HHS Public Access

Author manuscript

Stat Methods Med Res. Author manuscript; available in PMC 2015 November 24.

Published in final edited form as:

Stat Methods Med Res. 2016 April ; 25(2): 644–658. doi:10.1177/0962280212463415.

A Bayesian semiparametric approach with change points for spatial ordinal data

Bo Cai[1], **Andrew B. Lawson**[2], **Suzanne McDermott**[3], and **C. Marjorie Aelion**[4]

[1]Department of Epidemiology and Biostatistics, University of South Carolina, USA

[2]Division of Biostatistics & Epidemiology, Medical University of South Carolina, USA

[3]Department of Family and Preventive Medicine, University of South Carolina, USA

[4]School of Public Health and Health Sciences, University of Massachusetts, Amherst, USA

Abstract

The change-point model has drawn much attention over the past few decades. It can accommodate the jump process, which allows for changes of the effects before and after the change point. Intellectual disability is a long-term disability that impacts performance in cognitive aspects of life and usually has its onset prior to birth. Among many potential causes, soil chemical exposures are associated with the risk of intellectual disability in children. Motivated by a study for soil metal effects on intellectual disability, we propose a Bayesian hierarchical spatial model with change points for spatial ordinal data to detect the unknown threshold effects. The spatial continuous latent variable underlying the spatial ordinal outcome is modeled by the multivariate Gaussian process, which captures spatial variation and is centered at the nonlinear mean. The mean function is modeled by using the penalized smoothing splines for some covariates with unknown change points and the linear regression for the others. Some identifiability constraints are used to define the latent variable. A simulation example is presented to evaluate the performance of the proposed approach with the competing models. A retrospective cohort study for intellectual disability in South Carolina is used as an illustration.

Keywords

Bayesian semiparametric model; change point; intellectual disability; soil metal exposure; regression splines

1 Introduction

Intellectual disability (ID) is a long-term disability that impacts performance in cognitive aspects of life and usually has its onset prior to birth. Many factors may cause ID, including genetic and chromosomal abnormalities, infections, chemical exposures, intentional and unintentional injuries. Chemical exposures such as arsenic (As), lead (Pb) and mercury (Hg) are developmental toxicants that have been associated with neurobehavioral dysfunctions

Corresponding author: Bo Cai, Department of Epidemiology and Biostatistics, University of South Carolina, USA. bcai@sc.edu.

and have been found to have adverse effects on intelligence in children, even at low levels of exposure.[1–3] Substantial evidence shows that the chemical metals cross the placenta and accumulate in fetal tissues.[4] The exposure route for the pregnant woman has been identified as oral and dermatologic.[5–7] Previous studies demonstrate elevated soil metal concentrations in urban areas from industrial and transportation sources and elevated rural soil concentrations of metals from natural geologic sources, pesticides, and industrial facilities.[8–10] A significant association is verified between soil and blood concentrations of Pb and As in children through hand to mouth contamination.[1,3,11]

Our research is motivated by a retrospective cohort study of pregnant women who were insured by South Carolina Medicaid from 1996 through 2002 and resided in one of ten residential study areas during pregnancy. In this study, we want to identify the soil chemicals as risk factors for unknown cause ID. ID is categorized in order based on intellectual quotient (IQ), containing three ordered levels: normal, mild and moderate/severe. Although some association between soil chemicals and ID has been unveiled,[12,13] it is of substantial interest to assess the elevated risk of ordinal ID associated with the detectable concentrations of soil chemicals with geographical information.

To model ordinal outcomes, cumulative logit models can be utilized.[14] Particularly the proportional odds model is often adopted, which assumes an identical effect of the predictors for each cumulative probability.[15] Most of the existing methods are primarily based on linking categorical data to latent continuous variables, which have an underlying normal regression structure.[16] Among the methods from the Bayesian perspective, Albert and Chib[17] proposed Bayesian methods for analysis of binary and polychotomous response data through the data augmentation with Gibbs sampling. The probit regression model for binary outcomes is seen to have an underlying normal regression structure on latent continuous variables. Values of the latent variables can be simulated from suitable truncated normal distributions. If the underlying continuous measurements are known, then the posterior distribution of the parameters can be computed using standard results for normal linear models. To accelerate Markov chain Monte Carlo (MCMC) convergence for the ordered probit model, one may use a multivariate Hastings-within-Gibbs update step to generate latent data and bin boundary parameters jointly, instead of individually from their respective full conditional posterior distributions.[18] However, these approaches cannot be extended straightforwardly to deal with spatially correlated ordinal data.

Compared to the models for spatially correlated continuous response data, approaches for spatially correlated ordinal data are less developed. The existing methods mainly focus on binary outcomes. Diggle et al.[19] modeled binary and count data by using generalized linear spatial models or generalized geostatistical models, where a spatial random effect term is included in the overall mean structure of a continuous latent variable relative to the categorical variable. To avoid nonidentifiability when estimating parameters of the underlying spatial correlation function, a unified Bayesian method[20] can be adopted for inference and prediction for binary spatial data. Higgs and Hoeting[21] proposed a parametric model for a point-referenced spatially correlated ordered categorical response. This approach relies on the approach by Albert and Chib[17] with incorporation of spatial

correlation. However, the approach is parametric which has less flexibility. In addition, the method has potential convergence problems[18] and does not consider the change-point issues.

The methods for change-point problems have drawn much attention over the past few decades. Most of the methods focused on time series related data where the change points are taken from discrete time points. Few change-point models were developed for spatial data. The most related work to our problem is the method by Majumdar et al.[22] who proposed a spatio-temporal change-point model for spatio-temporal continuous outcomes, allowing one time change point to reflect the overall mean change in both temporal and spatial associations. Since we focus on the association between the elevated risk of ordinal ID and potential multiple change points in concentrations of all soil contaminants, the approach by Majumdar et al.[22] cannot be directly applied.

To detect the unknown threshold concentrations of metals associated with the elevated risk of ordinal ID with spatial information, we propose a Bayesian semiparametric spatial approach with change points. The spatial continuous latent variable underlying the spatial ordinal outcome is modeled by the Gaussian process, which captures spatial variation and is centered at the nonlinear mean. The mean function is modeled by using the penalized smoothing splines for the soil metals with unknown change points and the linear regression for the demographic covariates. Some identifiability constraints are used to define the latent variable. The proposed approach provides a solution to the issues raised by the data.

The remainder of the article is organized as follows. Section 2 describes the latent variable model with change points. Prior specification, reparameterization, posterior implementation, and model comparison are also described. Section 3 evaluates the performance of the approach based on simulated examples. Section 4 illustrates the approach via an application to a retrospective cohort study of ID in South Carolina. Finally, Section 5 concludes with a summary and discussion.

