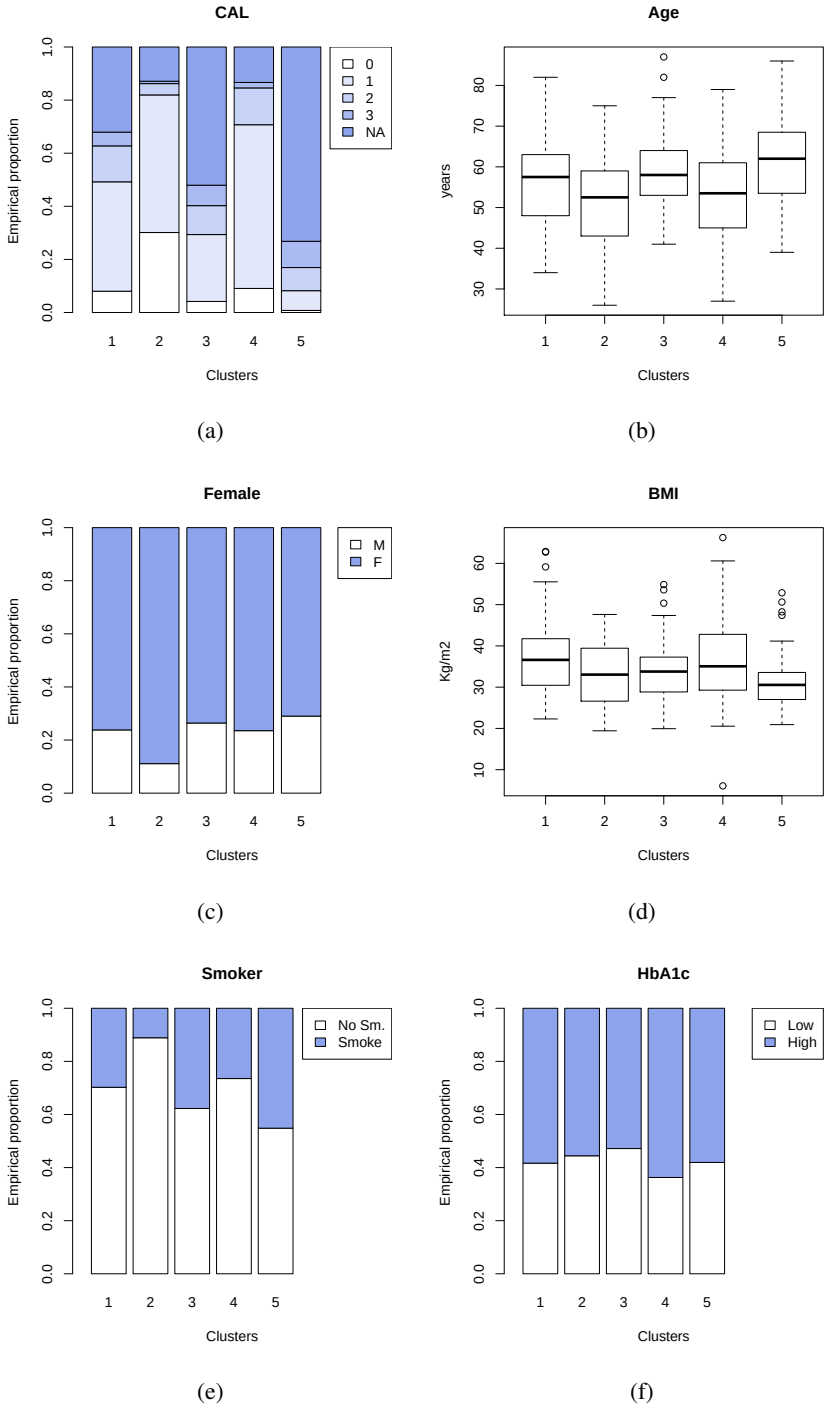


Figure 6: Empirical distributions of the response and the subject-level covariates across the 5 clusters. Panel (a) plots the empirical proportions of the ordinal CAL, where 'white' denotes the lowest category (no-PD), and the increasing darker shades represent the ordinal PD categories. The continuous covariates Age (Panel b) and BMI (Panel d) are represented by boxplots, whereas the categorical covariates are represented by barplots with the blue-shaded portions representing females (Panel c), smokers (Panel e) and uncontrolled HbA1c (Panel f).

