Title

A semi-parametric Bayesian model for semi-continuous longitudinal data

Permalink
https://escholarship.org/uc/item/3wf2t9gb

Journal
Statistics in Medicine, 41(13)

ISSN
0277-6715

Authors
Ren, Junting
Tapert, Susan
Fan, Chun Chieh
et al.

Publication Date
2022-06-15

DOI
10.1002/sim.9359

Peer reviewed

Published in final edited form as:
Stat Med. 2022 June 15; 41(13): 2354–2374. doi:10.1002/sim.9359.

A semi-parametric Bayesian model for semi-continuous longitudinal data

Junting Ren[*,1,4], **Susan Tapert**[2], **Chun Chieh Fan**[3,4], **Wesley K. Thompson**[4,5]

[1]Division of Biostatistics, Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla, USA

[2]Department of Psychiatry, University of California San Diego, La Jolla, USA

[3]Center for Human Development, University of California San Diego, La Jolla, USA

[4]Population Neuroscience and Genetics Lab, University of California San Diego, La Jolla, USA

[5]Department of Radiology, University of California San Diego, La Jolla, USA

Summary

Semi-continuous data present challenges in both model fitting and interpretation. Parametric distributions may be inappropriate for extreme long right tails of the data. Mean effects of covariates, susceptible to extreme values, may fail to capture relevant information for most of the sample. We propose a two-component semi-parametric Bayesian mixture model, with the discrete component captured by a probability mass (typically at zero) and the continuous component of the density modeled by a mixture of B-spline densities that can be flexibly fit to any data distribution. The model includes random effects of subjects to allow for application to longitudinal data. We specify prior distributions on parameters and perform model inference using a Markov Chain Monte Carlo (MCMC) Gibbs-sampling algorithm programmed in R. Statistical inference can be made for multiple quantiles of the covariate effects simultaneously providing a comprehensive view. Various MCMC sampling techniques are used to facilitate convergence. We demonstrate the performance and the interpretability of the model via simulations and analyses on the National Consortium on Alcohol and Neurodevelopment in Adolescence study (NCANDA) data on alcohol binge drinking.

Keywords

Semi-parametric; Bayesian; semi-continuous; longitudinal; B-spline; MCMC

1 | INTRODUCTION

Data that are a mixture of continuous values and a set of frequently-observed discrete values are often termed *semi-continuous*.[1] Discrete values often occur at zero and effectively continuous values are positive, right-skewed and with substantial heteroscedasticity.[2] Zero-

[*]Corresponding author Junting Ren. j5ren@ucsd.edu.

Conflict of Interest: None declared.

inflated semi-continuous data abound in practical applications, including questionnaire assessments,[3] medical costs,[4] microbiome data[5] and single cell gene expression.[6] An example of semi-continuous data is given by self-reported *number of binge drinking episodes in the past year* by participants of the cohort-sequential National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA) Study. Over 65% of the number of binge drinking episodes in the past year is concentrated at 0, but the count is widely distributed and highly right skewed, with a maximum of 170 episodes for one subject (see Figure 1b). Semi-continuous data present challenges in both model fitting and interpretation. Parametric distributions are inadequate for the extreme long right tails of the data, and mean effects of covariates obtained from traditional models fail to capture information for most of the population since means are heavily influenced by extreme values.

Two main approaches have been utilized to model semi-continuous outcomes, i.e. Tobit models[7,8] and two-part models.[9] A Tobit model assumes that outcomes are zero only because they fail to reach the detection level so the zeros are not true zeros, whereas a two-part model model treats the zero as true values and separately describes the probability of the outcome being positive and the magnitude of positive values. For many data types such as number of binge drinking episodes, the Tobit model is not appropriate since there is no meaningful definition of detection level. Therefore, here we focus on the two-part model:

- Part I: $P(Z > 0 \mid X) = p(X)$ and $P(Z = 0 \mid X) = 1 - p(X)$
- Part II: Modeling of the distribution $[Z \mid Z > 0, X]$

Different procedures have been implemented to model the positive (continuous) values as a function of covariates. Modeling the response as a log-normal distribution is a common approach to address the skewness and heteroscedasticity of the original distribution, but such strategy has issues on retransformation, interpretation, and inadequacy in fitting the data.[10,11] Another approach is to model the positive continuous values with exponential family distributions such as the generalized gamma distribution, log skew normal distribution or normal after Box-Cox transformation, and corresponding random effects models have been proposed.[4,2] Inferences from the positive component exclude zero values and only refer to the sub-population of those with positive outcomes. Therefore, a marginal interpretation of coefficients for the above two-component models has been proposed.[12] To take account of the heteroscedasticity of the exponential families as well as the non-linear effects of the covariates, others have proposed to model the means and variances as smooth function of covariates.[2,13,14] To address the problem that the above parametric approaches may not provide adequate fit for more extreme distributions, generalized estimating equations have been proposed, with mean and variance as smooth functions of the covariates.[15,16]

All the above methods focus primarily on covariate effects on the mean, and little attention has been paid to the other aspects of the distribution. For example, from both policy and clinical perspectives, the sub-population of "heavy drinkers" in the NCANDA Study is of great interest.[17,18] Moreover, the mean of an extremely skewed distribution can be a misleading summary statistic. From Figure 1a, only a handful of lines out of 400 subjects are above 25 and fluctuate (go up and down), but the highest values can go above 100, resulting in large mean number of binge-drinking episodes due to the extreme positive values whereas the median is relatively small.

To maintain both interpretability and adequate fit, we propose a two-part model, with the discrete component captured by a probability mass (at zero) and the continuous component of the density modeled semi-parametrically via a mixture of B-spline densities (B-spline basis functions normed to integrate to one) with weights dependent on covariates. Using the property that B-splines have local support,[19] we not only obtain a numerically stable and flexible estimation of the covariate-dependent density but also a local interpretation of the coefficients for the covariates. Others have shown that a mixture of B-splines densities provide better estimation compared to other parametric and non-parametric methods.[20,21,22] Furthermore, with our Bayesian Markov Chain Monte Carlo (MCMC) algorithm, we can perform inference on multiple functions of the posterior distribution simultaneously, e.g., the marginal interpretation in terms of mean, median or any quantile of the distribution simultaneously with credible intervals.

This paper contributes to both methodology and graphical interpretability in analyzing zero-inflated semi-continuous data. In the first part, we explain the details of our semi-parametric Bayesian zero-inflated mixture model that can account for cluster-correlated data. Then, we examine the performance of the algorithm in Monte Carlo simulations, comparing the performance to the generalized gamma model, and demonstrate its interpretability in a motivating application using the longitudinal number of binge-drinking episodes from the NCANDA dataset. The algorithm is implemented as an R script and freely available for download on GitHub: https://github.com/junting-ren/zero_inflated_b_spline.

2 | METHODS

Let Z_{ij} be random variables such that $Z_{ij} \geq 0$, $i = 1, \ldots, N$, $j = 1, \ldots, J_i$, where Z_{ij} denotes the value for the i^{th} subject at the j^{th} time point. We consider the scenario, where for each Z_{ij}, we also have an $(M+1)$-dimensional vector of covariates (including intercept) denoted by $\boldsymbol{x}_{ij} = (1_{,1ij}, x_{2ij}, \ldots, x_{Mij})^T$. The Z_{ij} are assumed independent conditional on $\boldsymbol{x}_{ij}$, *part I* random intercept $b_{\delta i}$ and *part II* random intercept $b_{\eta i}$, with marginal density f given by

$$f\left(z_{ij} \middle| \boldsymbol{x}_{ij}, b_{\delta i}, b_{\eta i}\right) = \pi_0\left(\boldsymbol{x}_{ij}, b_{\delta i}\right) \mathbb{1}\left(z_{ij} = 0\right) + \pi_1\left(\boldsymbol{x}_{ij}, b_{\delta i}\right) f_1\left(z_{ij} \middle| \boldsymbol{x}_{ij}, b_{\eta i}\right),$$

where $\pi_0\left(\boldsymbol{x}_{ij}, b_{\delta i}\right) = 1 - \pi_1\left(\boldsymbol{x}_{ij}, b_{\delta i}\right)$, $\mathbb{1}\left(z_{ij} = 0\right)$ is 1 when $z_{ij} = 0$ otherwise 0, and f_1 is a probability density with support on the positive real numbers which will be approximated via a finite mixture of B-spline densities. The part I random intercept captures individual-level variation in probability of z_{ij} being drawn from the zero or positive component, and the part II random effect $b_{\eta i}$ capture individual variation in the positive component. It is natural to conjecture that the two processes that generate semi-continuous data may be related. Thus, we assume the part I random intercepts $b_{\delta i}$ are correlated with the part II random effects $b_{\eta i}$ by specifying a joint multivariate Gaussian distribution prior. If they are truly associated, then the correlation structure will reduce bias in estimating the fixed effect coefficients.[23]

2.1 | Part I

We first introduce a global indicator vector $\boldsymbol{\delta} = \left(\delta_{11}, \ldots, \delta_{N\boldsymbol{J}_N}\right)^T$, where $\delta_{ij} = 1$ if $z_{ij} > 0$ and $\delta_{ij} = 0$ otherwise, and $\sum_{i=1}^{N} \boldsymbol{J}_i$ is the total number of observations. It is assumed that $\delta_{ij} \sim$ Bernoulli$\{\pi_1(x_{ij}, b_{\delta i})\}$, where

$$\pi_1\left(\boldsymbol{x}_{ij}, b_{\delta i}\right) = \boldsymbol{P}\left(\delta_{ij} = 1 \middle| \boldsymbol{\gamma}, \boldsymbol{x}_{ij}, b_{\delta i}\right) = \boldsymbol{\Phi}\left(\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + b_{\delta i}\right),$$

Φ is the standard normal cumulative distribution function (CDF) and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_M)^T$ is an $(M + 1)$-vector of unknown parameters.