2 The model

2.1 Latent variable model with change points

Let y_i ($i = 1, \ldots, n$) be an ordinal outcome with L categories, $y_i \in \{1, \ldots, L\}$, and $\mathbf{x}_i$ be the ith geographical location with $\mathbf{x}_i \in R^2$ (i.e. latitude and longitude). We can model the ordinal outcome as:

$$y_i \sim Categorical(\boldsymbol{\pi}_i)$$

where $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iL})$ denotes the vector of model probabilities that subject i is at level l, for $l = 1, \ldots, L$, and $\sum_{l=1}^{L} \pi_{il} = 1$. Since the categories of the outcome are ranked, the cumulative probability $p_{il} = Pr(y_i \leq l) = \sum_{r=1}^{l} \pi_{ir}$, for $l = 1, \ldots, L$. The cumulative probability p_{il} can then be modeled as

$$\text{logit}(p_{il}) = g_l(\mathbf{z}_i^*) + b_i, \quad l = 1, \ldots, L-1 \quad (1)$$

where $\mathbf{z}_i^* = \mathbf{z}^*(\mathbf{x}_i)$ denotes a vector of the covariates with linear or nonlinear effects at location $\mathbf{x}_i$, $g_l(\mathbf{z}_i^*)$ is a linear or nonlinear function of $\mathbf{z}_i^*$ at level l, and $b_i = b(\mathbf{x}_i)$ is the spatial random effect. However, modeling the ordinal outcome using logistic models could be complicated, especially when the linear predictor, η_{il}, is complex. Also, incorporating a realization from a spatial Gaussian process for each category is not straightforward. To allow for more convenient and efficient modeling on the spatial ordinal outcome, following Albert and Chib,[17] we use a latent continuous variable y_i^* distributed as N(η_i, 1) with $\eta_i = g(\mathbf{z}_i^*) + b_i$ and the variance being one for identifiability. The link between y_i and y_i^* is set as $y_i = 1$ if $y_i^* \in (\nu_{l-1}, \nu_l]$ with cut-points $\nu_0 < \nu_1 < \cdots < \nu_L$, where $\nu_0 = -\infty$ and $\nu_L = +\infty$, and $\nu_1 = 0$ for ensuring identifiability.

If function $g(\mathbf{z}_i^*)$ is a linear function, the model becomes a typical linear model. In general, the outcome can be linearly associated with some covariates while nonlinearly associated with the others. Thus, we further define $< g(\mathbf{z}_i^*) = \boldsymbol{\alpha}' \mathbf{u}_i + f(\mathbf{z}_i)$, where $\mathbf{u}_i = (u_{i1}, \ldots, u_{iq})'$ denotes the vector of covariates with linear effects, $\boldsymbol{\alpha}$ denotes the vector of coefficients for covariates $\mathbf{u}_i$, $\mathbf{z}_i = \mathbf{z}(\mathbf{x}_i) = (z_{i1}, \ldots, z_{ip})'$ denotes a vector of the covariates with potential nonlinear effects, and $f(\mathbf{z}_i)$ denotes a nonlinear function of $\mathbf{z}_i$. Among different nonlinear formulations, we consider a penalized smoothing spline function for $f(\mathbf{z}_i)$ due to its flexibility and efficiency. Then function $f(\mathbf{z}_i)$ can be written as

$$f(\mathbf{z}_i) = \sum_{j=1}^{p} \left(\boldsymbol{\beta}_j \mathbf{z}_{ij}^* + \boldsymbol{\gamma}_j (z_{ij} - \boldsymbol{\kappa}_j)_+^H \right)$$

where $\mathbf{z}_{ij}^* = (1, z_{ij}, z_{ij}^2, \ldots, z_{ij}^H)'$ with H being the order of splines, $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ denote the vectors of regression coefficients, $a_+ = \max(0, a)$ and $a_+^H = (a_+)^H$, and $\boldsymbol{\kappa}_j = (\kappa_{j1}, \ldots, \kappa_{jK})'$ denotes the vector of fixed knots. As mentioned in the last section, it is potentially the case that some spatial covariates (e.g. soil metals) may have extraordinary positive or negative effects on the outcome beyond a certain value. To allow for such change points, we model η_i as

$$\eta_i = \boldsymbol{\alpha}' \mathbf{u}_i + f(\mathbf{z}_i) + \sum_{j=1}^{p} \left(\boldsymbol{\beta}_j^* \mathbf{z}_{ij}^* + \boldsymbol{\gamma}_j^* (z_{ij} - \boldsymbol{\kappa}_j)_+^H \right) I(z_{ij} > d_j) + b_i \quad (2)$$

where $I(z > d)$ is an indicator which is 1 if $z > d$ and 0 otherwise, d_j is a change point for covariate j, and $\boldsymbol{\beta}_j^*$ and $\boldsymbol{\gamma}_j^*$ denote the vectors of regression coefficients. Equation (2) provides a general model to assess the effects of spatial covariates on ordinal outcomes with inclusion of potential change points. When $I(z > d) = 0$, there is no change point for a covariate in terms of its effect on the outcome. The model then reduces to the nonlinear regression model. When $H = 1$, the model reduces to the typical linear spline model with

change points. In this article, the nonlinear function is chosen as the penalized quadratic splines (i.e. $H = 2$).

Due to the spatial correlation, the realization of the spatial Gaussian process, $\mathbf{b} = (b_1, \ldots, b_n)'$, is assumed to be distributed as a multivariate Gaussian distribution, $\mathbf{b} \sim N_n(\mathbf{0}, \Sigma(\mathbf{x}, \boldsymbol{\theta}))$ where $\Sigma(\mathbf{x}, \boldsymbol{\theta})$ is the $n \times n$ spatial matrix with $\boldsymbol{\theta}$ being certain spatial parameters of a geostatistical model that may control smoothness (σ^2) and correlation decay (ρ) of the process. The spatial covariance matrix $\Sigma(\mathbf{x}, \boldsymbol{\theta})$ is taken to be $C(\|\mathbf{x}_i - \mathbf{x}_j\|)$ with $\|a\| = (a'a)^{1/2}$ and $C(r)$ being a member of the Matérn family of covariance functions. Generally, the Matérn covariance matrix function can be written as $\frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\|\mathbf{x}_i - \mathbf{x}_j\|\rho)^{\nu}\mathcal{K}_{\nu}(\|\mathbf{x}_i - \mathbf{x}_j\|\rho)$ with $\mathcal{K}_{\nu}$ being a modified Bessel function of order ν (e.g. Stein,[23] p.31). For a special case of $\nu = 3/2$, the Matérn covariance function becomes $C(r) = \sigma^2(1 + |r|\rho)\exp(-|r|\rho)$. We use this covariance function in our analysis because this expression is the simplest sub-family of the Matérn covariance functions that results in differentiable estimates.[24] Since the information contained in the categorical data pertains to the probability of being in a particular category, we cannot estimate both the variance of the (potentially hypothetical) latent continuous distribution and the thresholds for the categories.[20] Following Higgs and Hoeting,[21] the smoothing parameter σ^2 is set to be one. With the latent continuous variable $\mathbf{y}^*$, the complete-data likelihood can be expressed as

$$N_n(\mathbf{y}^*; \boldsymbol{\eta}, I_n)\prod_{i=1}^{n} 1\{\nu_{y_i-1} < y_i^* < \nu_{y_i}\} \quad (3)$$

where $\mathbf{y}^* = (y_1^*, , y_n^*)'$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$ and $1(x \in A)$ denotes an indicator function. This model is much simpler and easier to be implemented. One of the advantages of the latent variable model is that for three ordinal categories (i.e. $L = 3$) in the outcome of our data, there is only one unknown cut-point (i.e. ν_2).