Table 2: Simulation study results

Design	Sample Size (n)	Parameters	RB $\times$ 100		MSE $\times$ 100		95% CP	
			PSB	GW	PSB	GW	PSB	GW
1	50	β_1	0.412	-4.746	0.304	29.643	0.667	0.024
		β_2	0.414	-4.341	0.231	25.016	0.690	0.122
	100	β_1	0.260	-7.288	0.140	16.790	0.667	0.035
		β_2	0.403	-6.246	0.121	16.783	0.740	0.045
2	50	β_1	0.066	-4.679	0.152	30.966	0.889	0.080
		β_2	-0.279	-6.952	0.121	30.833	0.933	0.034
	100	β_1	0.091	-8.096	0.051	19.200	0.938	0.064
		β_2	-0.038	-7.674	0.047	18.293	0.942	0.049
3	50	β_1	-0.653	-5.235	0.105	36.714	0.917	0.057
		β_2	0.287	-6.458	0.096	33.172	0.976	0.034
	100	β_1	0.067	-8.469	0.051	19.669	0.926	0.068
		β_2	-0.147	-6.941	0.038	17.034	0.961	0.039

c. **Design 3**: Same PSB proposal as in Design 2, with $\alpha_h = (\alpha_{h1}, \ldots, \alpha_{hm})' \overset{ind}{\sim} \text{MVN}(0, \Sigma)$, where $\Sigma^{-1} \sim$ G-Wishart, i.e., a PSB structure with a G-Wishart proposal for α_h.

We simulate $B = 200$ datasets for each of these settings, and compare the parameter estimates derived from fitting the (a) PSB model and (b) the usual GWishart model using mean-square error (MSE), relative bias (RB) and coverage probability (CP). Defining $\hat{\beta}_j^{(b)}$ as the posterior mean of β_j from the bth simulated data set and β_j the true value, we calculate MSE $= \frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}_j^{(b)} - \beta_j)^2$, RB $= \frac{1}{B}\sum_{b=1}^{B}\frac{(\hat{\beta}_j^{(b)} - \beta_j)}{\beta_j}$ and CP $= \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}(\beta_j \in [\hat{\beta}_j^{(b,0.025)}, \hat{\beta}_j^{(b,0.975)}])$, where $\hat{\beta}_j^{(b,\delta)}$ is the estimated δ percentile from the b^{th} simulated dataset, and $\mathbb{1}$ is the indicator function.

The results from the simulation study are summarized in Table 2. On the overall, performances of β_1 and β_2 improved (RB and MSE reduced) for the higher sample size ($n = 100$) compared to $n = 50$. The absolute RB from the GW model far exceeds those from the PSB model for both parameters across all designs and sample sizes. Also, note that while the absolute RB reduced for both parameters in the PSB model for $n = 100$ as compared to $n = 50$, those for the GW model mostly increased, except for β_2 corresponding to Design 3. Similarly, the MSEs from the PSB model are substantially lower as compared to the GW model for both parameters across all sample sizes and designs. For the PSB model, the estimated CPs are close to the nominal (95%) level under Design 3, with a slightly better performance for $n = 100$ as compared to $n = 50$.

Note that both Designs 1 and 2 represent misspecified scenarios for our PSB framework; infact Design 1 (with the simple CAR structure) provides a higher degree of misspecification than Design 2. For the PSB, the CPs are slightly lower for Design 2 as compared to the nominal 0.95 level. Three of the four values are very close to 0.95, and the fourth isn't terrible as compared to the GW model. The CPs are however lower (between 0.67 and 0.74) under Design 1 for the PSB. As expected, the effect of misspecification on the GW model is considerable for both Designs. For example, under Design 1 and $n = 50$, MSE ($\times$ 100) for the GW model as high as 29.64 and the 95% coverage as low as 0.024 for β_1. The respective estimates under the PSB model are 0.304 and 0.667. Even under Design 3 whose data generation mimics our proposed model, the MSE ($\times$ 100) for β_1 and $n = 50$ under the GW model increases to 36.71 (corresponding estimate from the PSB model is 0.105), and the CP remains low at 0.057 (corresponding estimate from the PSB model is 0.917). From these results, we conclude that although the standard GW framework enjoys immense popularity due to the flexibility it provides, it is far less robust than our spatial PSB proposition in detecting

covariate effects under ordered data ground truth with spatial referencing.