Let $\boldsymbol{X}$ denote the $(M + 1) \times \sum \boldsymbol{J}_i$ covariate matrix with columns $\boldsymbol{x}_{ij}$. Then the joint density of $\boldsymbol{\delta}$ given $\boldsymbol{\gamma}$, random effect $b_{\delta i}$ and covariates $\boldsymbol{X}$ is given by

$$f_{\delta}(\boldsymbol{\delta} | \boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{b}_{\boldsymbol{\delta}}) = \prod_{i=1}^{N} \prod_{j=1}^{\boldsymbol{J}_i} \left\{ \left[\Phi\left(\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + b_{\delta i}\right)\right]^{\delta_{ij}} \left[1 - \Phi\left(\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + b_{\delta i}\right)\right]^{1 - \delta_{ij}} \right\}.$$

To implement the probit model, we introduce latent variable y_{ij} such that

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + b_{\delta i} + e_{ij},$$

where $e_{ij} \sim N(0,1)$. The latent variables y_{ij} are measured indirectly by the observed binary variable δ_{ij}, so y_{ij} can be defined by:

$$\delta_{ij} = \begin{cases} 0 & \text{if} \;\; y_{ij} \le 0 \\ 1 & \text{if} \;\; y_{ij} > 0 \end{cases}$$

Using a probit regression model, we obtain a closed form expression for the posterior of $\boldsymbol{\gamma}$ and $b_{\delta i}$, making the convergence faster and estimation more accurate when there are multiple global random effects, comparing to a logistic model. [24]

2.2 | Construction of B-spline density basis

Before we introduce the model for part II, we first discuss the construction of the B-spline density basis. A spline of order d+1 is a piecewise polynomial of degree d with continuous derivatives up to order d−1. This is only true when adjacent polynomial pieces have the same value at the knots up to the d–1th derivative. Each such condition constitutes a linear constraint. Therefore, a spline of degree d with internal knots $\xi_1, \xi_2, \ldots, \xi_Q$ is determined by $(Q+1)(d+1) - Qd = Q + d+1$ bases. Instead of fitting the spline by least squares subjected to linear constraints, different kinds of basis functions are used in practice. For example, the truncated power series basis for spline of degree d with internal knots $\xi_1, \xi_2, \ldots, \xi_Q$

are $\left\{1, z, \ldots, z^d, (z-\xi_1)_+^d, \ldots, \left(z-\xi_Q\right)_+^d\right\}$. However, the supports of truncated power series are not local, with some of the basis defined over the whole range. This might lead to high correlations among some basis, leading to numerical instabilities in estimation. Hence, it is often preferred to use a B-spline basis. Given the internal knots $\xi_1, \xi_2, \ldots, \xi_Q$ and boundary knots ξ_0, ξ_{Q+1}, we define the following new knot sequence:

$$\tau_q = \begin{cases} \xi_0 & \text{for } q = 1, \ldots, d+1 \\ \xi_{c-d-1} & \text{for } q = d+2, \ldots, d+Q+1 \\ \xi_{Q+1} & \text{for } q = d+Q+2, \ldots, 2d+Q+2 \end{cases}$$

Then, the B-spline basis functions of degree d are defined by the recursive formula

$$B_q^d(z) = \frac{z-\tau_q}{\tau_{q+d}-\tau_q} B_q^{d-1}(z) - \frac{\tau_{q+d+1}-z}{\tau_{q+d+1}-\tau_{q+1}} B_{q+1}^{d-1}(z)$$

$$q = 1, \ldots, Q+d+1$$

where

$$B_q^0(z) = \begin{cases} 1, & \tau_q \leq z < \tau_{q+1} \\ 0, & \text{else} \end{cases}$$

and $B_q^0(z) \equiv 0$ if $\tau q = \tau q+1$. It follows that they are larger than zero in intervals spanned by d + 2 knots and zero elsewhere, which results in high numerical stability.[25,20,26]

For our application, we need to construct B-spline basis for a density function, which adds the constraint that the area under each B-spline basis function adds up to 1. Therefore, we normalize the B-spline basis to integrate to one:

$$g_q(z) = \frac{d+1}{\tau_{q+d+1}-\tau_q} B_q^d(z),$$

since $\int_{\tau_q}^{\tau_{q+d+1}} B_q^d(z)dz = \frac{\tau_{q+d+1}-\tau_q}{d+1}$. We constructed B-spline density basis on the range of the positive observations (for example the binge drinking number) that can be used to model the likelihood function for z. Our model takes the number of basis functions $K > 4$ as a user-specified input, and constructs a cubic (i.e., $d = 3$) B-spline density basis with $K - 2$ internal knots and hence $K - 2 + d + 1 = K + 2$ basis functions. However, the last two B-spline density functions are not needed for most data since the probability of observing data at the right end of the distribution is typically low. Therefore, in our applications we remove the last two basis functions supported in the right tail, resulting in K remaining B-spline density basis functions.

For example, with the NCANDA data, we constructed $K = 5$ B-spline densities that are supported from 0.01 to the maximum observed binge drinking counnt of 170, giving three equally-spaced internal knots at 42.51, 85.01, and 127.51 (Figure 2). Observe that the B-spline density functions are intrinsically ordered on their intervals of support, e.g., the first B-spline density on the left has its mode at 0.01, whereas the second B-spline density from the left has a mode at around 17, etc. We utilize this ordering in building part II of our model.

2.3 | Part II

The density f_1 for the positive part is approximated by a finite mixture of B-spline densities with weights that vary as a function of covariates. In the remainder of the paper, we use cubic B-spline densities with knot number and position fixed by the researcher. Rather than focus on knot position selection, we include enough knots to allow a flexible fit.[27,28] The validity of this approach is evaluated in the simulation studies.

Specifically, the density of the non-zero observations is approximated by

$$f_1\left(z_1 \middle| \boldsymbol{X}_1, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{b}_{\boldsymbol{\eta}}\right) = \prod_{ij: \delta_{ij} = 1} \left\{ \sum_{k = 1}^{K} c_{kij} g_k\left(z_{ij}\right) \right\},$$

where $\boldsymbol{z}_1$ is the vector of observations corresponding to positive values. Let $N_1 = \sum_{i = 1}^{N} \sum_{j = 1}^{\boldsymbol{J}_i} \delta_{ij}$ and let $\boldsymbol{X}_1$ denote the corresponding $(M + 1) \times N_1$ matrix of covariates. The g_k are the cubic B-spline densities and c_{kij} are non-negative weights with the constraint $\sum_{k = 1}^{K} c_{kij} = 1$. Equivalently, c_{kij} is the probability that the i^{th} subject's j^{th} time point belongs to the k^{th} B-spline density, conditioned on $\delta_{ij} = 1$. To speed up the convergence of MCMC algorithm, we introduce a local latent indicator vector $\boldsymbol{\eta} = \left(\eta_{11}, \eta_{12}, \ldots, \eta_{NJ_N}\right)^T$. The element $\eta_{ij} \in \{0,1, \ldots, K\}$ for $i = 1, \ldots, N, j = 1, \ldots, J_i$. Latent indicator $\eta_{ij} > 0$ only if $\delta_{ij} = 1$, and hence specifies which B-spline density the non-zero value for subject i at time point j is generated from.

We consider two different modelling approaches to estimate the weights c_{kij}: 1) an ordinal probit model; and 2) a multinomial logistic model. In the main text we focus on the ordinal probit model. We provide the formulation, estimation, code and simulation results for the multinomial logistic model in the Supplementary Materials.