2.2 Prior specification and reparameterization

To complete the Bayesian specification of our model, we need to choose prior distributions for the parameters. We assume a priori independence for these different parameter vectors. Following the conventional choice of priors for the regression coefficients, we let $\boldsymbol{\alpha} \sim N(\mathbf{0}, R_{\alpha})$, $\boldsymbol{\beta}_j \sim N(\mathbf{0}, R_{\beta})$, $\boldsymbol{\gamma}_j \sim N(\mathbf{0}, R_{\gamma})$, $\boldsymbol{\beta}_j^* \sim N(\mathbf{0}, R_{\beta^*})$, and $\boldsymbol{\gamma}_j^* \sim N(\mathbf{0}, R_{\gamma^*})$. Since the potential change points can be any value within the range of the corresponding covariate, we assume that potential change point follow a uniform prior with its range, $d_j \sim \text{Uniform}(\min(\mathbf{z}_j), \max(\mathbf{z}_j))$. The prior for the decay parameter ρ is chosen as $\mathcal{G}(a, b)$ to ensure that it is positive. According to French and Wand,[24] one may also fix ρ at $1/\max_{\leq i,j \leq n}\|\mathbf{x}_i - \mathbf{x}_j\|$. A $\mathcal{G}(a, b)$ random variable is parameterized to have expected value a/b and variance a/b^2.

The use of standard Gibbs sampler in the model with ordered categorical response variables suffers from the slow convergence problem. Nandram and Chen[25] developed an improved algorithm by using Dirichlet proposal for re-scaled cut-point parameters. Weadopt their reparameterization approach for our model with ordered categorical responses. Let $\phi = \frac{1}{\nu_{L-1}}$,

we define $\tilde{\nu}_l = \varphi\nu_l$, for $l = 0, \ldots, L$. φ is a re-scaling parameter which transfers the cut-points, $0 = \nu_1 < \cdots < \nu_{L-1}$, to the values between 0 and 1. This re-scaling step reduces the correlations between the cut-points and the latent variables and thus accelerates the convergence of the algorithm. We then define the following transformations as

$$\tilde{y}_i^* = \phi y_i^*,\ \tilde{\boldsymbol{\alpha}} = \phi\boldsymbol{\alpha},\ \tilde{\boldsymbol{\beta}}_j = \phi\boldsymbol{\beta}_j,\ \tilde{\boldsymbol{\gamma}}_j = \phi\boldsymbol{\gamma}_j,\ \tilde{\boldsymbol{\beta}}_j^* = \phi\boldsymbol{\beta}_j^*,\ \tilde{\boldsymbol{\gamma}}_j^* = \phi\boldsymbol{\gamma}_j^*,\ \tilde{b}_i = \phi b_i$$

With this reparameterization, the realization $\tilde{\mathbf{b}} \sim \mathrm{N}_n(\mathbf{0}, \tilde{\Sigma}(\mathbf{x}, \boldsymbol{\theta}))$ where $\tilde{\Sigma}(\mathbf{x}, \boldsymbol{\theta}) = C(\|\mathbf{x}_i - \mathbf{x}_j\|) = \phi^2(1 + |t|\rho)\exp(-|t|\rho)$. By integrating out $\tilde{\mathbf{b}}$, the likelihood becomes

$$\mathrm{N}_n(\tilde{\mathbf{y}}^*; \tilde{\boldsymbol{\eta}}_{-\tilde{\mathbf{b}}}, \tilde{\sum}(\mathbf{x}, \boldsymbol{\theta}) + \phi^2 I_n) \prod_{i=1}^{n} 1\{\tilde{\nu}_{y_i - 1} < \tilde{y}_i^* < \tilde{\nu}_{y_i}\}, \quad (4)$$

where $\tilde{\boldsymbol{\eta}}_{-\tilde{\mathbf{b}}}$ is $\tilde{\boldsymbol{\eta}}$ without $\tilde{\mathbf{b}}$. For simplicity, we suppress the subscript $-\tilde{\mathbf{b}}$ for the remainder of the article. The prior for φ^2 is chosen as $\mathcal{IG}(c, d)$. It is clear that when $L = 3$, it becomes $\tilde{\nu}_0 = -\infty < \tilde{\nu}_1 = 0 < \tilde{\nu}_2 = 1 < \tilde{\nu}_3 = \infty$ and there is no unknown threshold scale needed to be updated.

2.3 Posterior computation

The full conditional posterior distributions for all the parameters can be derived based on the likelihood, the priors and the reparameterization specified in Section 2.2. Our posterior computation relies on the Gibbs sampler and Metropolis–Hastings algorithms. After specifying initial values for the parameters and latent variables, the proposed MCMC algorithm proceeds by updating the unknown parameters sequentially as shown in the Appendix.

Samples from the joint posterior distribution of the parameters and the latent variables are generated by repeating the steps for a large number of iterations after apparent convergence. It is noticed that the full conditional posterior distributions for all the parameters except ρ are conjugate, which provides an efficient sampling scheme. For ρ, we adopt the Metropolis–Hastings algorithm. In this article, multiple chains with different initial settings are carried out, which are used for the final summary. The implementation is carried out by using R.[26] The code is available from the first author upon request.

2.4 Model comparison

We consider the deviance information criterion (DIC),[27] a widely used criterion for model selection in hierarchical Bayesian models. Similar to the Bayesian information criterion (BIC) and Akaike's information criterion (AIC), the DIC combines a measure of fit and complexity. The measure of fit is given by the posterior mean of the deviance and the measure of complexity by the difference between the posterior mean of the deviance and the deviance based on the posterior means of the parameters. The Bayesian deviance, $D(\boldsymbol{\Theta})$, is defined as

$$D(\boldsymbol{\Theta})=-2logL(\boldsymbol{\Theta}|\mathbf{y})+2\log(Q(\mathbf{y}))$$

where $\boldsymbol{\Theta}$ denotes the set of all parameters in the model of interest, $L(\boldsymbol{\Theta}|\mathbf{y})$ denotes the likelihood for the observed data $\mathbf{y}$, and $Q(\mathbf{y})$ denotes some specified standardizing term that is a function of the data alone and hence does not affect model comparison. The DIC is defined as $\overline{D(\boldsymbol{\Theta})}+p_D$, where $\overline{D(\boldsymbol{\Theta})}=E_{\boldsymbol{\Theta}|\mathbf{y}}(D(\boldsymbol{\Theta}))$ denotes the posterior expectation of the deviance, and p_D denotes the effective number of parameters which is defined by difference between the expected deviance and the deviance evaluated at the posterior expectation, $\overline{D(\boldsymbol{\Theta})}-D(\overline{\boldsymbol{\Theta}})$. The DIC can be easily calculated by taking the difference of the sample mean of the simulated values of D and the deviance based on the sample means of the simulated values of $\boldsymbol{\Theta}$ obtained via MCMC.