6 Discussions

In this paper, we introduced a Bayesian nonparametric framework for spatially-associated clustered discrete ordinal data, motivated by applications to PD. Using PSB priors with a G-Wishart specification for the random terms, our computationally convenient approach uses standard MCMC tools and is flexible enough to capture interesting tooth-site level spatial clustering, which remained undetected under the standard GW specification. We adopted the Kottas et al. (2005) nonparametric approach for ordinal probit models with the ordinal cut-points pre-specified in advance to ease the computational burden, yet providing negligible effect on the posterior estimates. The model proposed is generalizable, and can be applied to a variety of other situations involving ordered discrete data in public health and medicine.

Note that the CAR structure is assumed for S, the symmetric positive definite scale matrix of the G-Wishart density. This is equivalent to assuming the CAR structure only as prior expectation for Σ^{-1}. Eventually, the G-Wishart confirmed its flexibility in our data application by selecting (clustering) sites that are not considered adjacent in the CAR proposition, such as clustering mandibular and maxillary (opposite jaw) sites. Certainly, in that sense, the CAR as a reference point (or, prior) was helpful. In this paper, our focus was to achieve flexibility over the standard CAR assumptions in the (replicated) areal data framework observed typically in PD studies. Indeed, there can be other densities/structures for S as worthwhile competitors to the CAR, however, the corresponding posterior sensitivity checks require further investigation.

As in Section 5, using the same set of fixed cut-offs for both the PSB and the GW models may not be theoretically justifiable, given that the GW model does not have a mixture distribution. We thank a referee for pointing this out. In practice, the cut-points in the GW model should be estimated from the data. However, in the data application and simulation studies, we wanted to provide some additional favor to Model 1 by fixing its cut-points to their true values for reasons well-established now.

Despite the lack of sensitivity to the choice of cut-offs for this dataset, some remarks on its generalizability to other datasets is of essense. Typically in parametric ordered probit models, for the sake of identifiability, the first cut-off is fixed, and the mean is free to vary (or vice versa). Next, the remaining cut-offs are estimated, since, conditionally on the mean, this is the only way to induce flexibility to the probability mass assigned to the other categories. On the contrary, our (nonparametric) mixture model provides enough flexibility by fixing the cut-offs, and letting the model decide on how much probability mass to assign to each category. To better understand this, consider our cut-offs choice $(-4, 0, 4, 8)$, and a simple finite mixture of $G = K$ Gaussian kernels, where $K = 5$ is the number of ordinal levels. We also assume the mid-points of the intervals (between the categories) as location parameters to assign the mixture components corresponding to the intervals. Now, with $\sigma = 1$, we have that each kernel can place more than 0.95 mass in each interval. Then, the values of π_k (the probability masses of each category) are such that $|\pi_k - w_k| < 0.05$, where w_k are the probability weights of each mixture component. Clearly, the distance between the cut-offs and the value of σ have an effect. Our choice restricts the π_k to be at most 0.95 for $k = 2, \ldots, K-1$, keeping the first and last cell probabilities unrestricted. However, if some other cell probability limits are desired a priori, the interval width can be increased or decreased accordingly.

Note that a variety of linear mixed model propositions exist in the literature (Bandyopadhyay et al., 2010; Reich and Bandyopadhyay, 2010; Reich et al., 2013) that quantifies the association of PD status with important clinical covariables. However, considering the inherent ordering exhibited by the CAL responses, we take an alternative modeling route here. Furthermore, the terminal 'Missing' category attributes the tooth to be missing mostly due to PD – the leading cause of tooth loss for this population with low socioeconomic status, presence of susceptible disease, and a unique genetic background (Fernandes et al., 2009; Reich and Bandyopadhyay, 2010). In reality, missingness can occur due to a variety of other reasons, such as past incidence of other types of dental diseases like caries, mechanical impact (such as collision, fall, etc),