For the ordinal probit model, we treat η_{ij} as an ordinal categorical variable, which means that it represents an assignment into one of K mutually exclusive and exhaustive ordered categories of B-spline densities. The ordinal probit regression model can be written in terms of an additional latent variable l_{ij} conditional on $\delta_{ij} = 1$ as follows:

$$l_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} + b_{\eta i} + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0,1)$, and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_M)$ are the fixed effects for part II, and $b_{\eta i}$ is the random intercept. Similarly to part I, l_{ij} can be measured indirectly by the ordinal category η_{ij}, but the difference is that η_{ij} itself is also a latent variable and can be defined in terms of l_{ij}:

$$\eta_{ij} = \begin{cases} 1 & \text{if} -\infty < l_{ij} \le \lambda_1 \\ 2 & \text{if } \lambda_1 < l_{ij} \le \lambda_2 \\ \vdots & \\ K & \text{if } \lambda_{K-1} < l_{ij} < \infty \end{cases}$$

The first threshold, λ_1, defines the upper bound of the interval corresponding to $\eta_{ij} = 1$, indicating that the observed z for subject i at time point j comes from the first B-spline density component. Threshold λ_{K-1} defines the lower bound of the interval corresponding $\eta_{ij} = K$. Threshold parameters are $\boldsymbol{\lambda}^T = (\lambda_0 = \lambda_{min} < \lambda_1 < \ldots < \lambda_{K-1} < \lambda_K = \lambda_{max})$ with $\lambda_{min} = -\infty$ and $\lambda_{max} = \infty$. For identification of the parameters, we can set $\lambda_1 = 0$ or fix the intercept $\alpha_0 = 0$.[29,30] The choice of setting $\lambda_1 = 0$ has the benefit of facilitating posterior sampling (since sampling α_0 is generally easier than sampling λ_1) and also makes the ordinal probit model theoretically consistent with part I binary probit model. The cumulative response probability for the k^{th} B-spline density is:

$$\mathbb{P}\left(\eta_{ij} \le k\right) \middle| \boldsymbol{\alpha}, b_{\eta i}\right) = \mathbb{P}\left(l_{ij} \le \lambda_k\right) \middle| \boldsymbol{\alpha}, b_{\eta i}\right) = \mathbb{P}\left(\epsilon_{ij} \le \lambda_k - \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} - b_{\eta i}\right) = \Phi\left(\lambda_k - \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} - b_{\eta i}\right).$$

Therefore,

$$c_{kij} = \mathbb{P}\left(\eta_{ij} = k\right) \middle| \boldsymbol{\alpha}, b_{\eta i}\right) = \Phi\left(\lambda_k - \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} - b_{\eta i}\right) - \Phi\left(\lambda_{k-1} - \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} - b_{\eta i}\right).$$

Conditional on $\delta_{ij} = 1$, $\eta_{ij} \sim \text{Multinomial}(\boldsymbol{c}_{ij})$, where $\boldsymbol{c}_{ij} = (c_{1ij}, \ldots, c_{Kij})^T$. The joint density of $\boldsymbol{\eta}$ given $\boldsymbol{\delta}$, $\boldsymbol{\alpha}$, $\boldsymbol{b}_{\eta}$, and $\boldsymbol{X}_1$ is given by

$$\begin{aligned} f_\eta\left(\boldsymbol{\eta} \middle| \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{b}_{\boldsymbol{\eta}}, \boldsymbol{X}_1\right) &= \prod_{ij:\delta_{ij}=1} \prod_{k=1}^{K} \left\{ \mathbb{P}\left(\eta_{ij} = k \middle| \boldsymbol{\alpha}, \boldsymbol{b}_{\boldsymbol{\eta i}}\right)^{\boldsymbol{I}\left(\eta_{ij} = k\right)} \right\} \\ &= \prod_{ij:\delta_{ij}=1} \prod_{k=1}^{K} \left[\left\{ \Phi\left(\lambda_k - \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} - b_{\eta i}\right) - \Phi\left(\lambda_{k-1} - \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} - b_{\eta i}\right) \right\}^{\boldsymbol{I}\left(\eta_{ij} = k\right)} \right]. \end{aligned}$$

In summary, in part I of the model, the covariates and part I random effect modulate the probability of the zero and positive status of each observation. In part II of the model (for the positive values such that $\delta_{ij} = 1$), the covariates and part II random effects modulate the probability of which B-spline component the observation $z_{ij} > 0$ is drawn from.

Here, we list a few important reasons that an ordinal model is preferred over a multinomial logistic model for modeling the weights. First, the ordinal probit model for part II is consistent with the part I binary probit model. Second, the B-spline density bases' modes are ordered across the support of the outcome z as shown in Figure 2, so it is natural to assume an underlying latent linear model such that larger values for latent variable I correspond to observations generated from B-spline densities supported on larger values of the observed outcome range. This approach is commonly utilized, for example, in genetic models for ordinal outcomes.[31,32] Third, multinomial logistic regression requires significantly more parameters as the number of B-spline density basis increases, for example a model with M covariates and K B-spline densities, the multinomial logistic regression requires $(K-1) \times M$ fixed effect parameters and $K-1$ random intercepts comparing to only $M+K-1$ fixed effect parameters and 1 random intercept for the ordinal probit model. Fourth, correlation between the random intercepts in part I and part II can be easily accounted for in the ordinal model, whereas it is impractical to estimate the correlation for a large number of random effects in a multinomial regression model. Fifth, in our simulations the ordinal probit model can handle longitudinal data with as few as 5 observations per subject, whereas the multinomial logistic model requires at least 20 observations per subject to give adequate estimates of the random effects (as demonstrated in the Supplementary Materials). Finally, it will be much easier (in future work) to adapt the ordinal probit model in applications to high-dimensional datasets, e.g., when the number of features is greater than the number of observations (such as is the case in many genetic data applications).

The trade-off is that the multinomial logistic model requires fewer assumptions and is more flexible than the ordinal model. This flexibility is demonstrated when fitting the two models to the generalized gamma distribution data in the cross-sectional settings model comparison section. In the main text, unless specified otherwise, we used the ordinal probit model for part II weights in simulations and real data analysis.

2.4 | Prior distribution

Here, we specify prior distributions for coefficients and the covariance of the random effects. For part I fixed effects $\boldsymbol{\alpha}$,

$$\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}),$$

and for part II fixed effects $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right),$$

where hyper-parameter $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ are specified by the user. In the simulations and data application, we set both to be diagonal with variance 2.5.

It is natural to conjecture that the two processes that generate semi-continuous data may be related. Thus, for the random effects $\boldsymbol{b} = (b_{\delta i}, b_{\eta i})$ we propose a multivariate Gaussian distribution prior

$$\boldsymbol{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{b}}).$$

Further, we assume non-informative priors on the covariance $\boldsymbol{\Sigma}_{\boldsymbol{b}}$ of the random intercepts:

$$\boldsymbol{\Sigma_b} \sim \text{Inverse Wishart}(2, \boldsymbol{I_2}) .$$

We thus sample $(b_{\delta i}, b_{\eta i})$ simultaneously from their joint conditional posterior, which induces correlation between the part I and part II random effects. This can lead to improved estimates of the fixed effects[23] when there is correlation between the random effects in the two parts.

The prior distribution for $K - 1$ thresholds λ are given as order statistics from $(\lambda_{min}, \lambda_{max})$ distribution,[32]

$$\mathbb{P}(\lambda) \propto \mathbb{1} \left(\lambda_{\min} < \lambda_1 < \ldots < \lambda_{K-1} < \lambda_{\max}\right) .$$

2.5 | Sampling Scheme

Here, we outline the MCMC sampling algorithm for the ordinal probit model. Details on the conditional posterior distributions are given in the Supplementary Materials. To reduce autocorrelation between random effects and fixed effects, we implemented block sampling algorithm for the both part I and part II part of the model.24 The sampling scheme consists of the following steps:

1. Block sample $\boldsymbol{y}$, $\boldsymbol{l}$, $\boldsymbol{\gamma}$, $\boldsymbol{a}$ and $\boldsymbol{b}_\delta, \boldsymbol{b}_\eta$ from $f(\boldsymbol{y}, \boldsymbol{l}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{b}_\delta, \boldsymbol{b}_\eta | \Sigma_{\boldsymbol{b}}, \Sigma_{\boldsymbol{a}}, \Sigma_{\boldsymbol{\gamma}}) \propto f(\boldsymbol{b}_\delta, \boldsymbol{b}_\eta | \boldsymbol{y}, \boldsymbol{l}, \boldsymbol{\gamma}, \boldsymbol{a}, \Sigma_{\boldsymbol{b}}) f(\boldsymbol{y}, \boldsymbol{l}, \boldsymbol{\gamma}, \boldsymbol{a} | \Sigma \boldsymbol{b}, \Sigma \boldsymbol{a}, \Sigma \boldsymbol{\gamma})$

2. Sample $\boldsymbol{\Sigma_b}$ from $f(\Sigma_{\boldsymbol{b}} | \boldsymbol{b}_\delta, \boldsymbol{b}_\eta)$

3. Sample thresholds $\boldsymbol{\lambda}$ from $f(\boldsymbol{\lambda} | \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{a}, \boldsymbol{b}_\eta)$ integrating over the latent variable $\boldsymbol{l}$ to speed up the convergence.

