Bayesian nonparametric spatial modelling of ordinal periodontal data

Bayesian nonparametric spatial modelling of ordinal periodontal data

Modelli bayesiani nonparametrici spaziali per dati ordinali parodontali

Dipankar Bandyopadhyay and Antonio Canale

Abstract Clinical attachment loss (CAL) is a measure often used to assess periodontal disease (PD) status at a tooth site. While being ideally continuous, CAL measures are usually rounded and recorded as ordered categorical data. In addition, these CAL measures are hypothesized to be spatially-referenced. Traditional analysis model this integer-valued CAL via a linear mixed model with appropriate spatial random effects. In this paper, we propose a flexible nonparametric Bayesian approach to model the ordinal categories using an ordered probit model, simultaneously accounting for the within mouth spatial-referencing, yet preserving computational simplicity. An application to a real dataset on PD is presented.

Abstract *La perdita di attacco epiteliale (o clinical attachment loss, CAL) è una misura utilizzata per valutare lo stato di parodontite in un data locazione di un dente. Pur essendo idealmente continue, le misure di CAL sono solitamente arrotondate e registrate come dati ordinali. Si ipotizza inoltre una qualche referenzazione spaziale. Analisi classiche trattano questi valori interi attraverso modelli lineari a effetti misti con un'opportuna componente casuale che considera la dipendenza spaziale. In questo articolo, viene proposto un approccio bayesiano non parametrico per modellare le varibili ordinali attraverso un modello probit ordinato che tiene conto della referenzazione spaziale della bocca pur mantenendo semplicità di calcolo. È discussa inoltre un'applicazione a un dataset reale sulla parodontite.*

Key words: Clinical attachment loss; Conditionally-autoregressive (CAR); Probit stick-breaking; Nonparametric random effect;

Dipankar Bandyopadhyay
Division of Biostatistics, University of Minnesota, Minneapolis, MN
e-mail: dbandyop@umn.edu

Antonio Canale
Department of Economics and Statistics, University of Turin and Collegio Carlo Alberto, Italy
e-mail: antonio.canale@unito.it

1 Introduction

Periodontal disease (PD) is an inflammatory disease affecting periodontium, the tissues that both surround and support the teeth and maintain them in the maxillary and mandibular bones. Clinical dental research generates large amounts of data with potentially complex correlation structure from measurements recorded at several sites throughout the mouth. Clinical attachment loss (CAL) is one such measure popularly used to assess PD status. Interestingly, although these CAL measures are continuous (in *mm*), they are usually rounded and recorded as whole numbers, producing error-prone responses at the onset. In addition, it is hypothesized that these CAL values are spatially-referenced, i.e., a diseased site with a high CAL has more potential to influence the PD status of a set of neighboring sites as compared to other sites which are located distantly.

Traditional analysis consider modeling these error CAL measures via. linear mixed models with obvious modifications to handle spatial-referencing (Reich et al., 2007; Reich and Hodges, 2008; Reich and Bandyopadhyay, 2010; Reich et al., 2013). In this paper, we model the ordinal categories of these spatially-referenced CAL responses via a nonparametric Bayesian approach motivated by the probit stick-breaking process of Rodriguez and Dunson (2011). Preserving computational simplicity and flexibility, we assume that the components of the stick-breaking weights are generated from a Markovian conditionally auto-regressive (CAR) process with a specified neighborhood structure.

In the next section, we describe the model specification. Finally, Section 3 applies the model to the motivating PD dataset.

2 Model

Suppose $y_i(s_j)$ be the j-th site-level ordinal measure of CAL for i-th subject. One can categorize the CAL ordinal measures as $y_i(s_j) \in \{0,1,2,3,4\}$. Let us also assume $x_i(s_j)$, the associated covariates for subject i at the j-th location. Our main goal here is to quantify the effect of $x_i(s_j)$ on $P\{y_i(s_j) = k\}$, $k = 0,1,...,4$, accounting for the spatial associations between observations robustly. The standard ordinal model assumes that there is a multinomial selection process, where we observe $y_i(s_j)$ independently according to,

$$y_i(s_j) \overset{ind}{\sim} \text{Multinomial}\left(1, \left(\pi_{1ij}, ..., \pi_{5ij}\right)\right), \; i = 1,\ldots,n; \; j = 1,\ldots,m \qquad (1)$$

with $\sum_{k=1}^{5} \pi_{kij} = 1$. The independence assumption in (1) is questionable due to the clustering of the $y_i(s_j)$, and researchers often quantify association structure among these $y_i(s_j)$ based on some latent variables. In the most popular approach, the response variable is expressed in terms of a continuous latent variable $y_i^*(s_j)$ taking values on $(-\infty, \infty)$ as follows,

$$y_i(s_j) = k \text{ iff } y_i^*(s_j) \in (a_{k-1}, a_k],\ k = 1, \ldots, 5 \tag{2}$$

where, $-\infty = a_0 < a_1 < \cdots < a_5 = \infty$. From (2), the cell probability is $\pi_{kij} = P\{a_{k-1} < y_i^*(s_j) \leq a_k\}$. The cumulative probability $\lambda_{kij} = P\{y_i(s_j) \leq k\} = P\{y_i^*(s_j) \leq a_k\}$ is customarily modeled as,

$$\Phi^{-1}(\lambda_{kij}) = a_k - x_i(s_j)'\beta - u_i(s_j) \tag{3}$$

where $\Phi(\cdot)$ is standard normal cumulative distribution function.