3 A simulated example

We evaluated the performance of the proposed approach based on a simulation with 100 replications. Without loss generality, we generated 500 observations with categorical outcomes and four covariates with nonlinear effects. The response variable y_i was categorized into three categories described in Section 2.2 with $\eta_i=\sum_{k=1}^{4} f_k(z_{ik})+b_i$, where z_{i1} ~ Uniform(0, 1), z_{i2} ~ Uniform(0, 4), z_{i3} ~ Uniform(0, 10), z_{i4} ~ Uniform(0, 7), and the associations between the covariates and the linear predictor follow different functions with change points,

$$\begin{gathered} f_1(z)=\sin(\pi z)+0.8I(z \geq 0.4), \\ f_2(z)=0.1\exp(z)+5I(z \geq 2), \\ f_3(z)=0.1(0.2(z/10)^{11}(10-z)^6+10z^3(1-z/10)^{10})+2I(z \geq 5), \\ f_4(z)=0.2+6I(z \geq 4). \end{gathered}$$

The spatial locations, $\mathbf{x}$, were randomly generated from Uniform(0, 5). The realization of the spatial Gaussian process $\mathbf{b}$ was generated from a multivariate Gaussian distribution, $N_n(\mathbf{0}, \Sigma(\mathbf{x}, \theta))$, where Matérn covariance function in $\Sigma(\mathbf{x}, \theta)$ is described in Section 2.1 with $\rho = 1$. With generated $\mathbf{b}$, we generated samples for the continuous variable $\mathbf{y}^*$ ~ $N_n(\zeta, I_n)$. The samples for the ordinal categorical response variable were then generated as $y_i = 1$ if $y_i^*<0$, $y_i = 2$ if $0<=y_i^*<1$ and $y_i = 3$ if $y_i^*>=1$. Figure 1 depicts the plots of covariates vs. the response variable in a simulated data set.

We fitted the proposed model with change points to the simulated data. We chose a flat normal prior N(0, 100) for the coefficients. The gamma prior $\mathcal{G}(2, 1)$ was chosen for ρ. Since the response variable has three ordered categories, φ is fixed as one. For each replication, we ran the Gibbs sampling algorithm described in Section 2.3 for 20,000 iterations after a burn-in of 5000 iterations. The diagnostic tests[28,29] were carried out which showed good convergence and efficient mixing.

For comparison, we also fitted the generalized additive model (GAM) and the proposed model without change points to the simulated data. The GAM model used here was proposed by Wood,[30] which relies on approximations to the thin plate splines. The smoothers provide optimal low rank approximations to generalized smoothing spline models that are both computationally efficient and stable. The GAM approach is available in the R package mgcv, available from www.cran.r-project.org. In the package, the smooth class 'tp' was used, indicating optimal low rank approximation to thin plate spline.

Figure 2 shows the averaged fitted nonlinear curves for the four covariates based on the proposed method with change points, the method without change points and the GAM model. The shaded band indicates the 95% pointwise credible intervals from the simulated data sets. Since the estimate of the nonlinear curves based on the proposed model without change points and the GAM could not fully reflect the true curves, the results may potentially be misleading. In contrast, the proposed semiparametric model with change points provides the best fit for the designed covariates functions. In addition, the estimate of ρ from the proposed model is 0.92 with 95% credible interval (0.58, 1.34). The estimates and 95% credible intervals of the change points for the four covariates are 0.37(0.31,0.43), 1.94 (1.46,2.40), 4.86 (4.20, 5.44) and 4.07 (3.65, 4.82), respectively. The estimated change points for the covariates are close to the true values.

To compare the goodness of fit of the proposed model with change points to that of the model without change points, we also calculated the deviance information criterion (DIC)[27] (there is no DIC available for the frequentist GAM model). The DICs for the model with change points and the model without change points are 130.7 and 144.6, respectively, implying that the proposed model with change points is better than the other model. For each replication of the simulation, we also assessed the sensitivity of the results to the prior specification by repeating the analysis with the different hyperparameters. We noticed that the estimates of the parameters varied slightly with different specifications of hyperparameters.

4 Application

We applied the proposed method to the ID data introduced in Section 1. The women in this study were followed through pregnancy and delivery. Then they were longitudinally followed to see if their child received a diagnosis of ID. The study areas were dispersed throughout South Carolina, in rural and urban areas and each area contained a low and high prevalence area for the outcome of ID. In order to maintain the confidentiality agreement, the soil samples were collected according to grids throughout the residential study areas. Nine chemicals were measured in soil samples collected, arsenic (As), barium (Ba), beryllium (Be), chromium (Cr), copper (Cu), mercury (Hg), manganese (Mn), nickel (Ni) and lead (Pb). Other available mother and baby covariates include mother's age, mother's ethnicity (white, black and other), mother's alcohol consumption during pregnancy (yes/no), the number of prior births (parity)(0, 1, 2 and 3 or more), birth weight, child sex and weeks of gestation. Clusters of ID for each gestational month of pregnancy were identified based on maternal addresses. Ten distant geographic areas of land that included a gradient of risk

of ID were then selected. Based on a unified grid system, soil samples were collected and analyzed for nine metals and a general toxicity indicator within the areas.

In our analysis, we included 9440 individual samples with 7 demographic covariates related to mothers and children's information. In each area, grided soil samples were collected with 8 chemical concentrations being measured. Figure 3 shows the ID outcome locations in Area 5. In the application, the mother's age range is from 12 to 44. Previous studies show that mothers over 35 would have higher risks of having ID babies than younger mothers. However, the percentage of mothers' age over 35 in our data is only 2.83%. When we categorized the age into three categories, [12, 20), [20, 24) and [24, 44], the proportions of having ID for these three categories are 30%, 35% and 35% respectively, indicating that they are in slightly increasing trend. It is known that there is a dramatically increasing risk of the ID when mothers' age is over 35 (e.g. Newberger[31]). However, the percentage of mothers' age over 35 in our data is only 2.83%. Thus, it seems plausible to assume the linear effect of mother's age. Preliminary analysis based on the logistic regression shows that there is no significant interactions between the demographic covariates.

We chose flat priors for the coefficients similar to those in the simulation. The prior for ρ was chosen as a gamma distribution $\mathcal{G}(3, 1)$ to allow for a reasonable range. We ran the Gibbs sampling algorithm detailed in the appendix for 40,000 iterations after a burn-in of 10,000 iterations. Convergence was deemed adequate based on the diagnostic tests used in the simulation study.

Table 1 shows the summary of regression coefficients for the demographic covariates based on the three models. The results from the three models in Table 1 show that the older mother, the black mother, male baby and lower gestational age would significantly increase the risk of children having ID. Although the three models have similar results of effects of demographic covariates, it is shown that the proposed model with change points provides narrower credible intervals.

Table 2 shows the posterior means and 95% credible intervals of the change points for the soil metals. It is observed that the 95% credible intervals for Ba, Cr, Cu, Mn, Ni and Pb are quite wide which cover most of their ranges. This implies that there is no obvious evidence of change points for the dramatic changes of ID for the six metals based on the data. We actually produced the histograms of the change-point samples for the six metals, which show fairly flat shapes. In contrast, As and Hg seem to have change points for elevated risk of ID due to much narrower 95% credible intervals compared to the other soil chemicals. Besides showing the significant impact of the two metals on children's ID, which is consistent with that in the previous studies,[12,13] the proposed approach allows for detecting a sudden change in the effects of the metals on the risk of ID.