4. Sample η_{ij} for $i = 1, .., N$ and $j = 1, \ldots, J_i$ from $f(\eta_{ij} | \delta_{ij} = 1, b_{\eta i}, \boldsymbol{\lambda}, \boldsymbol{a})$

2.6 | Evaluation of covariate effects

An important advantage of estimating the joint posterior using MCMC is that we can obtain the posterior distribution of any function of the parameters. Therefore, in addition to making inferences, e.g., for the change of the mean of Z for one unit increase in a covariate, we can estimate the posterior distribution for the change of any quantile of Z for one unit increase in a covariate. Furthermore, similar to the linear regression model, a positive coefficient $\boldsymbol{a}_m$ implies that as the m^{th} covariate increases, the probability of observing a larger outcome increases. This greatly improves the interpretability of model results for zero-inflated and highly skewed data, as demonstrated in the simulation study and real data example.

3 | SIMULATIONS

We performed Monte Carlo simulation studies to investigate the performance of the proposed two-part probit model. First, to evaluate the basic performance for the model, a total of 500 datasets was generated with longitudinal settings, varying the number of subjects N and time points $\boldsymbol{J}$ for each subject, as well as cross-sectional settings with different number of subjects N with $\boldsymbol{J}$ fixed at 1. Then, we investigated how the number of B-splines basis functions and placement of the interior knots affects inferences. Finally, we compared the performance of our model to the generalized gamma model[4,2] using data generated from either a mixture of B-spline densities or a generalized gamma distribution.

3.1 | Longitudinal setting

For the longitudinal simulations, we mimicked the skewed longitudinal binge drinking data from NCANDA. One of the most prominent characteristic of the NCANDA data is that most of the subjects' observations are 0, with only a handful of subjects' observations being extremely large. Consequently, increases in covariates will result in increase, e.g., in the mean and 75th quantile of the outcome but not the median. Therefore, we set our both part I and part II fixed coefficients to be relative small compared to the thresholds. In this case, without the random intercepts, the linear combinations of the fixed effect coefficients will be much lower than the threshold for B-spline density basis on the far right of the distribution, resulting in simulated data with most of the subjects having zeros or small values and a few subjects with consistently large values of the outcome. The simulation setup is the following:

1. Two independent explanatory variables where $X_1 \sim N(0,1)$ and $X_2 \sim Bernoulli(p = 0.5)$.
2. For the part I Binomial probit model, we generate the probability of non-zero observation using coefficients $\boldsymbol{\gamma} = (-1.0,1.2,1.0)$, where the first coefficient is the value of intercept.
3. For the part II mixture model, the data were generated from $K = 5$ B-splinedensitieswithsupportfrom0.01to150.00with 3 knots equally spaced. The ordinal probit model fixed effect coefficients are $\boldsymbol{\alpha} = (-1.00,0.35,0.60)$ and the thresholds are $\boldsymbol{\lambda} = (0.00,1.90,3.00,3.50)$.
4. For the part I and part II random effects $\boldsymbol{b}_i = (b_{\delta i}, b_{\eta i})$, we set the correlation to be either at 0.20 or 0.60 and the variances to be 1.10^2 and 2.00^2 respectively.
5. By the above setup, the baseline ($X_1 = 0_, X_2 = 0$) population mean, median and 75th quantile are 1.58, 0.00 and 0.00 respectively. For every one unit increase in X_1 or X_2, the population mean will increase 5.09 or 4.89 comparing to the baseline, respectively; the population median will increase 1.84 or 0.00, respectively; the population 75th quantile will increase 10.01 or 9.48 comparing to the baseline, respectively.

Here, we only showcase three possibilities when calculating the functions of the distribution (mean, median and 75th quantile). In practice, researchers can calculate any function of the distribution in one run.

Table 1 presents the bias (sample mean departure from the truth due to the finite sample size), standard deviation, mean square error and coverage percentage of the credible intervals over 500 simulations for covariate effects on the functions of distribution for each simulation setup. We simulated data of total sample size ranging from $N_{total} = 500$ to $N_{total} = 2000$. Within each sample size N_{total}, each subject has a number of longitudinal observations that ranges from $J = 5$ to $J = 20$. For example, if the sample size is $N_{total} = 1000$ and number of observation per subject is $J = 5$, then we would have $N = N_{total}/J = 1000/5 = 200$ subjects.

For all simulation scenarios, the coverage rates were all close to the nominal 95% level. As the total sample size increased from 500 to 2000, the bias and variances of the covariate effect estimators decreased as expected. For the mean covariate effects, the bias and variance of the estimators are acceptable (relative to the original scale of the true covariate mean) even at the smallest sample size $N_{total} = 500$. For instance, for $N_{total} = 500$ and $J = 5$, the true value for the baseline mean is 1.58, whereas the estimator has a bias of 0.32 and standard deviation of 0.61. We also investigated whether increasing the correlation of the part I and part II random intercepts would lead to increase of bias, so for $J = 5$, we simulated data with either a correlation of 0.2 or 0.6 (the variances are fixed same as before). There is no observed difference between the two different cases. Within a fixed total sample size, as the number of per subject observations J decreases, the bias and variance decreases. For example, at total sample size $N_{total} = 2000$, when $J = 5$, the mean square error (MSE) is 2.22 for the 0.75 quantile effect of one unit increase in X^2, but for $J = 20$, the MSE is 4.32.

Estimation of parameters $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ are displayed in Supplementary Material Table S1. The fixed effects $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ for all sample size and number of observation per subject are estimated with low bias, standard deviation and posterior credible intervals achieved the nominal coverage level. As for the thresholds $\boldsymbol{\lambda}$, the bias decreases and coverage increases as the total number of sample or number of observations per subject increases. When the sample size is small, the higher the threshold is, the more bias in estimation. This is due to observations coming from the B-spline density basis supported at the larger values are sparse, thus making it harder to estimate the threshold values on the high end. Interestingly, almost all threshold estimations only achieved around 85% coverage rate, but the coverage rate for the different effect of the covariates still maintained the nominal coverage level 95%. We speculate that this is due to the high correlation between the cubic B-spline density basis that led to identification problems for auxiliary parameters but the functions of the auxiliary parameters are still identifiable.[33,25] The part I and part II random effect variance coverage rates can be found in supplement Figure S1. Part I random effect variance coverage rates stayed stable at 95% nominal level regardless of the sample size. For part II random effect variance, as sample size or number of observations per subject increased, the coverage rate increased.

For Figure 3, the x-axis, ranging from 0 to 150, is the whole support of positive part of the semi-continuous distribution; The y-axis, starting from 0, is the probability density of specific point in the support and the gray shaded area is the 95% credible band. Figure 3a displays overall model fit when $X_1 = 1$, $X_2 = 1$ when the total sample size is $N_{total} = 2000$ and number of observations per subject is $J = 20$ for one realization of the simulated data. Figure 3b displays model fits for 9 randomly selected subjects in the same realization. We took the average of the selected subject's covariates across the different time points. Using the estimated mean fixed effects, thresholds and subject's random intercept, we calculated the weight for each B-spline density basis at the average covariate values for the specific subject. Using the estimated weights and the true weights, we plotted both the estimated and true positive valued distribution for that specific averaged covariates values and individual random effect. Although some of the individual distributions are very different from the population distribution, the estimated model density did a good job capturing the true individual density because of the subject-specific random intercept.

3.2 | Cross-sectional setting

We increased the coefficients for B-spline densities to $\boldsymbol{\alpha} = (-1,1.8,1.6)$ and kept the other parameters $(\lambda, \boldsymbol{\gamma})$ the same as the longitudinal setting. If we keep the same $\boldsymbol{\alpha}$ as in the longitudinal setting, the observations would be limited to the left half of the distribution (small values) due to the finite sample size and no random effects to increase the probability of observing values coming from the B-spline densities supported at the right tail of the distribution. For this new setting, the baseline $(X_1 = 0_, X_2 = 0)$ population mean, median and 75th quantile are 1.58, 0.00 and 0.00 respectively. For every one unit increase in X_1 or X_2, the population mean will increase 11.69 or 8.86, respectively; the population median will increase 5.54 or 0.00, respectively; the population 75th quantile will increase 22.21 or 17.42, respectively.

As shown in Table 2, the performance of the cross-sectional model is even better than the longitudinal model: for the same baseline true mean value of 1.58, the MSE of the estimator is 0.11 in the cross-sectional model but 0.36 for the best longitudinal model ($J = 5$) at the same total sample size of 500. As the sample size increased by a factor of two, the MSE of the all estimators also decreases by a factor of two, as observed in both the longitudinal and cross-sectional simulation result Table 1 and 2. The auxiliary parameters' bias, standard error and coverage rate is displayed in the supplement material Table S1. The fixed effects for part II are estimated with high accuracy for all sample sizes. The bias and variance for threshold decreased as the sample size increased. The coverage rate for the threshold increased as sample size increased, but did not reach the nominal level due to the high correlation of the cubic B-spline density basis, similar to the longitudinal setting.