or in accordance to patient recall. Thus, formulating the 'Missing (M)' component in our model as the extreme category representing PD may not be always generalizable (comments from a reviewer). Quite often, the dental hygienists rely on the subject's input during data collection, and this hinders an error-free determination of the true cause of missingness. However, in close conversation with periodontists who closely studied the Gullah PD dataset, we determined that the 'M' category for this dataset can be assumed to be *missing, only due to past PD incidence*, given that the proportions contributed from other causes of missingness is negligible. Hence, an ordinal model with missing as the extreme category seems feasible, avoiding unnecessary complications. Alternative approaches can model the missingness separately, such as a modification of the two-stage model of Li et al. (2011), with the missing tooth-surfaces modeled at the first stage, and the ordinal categories at the second stage, conditional on available tooth-surfaces. These two stages can be connected via a copula or some other structures with similar nonparametric formulation of the spatial referencing.

A second reviewer stated that there is a lot of enthusiasm currently to develop gradient-based Hamiltonian MC (HMC) techniques as an alternative to the standard Gibbs and Metropolis-Hastings based MCMC algorithms. However, implementing these HMC techniques are often preceded by unwieldy (gradient-based) calculations that varies with every problem. The learning curve can be steep for a reader not accustomed with these recent techniques. A recent paper (Orchard et al., 2013) explores the HMC techniques for the G-Wishart specification. However, our enthusiasm was dampened by the fact that adapting that scheme to our PSBP setup is not straightforward and computationally any smarter. Gibbs sampling, however old it is, is a widely accepted Bayesian tool, and our choice balances computational efficiency with implementation complexities compared to competing techniques available. It is not an unknown fact that chains take a longer time to converge in a typical Bayesian nonparametric problem. We were able to successfully derive relatively error-free and straightforward reader accessible Gibbs steps for each update, and someone conversant in standard Bayesian computing tools can efficiently replicate our setup to another problem. Both theoretical and practical exploration on the relative efficiency of our samplers with its competitors is certainly of essence, but beyond the current scope and focus. We plan to consider it elsewhere.

Our current proposition uses the probit link, and therefore somewhat limited in terms of the 'direct' interpretation afforded by using other link functions (such as the logit link with the corresponding odds ratios). However, the theoretical framework for adapting the fixed-cutpoints proposal with the logit link is non-trivial, and has not been considered elsewhere. Assessing link misspecification is beyond the scope of this paper, and hence we do not pursue it here. In addition, the 'ordering' assumption can be relaxed to consider a more general multinomial framework under non-Gaussian Markov random fields assumption via a Potts model (Green and Richardson, 2002). Although this removes the necessity of modeling latent space and deciding cut-offs, estimation in the Potts framework is often complicated due to the intractable normalizing constant (Møller et al., 2006), and require algorithms specifically tuned to the particular application. Furthermore, our current dataset is clustered at only one time-point. Clinical trials on PD to assess treatment effects often generate clustered-longitudinal data with site-level responses recorded at various time-points for a subject. Careful considerations are necessary (in terms of identifiability, etc) in extending our current setup to this framework by accommodating various sources of heterogeneity. All these remain viable avenues for future research.

Acknowledgments

The authors thank the Associate Editor and two reviewers whose insightful comments led to a substantially improved presentation of the manuscript. They also thank the Center for Oral Health Research at the Medical University of South Carolina for providing the motivating data, and Profs. Peter Müller, Sudipto Banerjee and Terrance D. Savitsky for interesting insights. Bandyopadhyay's research was partially supported by

grants R03DE023372 and R01DE024984 from the US National Institutes of Health.

References

Agresti, A. and Natarajan, R. (2001). Modeling clustered ordered categorical data: A survey. *International Statistical Review*, 69(3):345–371.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Armitage, G. C. (1999). Development of a Classification System for Periodontal Diseases and Conditions. *Annals of Periodontology*, 4(1):1–6.

Bandyopadhyay, D., Lachos, V. H., Abanto-Valle, C. A., and Ghosh, P. (2010). Linear mixed models for skew-normal/independent bivariate responses with an application to periodontal disease. *Statistics in Medicine*, 29(25):2643–2655.

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, Boca Raton, FL, second edition.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36:192–236.

Boehm, L., Reich, B. J., and Bandyopadhyay, D. (2013). Bridging Conditional and Marginal Inference for Spatially Referenced Binary Data. *Biometrics*, 69(2):545–554.