Figure 4 depicts the estimated curves and 95% pointwise intervals for As and Hg based on the proposed model with change points. The lower panel shows the fitted curves and 95% credible intervals for As and Hg across the entire observed ranges, where it is unclear if there exist the change points that accelerate the risk of ID. However, in the upper panel where the zoomed estimated curves are displayed, it is obvious that the change points exist

which have accelerated risk effects on ID occurrence in short ranges. This implies that when the two soil metals reach over their thresholds, there would be a sudden increase of risk of ID. This is an interesting finding as the detected thresholds of As and Hg in the soil suggest that some interventions might be needed to reduce the accelerated risk of ID in those areas. Since no previous studies focused on the change point detection for association between ID and the soil chemicals, more investigations are needed to unveil the scenario in the future work.

Similarly, we calculated the DICs for the model with change points and the model without change points, which are 4072.21 and 4098.58, respectively. This indicates that the model with change points has a better fit than the model without change points.

5 Discussion

In this article, we propose a Bayesian semiparametric approach with change points for spatial ordinal data, which allows to detect the unknown threshold points. By using the latent variable underlying the spatial ordinal outcome and the penalized smoothing splines for the covariates, we are able to efficiently model the ordinal outcome nonlinearly associated with the predictors. The simulation study and the application have shown the improvement of modeling and estimation from the proposed approach compared to the other two models without change points.

Although the location of the change point can be potentially found by using diagnostic methods, they often fail to provide statistical inferences about structural change. This is because searching for change points with diagnostics risks mistaking random variation for structural change. In contrast, the proposed Bayesian analysis provides statistical inferences for both regression coefficients and the location of the change point. The Bayesian change point model is shown to be feasibly estimated by stochastically sampling from the conditional posteriors for the regression coefficients. The simulation study shows that when a sudden change occurs, the proposed approach outperforms the methods without change points. However, when there is no dramatic change (e.g. for several soil chemicals in the application), the proposed method provides an estimate for the change point with a wide credible interval, implying that no important threshold effects can be effectively detected by the proposed method under this scenario.

The proposed approach can be implemented in WinBUGS[32] with exponential covariance function. Conceptually WinBUGS can carry out the proposed approach with Matérn covariance matrix. However, we experienced slowness in computation time even for the model with very few latent covariates. It is noticed that there might be a potential estimation bias due to the measurement error. The occurrence of the measure error may lead to potential misleading conclusion. On-going work involves incorporating Berkson measurement error model for the latent spatial covariates[33] to reduce potential biases. In addition, although a relatively small sample size (300) was used in the simulation study which shows reasonably good performance of the proposed approach, a systematic study is needed to examine the sample size for detecting important threshold effects.

Acknowledgments

The authors thank the editor and the two referees for valuable comments, which greatly improved the presentation.

Funding

This work was supported by NIH/NIEHS grant 2R01ES012895-04A2.

References

1. Goldman LR, Koduro S. Chemicals in the environment and developmental toxicity to children: a public health and policy perspective. Environ Health Perspect. 2000; 108(Suppl 3):443–448. [PubMed: 10852843]
2. Sullivan, JB.; Krieger, GR. Clinical environmental health and toxic exposures. Philadelphia: Lippincott Williams & Wilkins; 2001.
3. Bellinger, DC. Human developmental neurotoxicology. New York: Taylor & Francis; 2006.
4. Miodovnik A. Environmental neurotoxicants and developing brain. Mount Sinai J Med. 2011; 78(1): 58–77.
5. Baghurst PA, McMichael AJ, Wigg NR, et al. Environmental exposure to lead and children's intelligence at the age of seven years. The Port Pirie Cohort Study. N Eng J Med. 1992; 327:1279–1284.
6. Wasserman G, Graziano JH, Factor-Litvak P, et al. Consequences of lead exposure and iron supplementation on childhood development at age 4 years. Neurotoxicol Teratol. 1994; 16(3):233–240. [PubMed: 7523846]
7. Davidson PW, Myers GW, Weiss B, et al. Prenatal methyl mercury exposure from fish consumption and child development: a review of evidence and perspectives from the Seychelles Child Development Study. Neurotoxicology. 2006; 27:1106–1109. [PubMed: 16687174]
8. Li X, Lee S, Wong S, et al. The study of metal contamination in urban soils of Hong Kong using a GIS-based approach. Environ Pollut. 2004; 129:113–124. [PubMed: 14749075]
9. Aelion CM, Davis HT, McDermott S, et al. Metal concentrations in rural topsoil in South Carolina: potential for human health impact. Sci Total Environ. 2008; 402:149–156. [PubMed: 18538375]
10. Davis HT, Aelion CM, McDermott S, et al. Identifying natural and anthropogenic sources of metals in urban and rural soils using GIS-based data, PCA, and spatial interpolation. Environ Pollut. 2009; 157:2378–2385. [PubMed: 19361902]
11. Wang SX, Wang ZH, Cheng XT, et al. Arsenic and fluoride exposure in drinking water: children's IQ and growth in Shanyin county, Shanxi province, China. Environ Health Perspect. 2007; 115(4): 643–647. [PubMed: 17450237]
12. Liu Y, McDermott S, Lawson AB, et al. The relationship between mental retardation and developmental delays in children and the levels of arsenic, mercury and lead in soil samples taken near their mother's residence during pregnancy. Int J Hygiene Environ Health. 2010; 213:116–123.
13. McDermott S, Wu J, Cai B, Lawson AB, et al. Probability of intellectual disability is associated with soil concentrations of arsenic and lead. Chemosphere. 2011; 84:31–38. [PubMed: 21450328]
14. Agresti, A. Categorical data analysis. New Jersey: Wiley & Sons; 2002.
15. McCullagh P. Regression models for ordinal data. J Roy Stat Soc Ser B. 1980; 42:109–142.
16. Anderson JA, Philips PR. Regression, discrimination, and measurement models for ordered categorical variables. Appl Stat. 1981; 30:22–31.
17. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc. 1993; 88:669–679.
18. Cowles MK. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. Stat Comput. 1996; 6:101–110.
19. Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics (with discussion). Appl Stat. 1998; 47:299–350.
20. De Oliveira V. Bayesian prediction of clipped Gaussian random fields. Comput Stat Data Anal. 2000; 34:299–314.

21. Higgs MD, Hoeting JA. A clipped latent variable model for spatially correlated ordered categorical data. Comput Stat Data Anal. 2010; 54:1999–2011.
22. Majumdar A, Gelfand A, Banerjee S. Spatiotemporal change-point modeling. J Stat Plan Infer. 2005; 130:149–166.
23. Stein, ML. Interpolation of spatial data: some theory for Kriging. New York: Springer; 1999.
24. French JL, Wand MP. Generalized additive models for cancer mapping with incomplete covariates. Biostatistics. 2004; 5:177–191. [PubMed: 15054024]
25. Nandram B, Chen M-H. Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. J Stat Comput Simul. 1996; 54:129–144.
26. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2007. http://www.R-project.org
27. Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit. J Roy Stat Soc Ser B. 2002; 64:1–34.
28. Geweke, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Berger, JO.; Bernardo, JM.; Dawid, AP.; Smith, AFM., editors. Bayesian statistics 4. Oxford: Oxford University Press; 1992. p. 169-193.
29. Raftery AE, Lewis SM. One long run with diagnostics: implementation strategies for Markov chain Monte Carlo. Stat Sci. 1992; 7:493–497.
30. Wood SN. Thin plate regression splines. J Roy Stat Soc Ser B. 2003; 65(1):95–114.
31. Newberger D. Down syndrome: prenatal risk assessment and diagnosis. Am Family Phys. 2000; 62:825–832.
32. Lunn DJ, Thomas A, Best N, et al. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput. 2000; 10:325–337.
33. Carroll, R.; Ruppert, D.; Stefanski, L., et al. Measurement error in nonlinear models: a modern perspective. Boca Ratom: Chapman & Hall/CRC; 2006.