3.3 | Model sensitivity to B-spline density specification

Our model performed well when we know the true number of B-spline density basis functions as well as the true knot placement, but it is natural to expect that researchers do not know the true underlying B-spline density basis number and knot placement. In this section, we investigate the performance of our model under the situation when we do not know the number of splines and knot position of the underlying data model. Using the same data generating model as the cross-sectional setting where the number of B-spline densities $K = 5$, the three internal knot positions are equally spaced and the total sample size fixed at $N_{total} = 1000$, we ran misspecified models with $K = 7$ (number of equally spaced internal knots $K - 2 = 5$), $K = 10$ (number of internal knots 8) and $K = 15$ (number of internal knots 13). The auxiliary parameters $\boldsymbol{\alpha}$ for part II are incorrectly estimated with bias as large as 3.5 and coverage rate as low as 0 shown in supplement material Table S1. However as the model complexity increases (K increases), we do not see a trend in increasing bias and variance comparing to the correctly specified model for the covariate effect parameters of interest as shown in Table 3. Moreover, the coverage rate maintain the nominal 95% level regardless of the B-spline density basis number and internal knot placement. This means that the probability distributions of, when $X_1 = 0$ and $X_2 = 0$, when $X_1 = 1$ and $X_2 = 0$, when X_1

= 0 and X_2 = 1, are well estimated which is illustrated in Supplementary Materials Figure S2.

We also included longitudinal simulations using the same set of parameters as the cross-sectional setting (K = 5) with additional correlated part I and part II random effects with standard deviation 1.1 and 2, and a correlation of 0.2. With K fixed at 10 and knots equally spaced, the misspecified B-spline model still achieved decent coverage rate with the lowest coverage rate hovering around 90% when J = 5. As the number of observations per subject increase from J = 5 to J = 20, the variance slightly increased, but all the covariate effects' coverage rate converged to the nominal level. This demonstrated one of the most prominent advantages of modeling the data density semi-parametrically: the true model does not need to be known beforehand in order to obtain good estimates of densities and effects of interest.

3.4 Model comparison

One of the most flexible parametric models for part II is the *generalized gamma distribution*. Therefore, we compared our semi-parametric model to generalized gamma model. First, let's introduce the generalized gamma distribution. Let $\Gamma(\cdot)$ denote the standard gamma function. The density of the generalized gamma distribution is given by:

$$f(z;\kappa,\mu,\sigma)=\frac{\eta^{\eta}}{\sigma z\Gamma(\eta)\sqrt{\eta}}\exp\left[u\sqrt{\eta}-\eta\exp(|\kappa|u)\right]$$

with three parameters: κ for shape, μ for location, and σ for scale. We have $\eta = |\kappa|^{-2} > 0$ and $u = \mathrm{sign}(\kappa)(\log z - \mu)/\sigma$. If Z denotes a random variable with the generalized density, then its mean and variance are respectively given by

$$E(Z)=\exp\left\{\mu+\frac{\sigma\log\left(\kappa^{2}\right)}{\kappa}+\log\left[\Gamma\left(1/\kappa^{2}+\sigma/\kappa\right)\right]-\log\left[\Gamma\left(1/\kappa^{2}\right)\right]\right\}$$

and

$$\mathrm{Var}(Z)=\left\{\exp(\mu)\kappa^{2\sigma/\kappa}\right\}^{2}\left\{\frac{\Gamma\left(1/\kappa^{2}+2\sigma/\kappa\right)}{\Gamma\left(1/\kappa^{2}\right)}-\left[\frac{\Gamma\left(1/\kappa^{2}+2\sigma/\kappa\right)}{\Gamma\left(1/\kappa^{2}\right)}\right]^{-2}\right\}$$

The generalized gamma distribution comprehensively includes the standard gamma, inverse gamma, Weibull, and log-Normal distributions as its special cases. For example, if the scale parameter $\sigma = k$, it reduces to the standard gamma distribution with density

$$f(z;\upsilon,\eta)=\frac{1}{\upsilon^{\eta}\Gamma(\eta)}z^{\eta-1}\exp(-z/\nu)$$

where shape parameter $\eta = |\kappa|^{-2}$ and scale parameter $\upsilon = \kappa^2 \exp(\mu)$. Alternatively, taking the limit of generalized gamma density as $k \rightarrow 0$, one obtains

$$f(z;\mu,\sigma) = \frac{1}{\sigma z\sqrt{2\pi}}\exp\left\{-\frac{(\log(z)-\mu)^2}{2\sigma^2}\right\}$$

a log-Normal density function with log-mean μ and log-standard deviation σ. Finally, the inverse gamma distribution is obtained by setting $k = -\sigma$ with $\sigma > 0$, while the Weibull distribution is obtained by setting $k = 1$. Further detail and SAS code may be found in Liu's paper.[2,4]

3.4.1 | Comparison setup

We generated cross-sectional data and longitudinal data from the zero-inflated generalized gamma distribution and the zero-inflated B-spline mixture distribution with total sample size $N_{total = 2000}$. We ran the semi-parametric Bayesian model and zero-inflated generalized gamma model[2,4] regardless of whether it is the correct model for the underlying data.

For the generalized gamma distribution data, we have part I model (binary probit model with $\boldsymbol{\gamma} = (1,1.2,-1)$, covariates distribution $X_1 \sim N(0,1)$, $X_2 \sim Bernoulli(p = 0.5)$ and random intercept with standard deviation of 1.1 for longitudinal settings. For part II of the parametric model, we generated from the generalized gamma distribution with coefficients $\psi = (0.3,1,-0.5)$, $\zeta = (-0.3,-0.5,-0.2)$ and fixed $k = 0.7$, where μ and σ vary with the covariates:

$$\mu_i = \boldsymbol{x}_{\boldsymbol{i}}^T\boldsymbol{\psi} + b$$

$$\sigma_i^2 = \exp\left(\boldsymbol{x}_{\boldsymbol{i}}^T\zeta\right)$$

Note that the dependence of σ_i on the covariates allows for possibility of heteroscedasticity and b is the second part random intercept with standard deivation of 1 for longitudinal simulation scenarios. The baseline ($X_1 = 0_, X_2 = 0$) population mean, median and 75th quantile are 1.25, 0.87 and 1.73 respectively. The median and 75th quantile is calculated using Monto Carlo simulation since close form calculation is not possible for generalized gamma distribution. For every one unit increase in X_1 or X_2, the population mean will increase 2.53 or −0.81, respectively; the population median will increase 2.20 or −0.87, respectively; the population 75th quantile will increase 3.10 or −1.05, respectively.

For the mixture B-spline distribution data, part I model is the same as the generalized gamma model. For part II, we generated from $K = 10$ B-spline density bases equally spaced starting from 0.01 to 10.00 with the same part II coefficients as in the generalized gamma model $\boldsymbol{\alpha} = \boldsymbol{\psi} = (0.3,1,-0.5)$ and threshold $\boldsymbol{\lambda} = (0,0.1,0.15,0.2,0.22,0.25,0.3,1,1.5)$. Therefore, the baseline ($X_1 = 0_, X_2 = 0$) population mean, median and 75th quantile are 3.44, 0.76 and 7.06 respectively. For every one unit increase in X_1 or X_2, the population mean will increase 3.33 or −2.12, respectively; the population median will increase 7.02 or −0.76, respectively; the population 75th quantile will increase 1.56 or −6.65, respectively. This is an extremely right skewed distribution, with most of the data concentrated at the right tail of the distribution.

For fitting the Bayesian semi-parametric model to the generalized gamma data, since we do not know the correct knot placement positions, the placement of the knots are equally spaced or determined by quantiles of the observed simulated data. For example, if we have K number of B-spline densities, then we would place $K - 2$ internal knots at the $\frac{t}{K-1}$ for $t = 1, \ldots, K - 2$ quantiles of the observed positive valued data. We fitted the generalized gamma data with our semi-parametric Bayesian probit model with $K = 10,15,20,30$ to investigate the degree of underfitting or overfitting. Furthermore, we included a run with our multinomial logistic model for part II with $K = 15$ and equally spaced knots.

3.4.2 | Comparison result

For the cross-sectional positive data generated from the B-spline mixture distribution, the semi-parametric Bayesian model fitted this extremely right skewed data perfectly and obtained the correct inference for covariate mean, median and 75th quantile effect. On the other hand, the generalized gamma model failed to estimate the correct covariate mean effect with coverage rate as low as 1.23% and failed to converge in multiple instances. With poor performance even for cross-sectional B-spline data, the performance of gamma model for longitudinal B-spline data are expected to be worse and therefore omitted. This simulation demonstrated that our semi-parametric Bayesian model can be used to fit extremely right-skewed data while giving correct inference on any quantile of covariate effects, whereas the generalized gamma model cannot.