Brown, L. J., Oliver, R., and Loe, H. (1990). Evaluating periodontal status of us employed adults. *Journal of the American Dental Association*, 121(2):226–232.

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):1528–1539.

Carvalho, C. M., Massam, H., and West, M. (2007). Simulation of hyper-inverse wishart distributions in graphical models. *Biometrika*, 94(3):647–659.

Chen, M.-H., Dey, D. K., and Ibrahim, J. G. (2004). Bayesian criterion based model assessment for categorical data. *Biometrika*, 91(1):45–63.

Chung, Y. and Dunson, D. B. (2009). Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488):1646–1660.

Cowles, M. K. (1996). Accelerating monte carlo markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6(2):101–111.

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433.

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

Ferguson, T. (1974). Prior distribution on spaces of probability measures. *The Annals of Statistics*, 2:615–629.

Fernandes, J. K., Wiegand, R. E., Salinas, C. F., Grossi, S. G., Sanders, J. J., Lopes-Virella, M. F., and Slate, E. H. (2009). Periodontal disease status in gullah african americans with type 2 diabetes living in south carolina. *Journal of Periodontology*, 80(7):1062–1068.

Green, P. J. and Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1055–1070.

Herring, M. E. and Shah, S. K. (2006). Periodontal Disease and Control of Diabetes Mellitus. *The Journal of the American Osteopathic Association*, 106(7):416–421.

Hugoson, A., Thorstensson, H., Faltt, H., and Kuylenstierna, J. (1989). Periodontal conditions in insulin-dependent diabetics. *Journal of Clinical Periodontology*, 16(4):215–223.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Jasra, A., Holmes, C., and Stephens, D. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20:50–67.

John, G. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4*, pages 169–193. Oxford University Press, Oxford.

Johnson, G. K. and Hill, M. (2004). Cigarette smoking and the periodontal patient. *Journal of Periodontology*, 75(2):196–209.

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

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105.

Kottas, A., Mueller, P., and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics*, 14(3):610–625.

Laffont, C. M., Vandemeulebroecke, M., and Concordet, D. (2014). Multivariate Analysis of Longitudinal Ordinal Data With Mixed Effects Models, With Application to Clinical Outcomes in Osteoarthritis. *Journal of the American Statistical Association*, 109(507):955–966.

Leon-Novelo, L., Zhou, X., Bekele, B. N., and Müller, P. (2010). Assessing toxicities in a clinical trial: Bayesian inference for ordinal data nested within categories. *Biometrics*, 66(3):966–974.

Li, X., Bandyopadhyay, D., Lipsitz, S., and Sinha, D. (2011). Likelihood methods for binary responses of present components in a cluster. *Biometrics*, 67(2):629–635.

Liang, F. (2010). A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.

Mealey, B. L. and Oates, T. W. (2006). Diabetes Mellitus and Periodontal Diseases. *Journal of Periodontology*, 77(8):1289–1303.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206.

Mitsakakis, N., Massam, H., and Escobar, M. (2011). A Metropolis-Hastings Based Method for Sampling From G-Wishart Distribution in Gaussian Graphical Models. *Electronic Journal of Statistics*, 5:18–31.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Orchard, P., Agakov, F., and Storkey, A. (2013). Bayesian Inference in Sparse Gaussian Graphical Models. *arXiv preprint arXiv:1309.7311*.

Pihlstrom, B. L., Michalowicz, B. S., and Johnson, N. W. (2005). Periodontal diseases. *The Lancet*, 366(9499):1809–1820.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1):7–11.

Reich, B. and Bandyopadhyay, D. (2010). A latent factor model for spatial data with informative missingness. *Annals of Applied Statistics*, 4:439–459.

Reich, B. J., Bandyopadhyay, D., and Bondell, H. D. (2013). A nonparametric spatial model for periodontal data with nonrandom missingness. *Journal of the American Statistical Association*, 108(503):820–831.

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6(1):145–178.

Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, 29:391–411.

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

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54.

Wang, H. and Carvalho, C. M. (2009). Bayesian analysis of matrix normal graphical models. *Biometrika*, 96:821–834.

Wang, H. and Li, S. Z. (2012). Efficient Gaussian graphical model determination under G-Wishart prior distribution. *Electronic Journal of Statistics*, 6:168–198.