Appendix

The full conditional distributions in Section 2.3 are:

Step 1: Update $\tilde{y}_i^*$ from its full conditional distribution,

$$\pi(\tilde{y}_i^*|\cdot) \propto \mathrm{N}(\tilde{y}_i^*;\mu_i,\psi_i)1\{\tilde{\nu}_{y_i-1}<\tilde{y}_i^*<\tilde{\nu}_{y_i}\}.$$

where $\mu_i=(\tilde{\mathbf{y}}_{-i}^*-\tilde{\boldsymbol{\eta}}_{-i})(\tilde{\sum}_{-i}+\phi^2 I_{n-1})^{-1}\boldsymbol{\sigma}_{i,-i}+\tilde{\eta}_i$, $\tilde{\boldsymbol{\Sigma}}_{-i}$ is the submatrix of $\tilde{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\theta})$ excluding the ith row and column, $\boldsymbol{\sigma}_{i,-i}$ is the vector of covariances between $\tilde{y}_i^*$ and $\tilde{\mathbf{y}}_{-i}^*$, and $\psi_i=\sigma_{ii}-\boldsymbol{\sigma}_{i,-i}'(\tilde{\sum}_{-i}+\phi^2 I_{n-1})^{-1}\boldsymbol{\sigma}_{i,-i}$.

Step 2: Update $\tilde{\boldsymbol{\alpha}}$ from its full conditional distribution,

$$\pi(\tilde{\boldsymbol{\alpha}}|\cdot) \propto \mathrm{N}(\tilde{\boldsymbol{\alpha}};\hat{\boldsymbol{\alpha}},\hat{R}_{\tilde{\alpha}}),$$

where $\hat{\boldsymbol{\alpha}}=\varphi^{-2}\,\hat{R_\alpha}\mathbf{U}'(\boldsymbol{\Sigma}(\mathbf{x},\boldsymbol{\theta})+I_n)^{-1}(\tilde{\mathbf{y}}^*-\tilde{\boldsymbol{\eta}_{-\alpha}}\tilde{\mathbf{U}})$ and $\hat{R}_\alpha=\left(R_{\tilde{\alpha}}^{-1}+\phi^{-2}\mathbf{U}'(\sum(\mathbf{x},\boldsymbol{\theta})+I_n)^{-1}\mathbf{U}\right)^{-1}$ with $\mathbf{U}=(\mathbf{u}_1,\ldots,\mathbf{u}_n)'$.

Step 3: Update d_j from its full conditional distribution. To update d_j, we first calculate the probability, $p_{ij}=\mathrm{P}(d_j=w_{ij}|\cdot)=L(d_j=w_{ij}|\cdot)/\sum_{r=1}^{n-1}L(d_j=w_{rj}|\cdot)$, where $\{w_{ij}\}_{i=1}^{n-1}$

denote the ordered $\{z_{ij}\}$ and $L(\cdot)$ denotes the likelihood. Then we draw d_j from the Categorical ($\mathbf{p}_j$), where $\mathbf{p}_j=(p_{1j},\ldots, p_{n-1,j})$.

Step 4: Update $\tilde{\boldsymbol{\beta}}_j$ from its full conditional distribution,

$$\pi(\tilde{\boldsymbol{\beta}}_j|\cdot) \propto \mathrm{N}(\tilde{\boldsymbol{\beta}}_j;\hat{\boldsymbol{\beta}}_j, \hat{R}_{\tilde{\boldsymbol{\beta}}}),$$

where $\hat{\boldsymbol{\beta}}_j=\phi^{-2}\hat{R}_{\tilde{\boldsymbol{\beta}}}\mathbf{z}_j^{*\prime}\left(\sum(\mathbf{x},\boldsymbol{\theta})+I_n\right)^{-1}(\tilde{\mathbf{y}}^*-\tilde{\boldsymbol{\eta}}_{-\tilde{\boldsymbol{\beta}}_j'\mathbf{z}_j^*})$ and

$\hat{R}_{\tilde{\boldsymbol{\beta}}}=\left(R_{\tilde{\boldsymbol{\beta}}}^{-1}+\phi^{-2}\mathbf{z}_j^{*\prime}\left(\sum(\mathbf{x},\boldsymbol{\theta})+I_n\right)^{-1}\mathbf{z}_j^*\right)^{-1}$ with $\mathbf{z}_j^*=(\mathbf{z}_{1j}^*,\ldots,\mathbf{z}_{nj}^*)'$.

Step 5: Update $\tilde{\boldsymbol{\gamma}}_j$ from its full conditional distribution,

$$\pi(\tilde{\boldsymbol{\gamma}}_j|\cdot) \propto \mathrm{N}(\tilde{\boldsymbol{\gamma}}_j;\hat{\boldsymbol{\gamma}}_j, \hat{R}_{\tilde{\gamma}}),$$

where $\hat{\boldsymbol{\gamma}}_j=\hat{R}_{\tilde{\gamma}}\mathbf{V}_j'\left(\tilde{\sum}^{-1}(\mathbf{x},\boldsymbol{\theta})+\phi^{-2}I_n\right)(\tilde{\mathbf{y}}^*-\tilde{\boldsymbol{\eta}}_{-\tilde{\boldsymbol{\gamma}}_j'\mathbf{v}_j})$ and

$\hat{R}_{\tilde{\gamma}}=\left(R_{\tilde{\gamma}}^{-1}+\mathbf{V}_j'\left(\tilde{\sum}^{-1}(\mathbf{x},\boldsymbol{\theta})+\phi^{-2}I_n\right)\mathbf{V}_j\right)^{-1}$ with $\mathbf{V}_j=(\mathbf{V}_{1j},\ldots,\mathbf{V}_{nj})'$ and

$\mathbf{V}_{ij}=\left((z_{ij}-\kappa_{j1})_+^H,\ldots,(z_{ij}-\kappa_{jK})_+^H\right)'$.