For cross-sectional positive data generated from the generalized gamma distribution, the generalized gamma model estimated the covariate mean effect with high accuracy and low variance as shown in Table 4. The generalized gamma model can not provide inference on the median nor the 75th quantile effects, since there is no closed-form representation for these quantiles. On the other hand, even without knowing the underlying distribution, our Bayesian semi-parametric model with knots placed at quantiles of the data or equally spaced did a good job in estimating covariate effects when the number of B-spline density functions was greater or equal to 20 using the ordinal probit model, with the coverage rate for the mean and 75th quantile effects at the 95% nominal level. For the median effect, the coverage rate was slightly lower, at around 80% coverage. As the number of B-spline density functions increased up to 30, the coverage rate of estimators began to decline slightly, indicating the model was starting to overfit the data, as shown in Table 4. However, the drop was quite small, demonstrating the robustness of the model to B-spline density number and knot placement. Furthermore, we compared the equally spaced knot ordinal probit model with equally spaced knot multinomial model with $K = 15$. The multinomial model had similar bias and higher variance comparing to the ordinal model. On the other hand, the multinomial model is more flexible and hence had higher coverage rate at around 95% for all three covariate effects. Another interesting observation is that the variances of the semi-parametric model were only slightly larger than the generalized gamma model.

For longitudinal data generated from the generalized gamma distribution, the bias decreased, and coverage increased as the number of observations per subject $\boldsymbol{J}$ increased for our Bayesian B-spline model, achieving nominal coverage rate for most of the parameters when number of observations per subject was at $\boldsymbol{J} = 20$, but the coverage rate for the median effect can be as low as around 60% when $\boldsymbol{J} = 5$. Due to the semi-parametric nature of our model and fitting to a parameter-rich generalized gamma data, we speculated that the individual random effects were estimated more accurately when the number of observations per subjects were large and thus leading to better estimation of the population effects.

4 | REAL DATA ANALYSIS

The National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA) was designed to disentangle the complex relationships between onset, escalation, and desistance of alcohol use and changes in neurocognitive functioning and neuromaturation.[34] We applied the proposed method to evaluate how number of binge drinking episodes in the past year changes with neuromaturation, using calendar age as our proxy. Number of binge drinking episodes is measured by the *Customary Drinking and Drug Use Record*,[35] administered on an annual basis, which asks: "During the past year, how many times have you consumed 4+ (females)/5+ (males) drinks within an occasion?" We considered three covariates: age, sex and education level. The analysis included $N = 820$ subjects aged from 12 to 26 at baseline (11 subjects without education level information were excluded). Each subject had at most 5 annual observations, with 3,261 total observations. Over 65% of the number of binge drinking episodes in the past year were equal at 0, but of the remainder there is large variation, with a maximum at 170. While number of binge drinking episodes is integer valued, we can approximate it as a continuous outcome because of the wide range of the values. From the spaghetti plot in Figure 1a, most of the subjects' binge drinking episodes in the past year remain below 25 regardless of age, but the overall average was high due to the extreme values.

We fitted our proposed model to the NCANDA data with $K = 5$ B-spline basis functions with equally-spaced knot placement (see Figure 2). Note, we also fitted the data with $K = 15$ B-spline densities (not shown), with no significant change in covariate effects. As seen in Figure 4a, the model yielded good fit for the overall sample at mean covariates values. The model-based distribution fitted the data well for smaller values of binge drinking episodes but somewhat under-fitted the data when the values were large. This is because the plot only showed the distribution where the covariates were fixed at the mean value, which did not take account of the random effects. The between-subject variation of the right end of the distribution is accounted for by the high variances of the random intercepts for part I (3.66, CI: 2.87 to 4.58) and part II (3.81, CI: 2.44 to 5.83).

Table 5 shows how the mean, median and 75th quantile of the population distribution changes with standardized age, sex and standardized education level. For the mean number of binge drinking episodes, the baseline value is 0.71 [95% Credible Interval (CI): 0.41 to 1.06], and it increases by 4.78 (95% CI: 4.09 to 5.52) as standardized age increases by one unit, and increases by 0.62 as standardized education increase by one unit. On the other hand, the baseline median number of binge drinking is 0 and only increases by 2.33

(95% CI: 0.59 to 3.75) for one unit increase of standardized age and increase by 0.00 (95% CI: 0.00 to 0.00) for one unit increase of standardized education. For heavy drinkers (75th quantile), who starts out with 0 number of binge drinking at baseline, will binge drink 8.80 (95% CI: 7.30 to 10.06) more as standardized age increases one unit but will not binge drink more as education level increased. The results from median and 75th quantile of the distribution suggested that heavy binge drinking subjects (75th quantile of the distribution) will have a large increase in binge drink number as age increases, whereas for over half of the population the increase is relatively minute. Moreover, while higher education levels increased the mean binge drinking number, for over 75% of the sample education level had no effect. Therefore, for zero-inflated and highly skewed data such as these, quantiles can be more informative summary statistics than the mean.

Table 6 shows the estimates and credible interval of the model parameters. As age increases, the probability of observing larger values increases. The part II coefficient for age is positive (1.18, CI: 0.89–1.49), indicating a higher probability of values generated from B-splines that are supported to the right of the distribution. For example, if we increase standardized age by 3 and fix other covariates at baseline (standardized age=0, sex=Female, standardized education=0), the part II latent variable mean will be $1.18 \times 3 - 3.71 = -0.17$ (the variance of the latent variable l for the ordinal model is fixed at 1). The probability of this latent variable reaching above $\lambda_1 = 0$ will increase significantly compared to the baseline, with the mean equal to −3.71. This leads to a larger probability of observing values from the second B-spline and lower probability of observing values from the first B-spline. This is illustrated in Figure 4b: when standardized age increases from 0 to 3, the probability density shifted to the right (the density region where the B-spline $k \geq 2$ is supported at increases, but the likelihood decreased for the region where $k = 1$ B-spline is supported).

5 | DISCUSSION

We proposed a semi-parametric Bayesian model for cluster correlated zero-inflated semi-continuous data. Instead of assuming a fixed family of distributions (e.g., Exponential Family)[2] or no form of distribution for response (e.g., generalized estimating equations),[16] we directly model the density of the positive part with mild assumptions for the number of B-spline densities and placement of internal knots that are robust to misspecification. Together with a Bayesian Gibbs sampler, we are able to model extremely skewed distributions that parametric models cannot fit and, moreover, obtain the posterior distribution of any function of the parameters. This is important, as obtaining the quantiles of the covariate effects allows researchers to more accurately interpret questions regarding covariate effects and individual variation, as illustrated in the simulations and real data example. Using the localized properties of B-spline density, researchers can use the model to intuitively quantify how covariates are related to the outcomes.

To reduce computational cost and improve estimation, we modeled the weights of the mixture using an ordinal probit model. To our knowledge, this is the first paper to formally introduce and investigate the algorithm of modelling the mixture B-spline density basis weights using an ordinal model. Another advantage of using the ordinal model is that we can easily penalize the coefficients by imposing a common prior for all the coefficients so that it can be fitted to high-dimensional data when the number of the features larger than the number of observations. On the other hand, we also provide the algorithm and code for estimating the B-spline density weights using a more commonly-applied multinomial logistic model.

Another advantage of our model is that it naturally handles heteroskedasticity. Here heteroskedasticity refers to non-constant variance of the outcome across different levels of the covariates. Our Bayesian semi-parametric model directly estimates the density (and hence the variance) of the outcome for each level of the covariates semi-parametrically. This is illustrated in Figure 3 and Figure S2, which show that the density of the specific levels of the covariates were estimated correctly. As a consequence, the non-constant variances of the outcome for different covariate sets were also correctly estimated.

In simulations, the semi-parametric Bayesian model was able to capture the covariate effects with high accuracy. Random effects variances were estimated well even when the number of observations per subjects were as low as 5. We also investigated the model performance when the number of B-spline density basis and the knot placement were misspecified in the cross-sectional setting. The estimated distribution for different covariate levels coincided with the true distribution as demonstrated in Figure S2 in the Supplementary Materials, giving good estimates for the covariate mean, median and 75th quantile effects. Finally, our semi-parametric model maintained decent performance even when the underlying data was generated from generalized gamma distribution. The interpretation of our semi-parametric model is more intuitive thanks to the localized support of the B-spline density basis and latent linear model, compared to the generalized gamma model where the mean covariate effects are dependent on the variance. Furthermore, if of scientific interest, a researcher could classify each observation into $\boldsymbol{K}$ categories according to the $\boldsymbol{K}$ locally supported B-spline bases using the posterior distribution of the corresponding B-spline density weights c_k.