Our main contribution here is to model the random effect distribution with a Bayesian nonparametric prior, namely

$$u_i(s_j) \sim G, \qquad G \sim \Pi \tag{4}$$

where G is a random probability measure with prior Π. To retain the data association structure offering a rich modeling perspective that alleviates the independence assumption between $y_i(s)$, we choose Π to be the probit stick-breaking prior of Rodriguez and Dunson (2011), specified as:

$$\begin{aligned} G &= \sum_{h=1}^{\infty} w_{hj}\delta_{\theta_h} \\ w_{hj} &= \Phi(\alpha_{hj}) \prod_{l<h} (1 - \Phi(\alpha_{lj})) \\ \alpha_h &= (\alpha_{h1}, \ldots, \alpha_{hm})' \overset{ind}{\sim} N(0, \Sigma). \end{aligned} \tag{5}$$

where the spatial dependence structure is modeled via the covariance matrix Σ. We consider an adjacency matrix W of the form $w_{jj'} = 1$ if s_j is a neighbor of $s_{j'}$ and $= 0$ otherwise. Uisng this notation, we let

$$\Sigma^{-1} = \frac{D - \rho W}{\tau^2} \tag{6}$$

where, D is a diagonal matrix with j-th diagonal entry signifying the number of neighbors of the location s_j. The latter construction give rise to the so called conditionally auto-regressive model (CAR)(Banerjee et al., 2004), where each conditional distribution of (5) is given as

$$\alpha(s_j) | \alpha(s_{j'}),\ j \neq j' \sim N\left(\frac{\rho}{m_j} \sum_{j \sim j'} \alpha(s_{j'}), \frac{1}{\tau^2 m_j}\right) \tag{7}$$

where, $j \sim j'$ specifies a neighborhood relationship between locations s_j and $s_{j'}$ and m_j denotes the number of neighbors of the location s_j. Clearly $\rho = 1$ in CAR prior imposes a singular Normal joint distribution and so one has to choose ρ appropriately. We additionally let $\rho \sim U(0,1)$ and $\tau^{-2} \sim \text{Ga}(a, b)$.

3 Application

We analyzed data obtained from a clinical study conducted at the Medical University of South Carolina to determine the PD status of type-2 diabetic Gullah-speaking African-Americans (Fernandes et al., 2009). The scientific interest is to determine the disease status of this population, accounting for patient level covariates, namely age (in years), gender (Male = 1, Female = 0), body mass index (BMI), smoking status (present or past smoker = 1, non-smoker = 0), glycemic status (determined by HbA1c). Additionally, we also included site-level covariates, such as the gap indicator (which is 1 if the site in in the gap, 0 = otherwise), and an indicator for the jaw (1 = maxilla, 0 = mandible).

In this study, CAL is measured at six pre-specified sites for each tooth, resulting in 168 measurements for each of the 288 subject in the study. Missing teeth (almost 20% in the sample) are denoted by an extreme category, since it has been shown that high level of CAL leads to destruction of the supporting bone around natural teeth, and eventually to tooth loss (see Reich and Bandyopadhyay, 2010, for a recent contribution).

The fixed effects parameters were given usual non-informative normal priors, ρ was given a uniform hyperprior between zero and one and τ was given an inverse gamma hyperprior. We implemented a Gibbs sampling scheme, and compute posterior summaries for the regression coefficients and the CAR covariance parameters. Our Gibbs sampler was run for 10,000 iterations, after 2,000 burn-ins.

Table 1 Regression coefficient posterior 95% credible intervals and posterior median for the standadized covariates.

	Lower	Median	Upper
Age	0.19	0.20	0.21
Female	-0.20	-0.17	-0.15
BMI	0.00	0.01	0.02
Smoker	0.14	0.16	0.18
Hba1c	0.03	0.05	0.07
Maxilla	0.22	0.24	0.26
Gap	-0.02	0.01	0.03

The results for the fixed effects are presented in Table 1, and are consistent with previous studies. From the table, we comment that increase in Age leads to worsening of the PD status. In addition, progression of PD is more prominent among males (compared to females), among smokers, and among subjects with uncontrolled gylcemic status. In addition, sites located in the upper jaw (maxilla) have higher degree of PD, as compared to sites in the lower jaw (mandible).

References

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, Boca Raton, FL.

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

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., D, B., and H, B. (2013). A nonparametric spatial model for periodontal data with non-random missingness. *Journal of the American Statistical Association*, 108:820–831.

Reich, B. J. and Hodges, J. S. (2008). Modeling longitudinal spatial periodontal data: A spatially adaptive model with tools for specifying priors and checking fit. *Biometrics*, 64(3):790–799.

Reich, B. J., Hodges, J. S., and Carlin, B. P. (2007). Spatial analyses of periodontal data using conditionally autoregressive priors having two classes of neighbor relations. *Journal of the American Statistical Association*, 102(477):44–55.

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6(1):145–177.