Step 6: Update $\tilde{\boldsymbol{\beta}}_j^*$ from its full conditional distribution,

$$\pi(\tilde{\boldsymbol{\beta}}_j^*|\cdot) \propto \mathrm{N}(\tilde{\boldsymbol{\beta}}_j^*;\hat{\boldsymbol{\beta}}_j^*, \hat{R}_{\tilde{\boldsymbol{\beta}}^*}),$$

where $\hat{\boldsymbol{\beta}}_j^*=\phi^{-2}\hat{R}_{\tilde{\boldsymbol{\beta}}^*}[\mathbf{z}_j^{*\prime}\left(\sum(\mathbf{x},\boldsymbol{\theta})+I\right)^{-1}(\tilde{\mathbf{y}}^*-\tilde{\boldsymbol{\eta}}_{-\tilde{\boldsymbol{\beta}}_j^{*\prime}\mathbf{z}_j^*})]_{z_{ij}>d_j}$ and

$\hat{R}_{\tilde{\boldsymbol{\beta}}^*}=\left(R_{\tilde{\boldsymbol{\beta}}^*}^{-1}+\phi^{-2}[\mathbf{z}_j^{*\prime}\left(\sum(\mathbf{x},\boldsymbol{\theta})+I_n\right)^{-1}\mathbf{z}_j^*]_{z_{ij}>d_j}\right)^{-1}$.

Step 7: Update $\tilde{\boldsymbol{\gamma}}_j^*$ from its full conditional distribution,

$$\pi(\tilde{\boldsymbol{\gamma}}_j^*|\cdot) \propto \mathrm{N}(\tilde{\boldsymbol{\gamma}}_j^*;\hat{\boldsymbol{\gamma}}_j^*, \hat{R}_{\tilde{\gamma}^*}),$$

where $\hat{\boldsymbol{\gamma}}_j^*=\phi^{-2}\hat{R}_{\tilde{\gamma}^*}[\mathbf{V}_j'\left(\sum(\mathbf{x},\boldsymbol{\theta})+I_n\right)^{-1}(\tilde{\mathbf{y}}^*-\tilde{\boldsymbol{\eta}}_{-\tilde{\boldsymbol{\gamma}}_j^{*\prime}\mathbf{v}_j})]_{z_{ij}>d_j}$ and

$\hat{R}_{\tilde{\gamma}^*}=\left(R_{\tilde{\gamma}^*}^{-1}+\phi^{-2}[\mathbf{V}_j'\left(\sum(\mathbf{x},\boldsymbol{\theta})+I_n\right)^{-1}\mathbf{V}_j]_{z_{ij}>d_j}\right)^{-1}$.

Step 8: Update ϕ^2 from its full conditional distribution,

$$\pi(\phi^2|\cdot) \propto \mathcal{IG}\left(a+\frac{1}{2}(n+1), b+\frac{1}{2}(\tilde{\mathbf{y}}^*-\tilde{\boldsymbol{\eta}})'\left(\sum(\mathbf{x},\boldsymbol{\theta})+I_n\right)^{-1}(\tilde{\mathbf{y}}^*-\tilde{\boldsymbol{\eta}})\right).$$

Step 9: Update ρ from its full conditional distribution,

$$\frac{\rho^{a-1}}{\phi|\sum(\mathbf{x},\boldsymbol{\theta})+I_n|^{1/2}}\exp\left(-b\rho-\frac{1}{2\phi^2}(\tilde{\mathbf{y}}^*-\tilde{\boldsymbol{\eta}})'(\sum(\mathbf{x},\boldsymbol{\theta})+I_n)^{-1}(\tilde{\mathbf{y}}^*-\tilde{\boldsymbol{\eta}})\right).$$

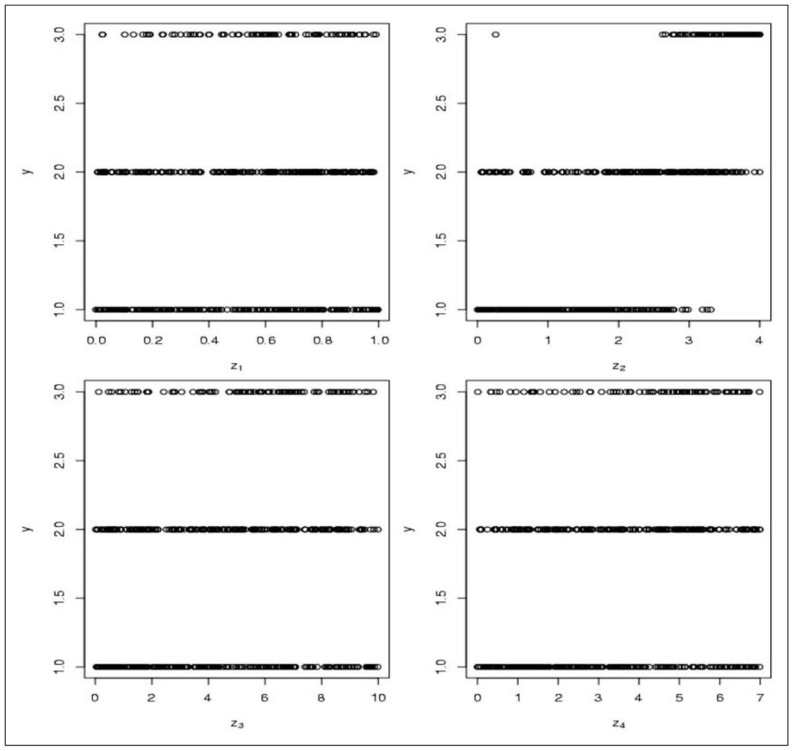

Figure 1.
Plots of covariates vs. the response variable in a simulated data set.

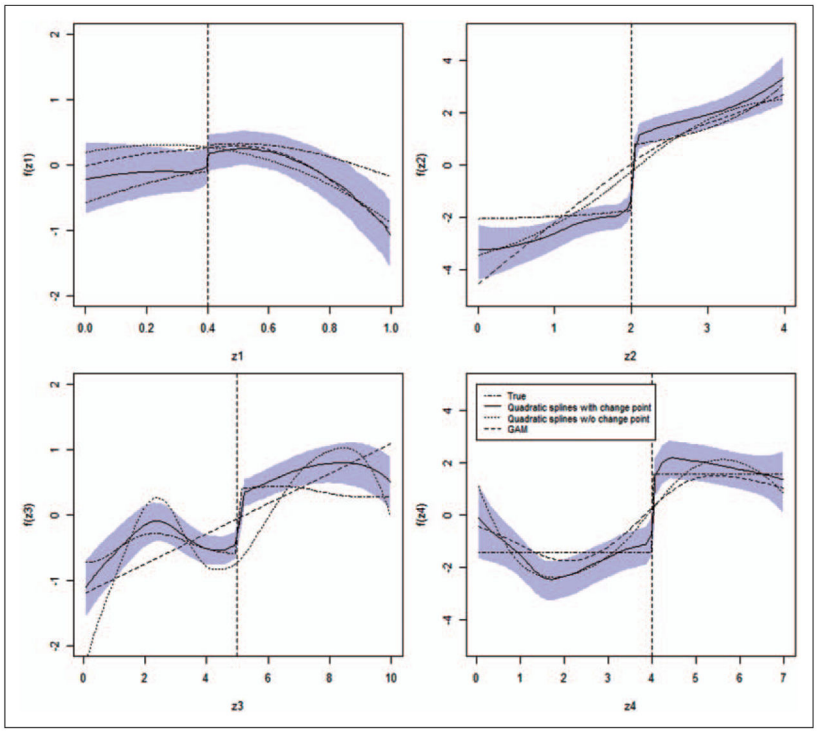

Figure 2.

True and fitted curves from the different methods for simulated data. The vertical line indicates the true value of the change point. The shaded band indicates the 95% pointwise credible intervals from the simulated data sets.