Several issues deserve further investigation. A systematic approach to select the number B-spline density basis functions and internal knot position would be helpful to improve the performance of the model when the underlying true data distribution is not known. Although the covariate effect estimates were unbiased and reached the nominal coverage rate, the low coverage rate of the threshold parameters needs further investigation. Due to the nature of MCMC algorithm, the model requires more computational time compared to the maximum likelihood models. Using AMD Ryzen 3900x CPU with 3800MHZ computing power, it took on average 8.5 minutes to finish one instance on a dataset with 2000 sample size and 2 covariates. Methods such as MCMC subsampling[36] and variational inference[37] could be implemented to speed up the algorithm.

6 | SOFTWARE

Software in the form of R code, together with simulation data set is available on github: https://github.com/junting-ren/zero_inflated_b_spline.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors thank the investigators of NCANDA for providing the data. The full list of investigators are available at http://ncanda.org/. In addition, the authors would like to express sincere thanks to the constructive comments and suggestions given by the reviewers and editor. Funding for the project was provided by grant from National Institutes of Health(NIH) 5U24AA021695-09, 5U24AA021697-09, MH122688 and MH120025.

References

1. Shmueli G, Jank W, Hyde V. Transformations for semi-continuous data. Computational statistics & data analysis 2008; 52(8): 4000–4020.
2. Liu L, Strawderman RL, Johnson BA, O'Quigley JM. Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. Statistical Methods in Medical Research 2016; 25(1): 133–152. [PubMed: 22474003]
3. Karcher NR, Barch DM. The ABCD study: understanding the development of risk for mental and physical health outcomes. Neuropsychopharmacology 2021; 46(1): 131–142. [PubMed: 32541809]
4. Liu L, Strawderman RL, Cowen ME, Shih YCT. A flexible two-part random effects model for correlated medical costs. Journal of health economics 2010; 29(1): 110–123. [PubMed: 20015560]
5. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics 2016; 32(17): 2611–2617. [PubMed: 27187200]
6. Finak G, McDavid A, Yajima M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome biology 2015; 16(1): 1–13. [PubMed: 25583448]
7. Tobin J Estimation of relationships for limited dependent variables. Econometrica: journal of the Econometric Society 1958: 24–36.
8. Moulton LH, Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. Biometrics 1995: 1570–1578. [PubMed: 8589241]
9. Manning WG, Morris CN, Newhouse JP, et al. A two-part model of the demand for medical care: preliminary results from the health insurance study. Health, economics, and health economics 1981: 103–123.
10. Manning WG, Mullahy J. Estimating log models: to transform or not to transform?. Journal of health economics 2001; 20(4): 461–494. [PubMed: 11469231]
11. Ghosh P, Albert PS. A Bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial. Computational statistics & data analysis 2009; 53(3): 699–706. [PubMed: 19763231]
12. Smith VA, Preisser JS. A marginalized two-part model with heterogeneous variance for semicontinuous data. Statistical methods in medical research 2019; 28(5): 1412–1426. [PubMed: 29451088]
13. Chen J, Liu L, Johnson BA, O'Quigley J. Penalized likelihood estimation for semiparametric mixed models, with application to alcohol treatment research. Statistics in medicine 2013; 32(2): 335–346. [PubMed: 22833388]
14. Basu A, Rathouz PJ. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. Biostatistics 2005; 6(1): 93–109. [PubMed: 15618530]
15. Chen J, Liu L, Zhang D, Shih YCT. A flexible model for the mean and variance functions, with application to medical cost data. Statistics in medicine 2013; 32(24): 4306–4318. [PubMed: 23670952]
16. Chen J, Liu L, Shih YCT, Zhang D, Severini TA. A flexible model for correlated medical costs, with application to medical expenditure panel survey data. Statistics in medicine 2016; 35(6): 883–894. [PubMed: 26403805]

17. Liu L, Shih YCT, Strawderman RL, et al. Statistical analysis of zero-inflated nonnegative continuous data: a review. Statistical Science 2019; 34(2): 253–279.

18. Infante MA, Zhang Y, Brumback T, et al. Adolescent Binge Drinking is Associated with Accelerated Decline of Gray Matter Volume. bioRxiv 2021.

19. Bde Boor C A practical guide to splines, revised edition. 2001.

20. Lopes HF, Dias R. Bayesian mixture of parametric and nonparametric density estimation: A Misspecification Problem. Brazilian Review of Econometrics 2011; 31(1): 19–44.

21. Schellhase C, Kauermann G. Density estimation and comparison with a penalized mixture approach. Computational Statistics 2012; 27(4): 757–777.

22. Zablocki RW, Levine RA, Schork AJ, et al. Semiparametric covariate-modulated local false discovery rate for genome-wide association studies. The Annals of Applied Statistics 2017; 11(4): 2252–2269.

23. Su L, Tom BD, Farewell VT. Bias in 2-part mixed models for longitudinal semicontinuous data. Biostatistics 2009; 10(2): 374–389. [PubMed: 19136448]

24. Chib S, Carlin BP. On MCMC sampling in hierarchical longitudinal models. Statistics and Computing 1999; 9(1): 17–26.

25. De Boor C, De Boor C, Mathématicien EU, De Boor C, De Boor C. A practical guide to splines. 27. Springer-Verlag New York. 1978.

26. Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. BMC medical research methodology 2019; 19(1): 1–16. [PubMed: 30611213]

27. Ruppert D Selecting the number of knots for penalized splines. Journal of computational and graphical statistics 2002; 11(4).

28. Thompson WK, Rosen O. A Bayesian model for sparse functional data. Biometrics 2008; 64(1): 54–63. [PubMed: 17573864]

29. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. Journal of the American statistical Association 1993; 88(422): 669–679.

30. Yi N, Banerjee S, Pomp D, Yandell BS. Bayesian mapping of genomewide interacting quantitative trait loci for ordinal traits. Genetics 2007; 176(3): 1855–1864. [PubMed: 17507680]

31. Bi W, Zhou W, Dey R, Mukherjee B, Sampson JN, Lee S. Efficient mixed model approach for large-scale genome-wide association studies of ordinal categorical phenotypes. The American Journal of Human Genetics 2021; 108(5): 825–839. [PubMed: 33836139]

32. Montesinos-López OA, Montesinos-López A, Crossa J, Burgueño J, Eskridge K. Genomic-enabled prediction of ordinal data with Bayesian logistic ordinal regression. G3: Genes, Genomes, Genetics 2015; 5(10): 2113–2126. [PubMed: 26290569]

33. Lavielle M, Aarons L. What do we mean by identifiability in mixed effects models?. Journal of pharmacokinetics and pharmacodynamics 2016; 43(1): 111–122. [PubMed: 26660913]

34. Brown SA, Brumback T, Tomlinson K, et al. The National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA): a multisite study of adolescent development and substance use. Journal of studies on alcohol and drugs 2015; 76(6): 895–908. [PubMed: 26562597]

35. Brown SA, Myers MG, Lippke L, Tapert SF, Stewart DG, Vik PW. Psychometric evaluation of the Customary Drinking and Drug Use Record (CDDR): a measure of adolescent alcohol and drug involvement. Journal of studies on alcohol 1998; 59(4): 427–438. [PubMed: 9647425]

36. Quiroz M, Kohn R, Villani M, Tran MN. Speeding up MCMC by efficient data subsampling. Journal of the American Statistical Association 2018.

37. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. Journal of the American statistical Association 2017; 112(518): 859–877.