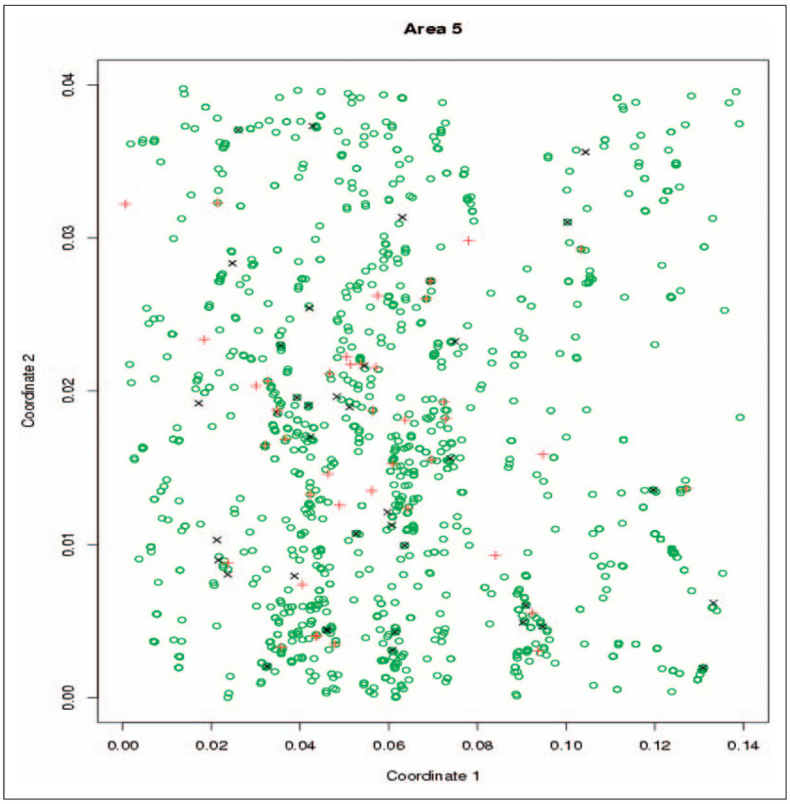

Figure 3.

The ID locations in Area 5. '○' denotes the non-ID case, '×' denotes the mild ID case and '+' denotes the moderate/severe ID case.

ID: intellectual disability.

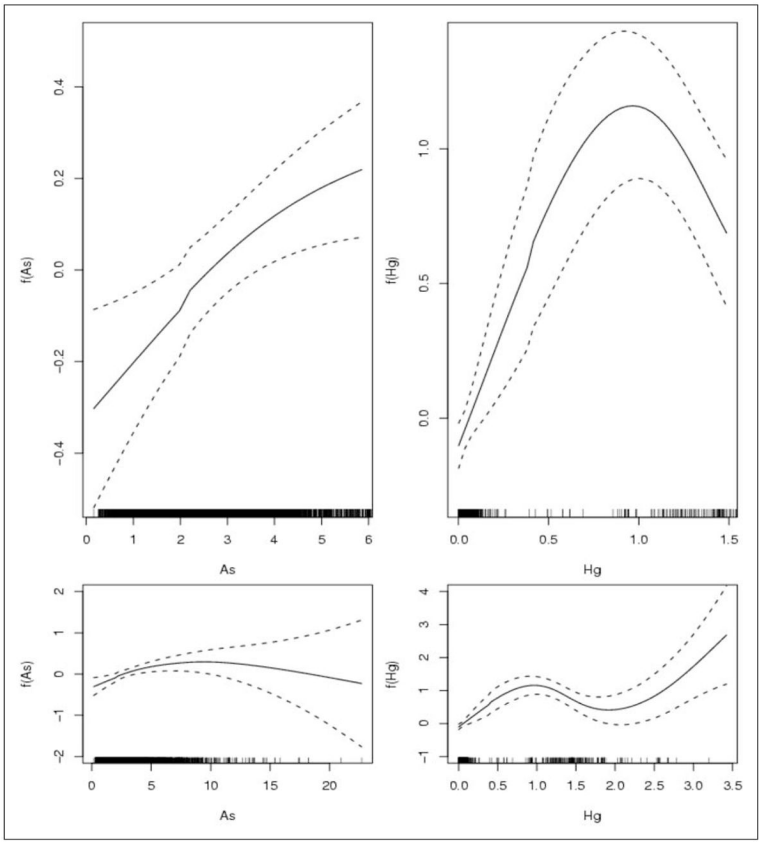

Figure 4.

The estimated curves and 95% pointwise intervals for As and Hg based on the proposed model with change points. The lower panel shows the estimated curves for the entire range of As and Hg. The upper panel shows the estimated curves for the dense range of As and Hg. The rug plots on the horizontal axis represent the data points for As and Hg.

Table 1

Summary of regression coefficients for the demographic covariates in the ID application based on the three models

	Proposed model with thresholds		Proposed model without thresholds		Generalized additive model	
Effect	**Posterior Mean**	**95% C.I.[1]**	**Posterior Mean**	**95% C.I.[1]**	**Mean**	**95% C.I.[2]**
Mother's age	0.03	(0.01, 0.04)	0.03	(0.01, 0.06)	0.04	(0.02, 0.05)
Mother's race[3] (black)	0.29	(0.08, 0.49)	0.28	(0.06, 0.52)	0.28	(0.05, 0.50)
Mother's race (other)	0.022	(−0.68, 0.92)	0.025	(−0.78, 0.80)	0.027	(−0.77, 0.82)
Female child	−0.80	(−0.95, −0.58)	−0.81	(−0.93, −0.53)	−0.73	(−0.90, −0.53)
Mother's alcohol consumption (yes)	−0.06	(−0.81, 0.66)	−0.07	(−0.77, 0.61)	−0.07	(−0.79, 0.64)
Number of prior births =1[4]	0.06	(−0.12, 0.27)	0.05	(−0.19, 0.33)	0.06	(−0.15, 0.28)
Number of prior births =2	0.18	(−0.08, 0.48)	0.18	(−0.10, 0.49)	0.16	(−0.12, 0.42)
Number of prior births =3 or more	0.08	(−0.20, 0.39)	0.09	(−0.23, 0.41)	0.10	(−0.23, 0.43)
Gestational age (week)	−0.07	(−0.09, −0.04)	−0.07	(−0.10, −0.04)	−0.06	(−0.11, −0.02)
Birth weight (kg)	−0.00	(−0.00, −0.00)	−0.00	(−0.00, −0.00)	−0.00	(−0.00, −0.00)

[1] Credible interval.

[2] Confidence interval.

[3] Reference category: Mother's race = white.

[4] Reference category: Number of prior births = 0.

Table 2

Posterior means and 95% credible intervals of change points of the soil metals

Soil metal (mg/kg)	Posterior mean	95% Credible interval
As	2.11	(1.47, 2.78)
Ba	50.25	(2.53, 155.40)
Cr	18.56	(0.81, 67.86)
Cu	32.06	(2.40, 68.47)
Hg	0.42	(0.02, 0.96)
Mn	104.0	(5.95, 286.45)
Ni	7.73	(0.03, 20.52)
Pb	105.86	(5.40, 325.67)