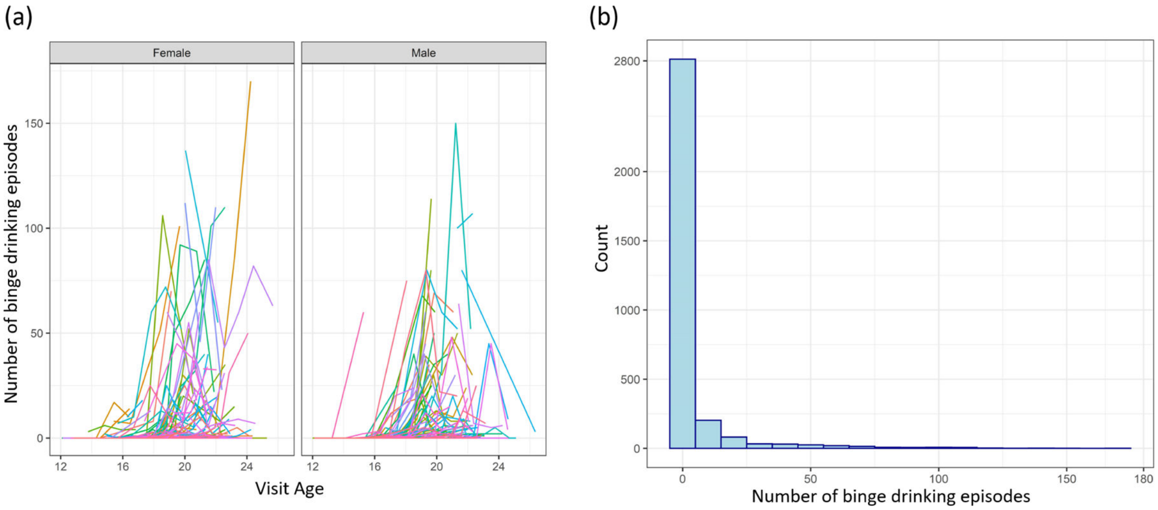
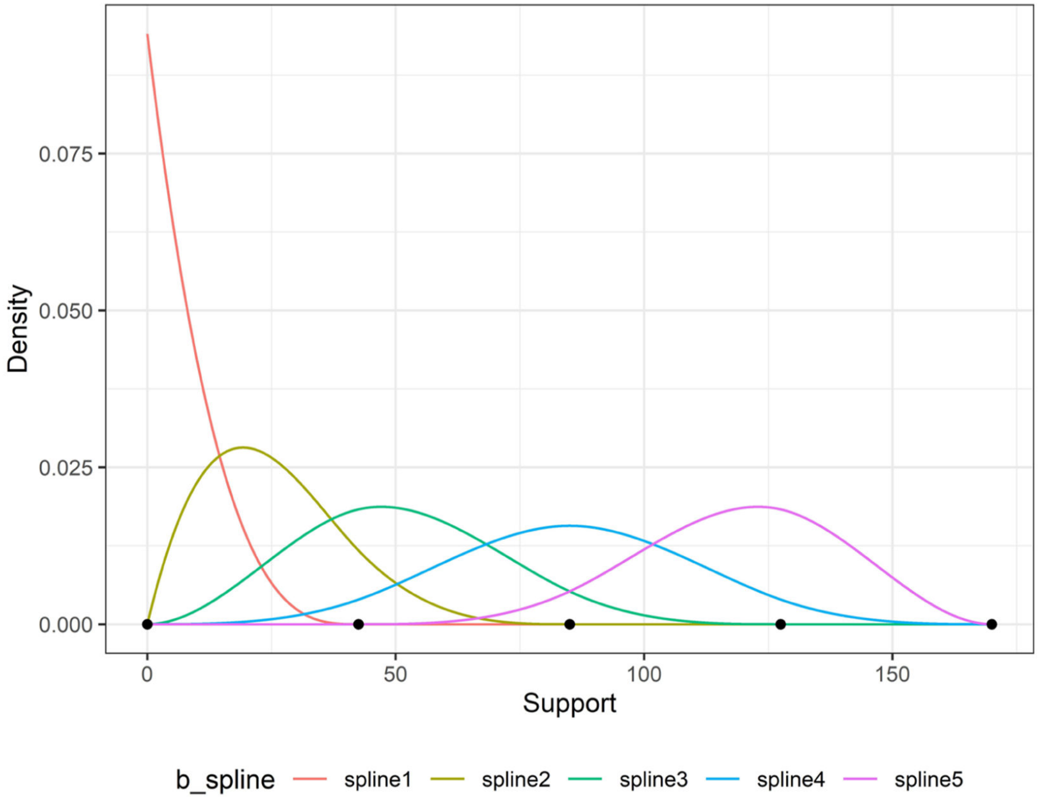


FIGURE 1.

NCANDA data descriptive plots: (a) Spaghetti plot of binge drinking episode number VS visit age for randomly selected 400 subjects and (b) Histogram for the number of binge drinking episodes for all subjects.



Spline	Start	Mode	End
k = 1	0.01	0.01	42.50
k = 2	0.02	19.27	85.00
k = 3	0.02	47.11	127.50
k = 4	0.02	85.01	170.00
k = 5	42.52	122.91	170.00

FIGURE 2.

B-spline density used in the NCANDA data model and their corresponding supports and modes. The black points on the x-axis are the corresponding two boundary knots and three internal knots.

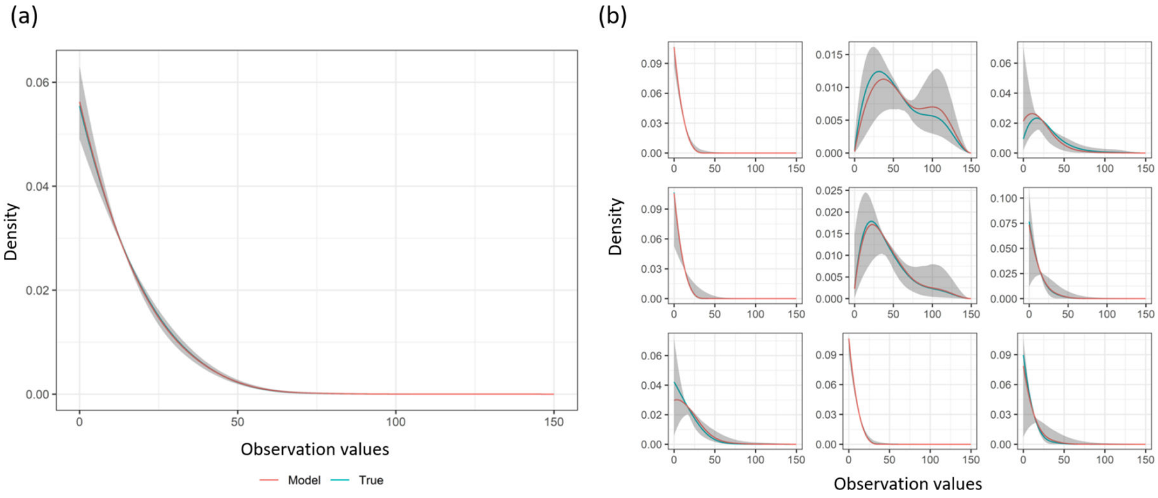


FIGURE 3.

Model fit when $N_{total} = 2000$ and $\boldsymbol{J} = 20$ in one realization for (a) Population positive distribution comparing model to true density when $X_1 = 1$, $X_2 = 1$; (b) Individual positive distribution for randomly selected 9 subjects taking random intercepts into account. The red line is the model estimate, 95% credible interval in gray, and the blue line is the true density.

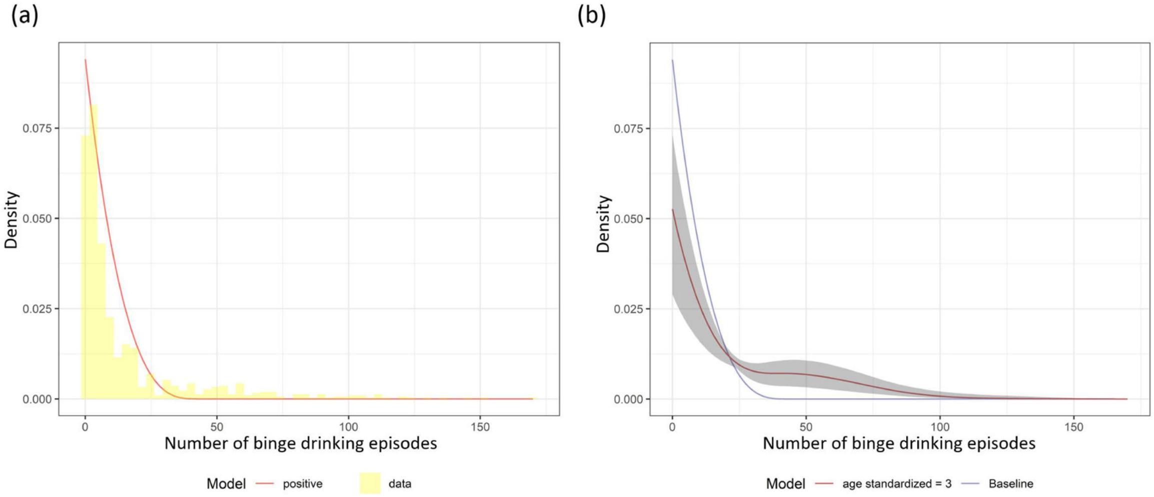


FIGURE 4.

NCANDA data analyses: model fit for (a) Population distribution at mean covariates values for the positive part. The yellow histograms are densities for the data, whereas gray shaded areas are the 95% credible band. (b) Population positive distribution comparing baseline to when standardized age increased 3 units. The gray shaded areas are the 95% credible band.

TABLE 1

Longitudinal simulation results with 500 replications. Mean, median and 75th quantile of the distribution of Z at the baseline X_1 = 0, X_2 = 0, and the covariate effects comparing to the baseline. True value for baseline, X_1 effect and X_2 effect are: 1.58, 5.09 and 4.89 for the mean; 0.00, 1.84 and 0.00 for the median; 0.00, 10.01 and 9.48 for 75th quantile, respectively. SD: standard deviation. MSE: mean square error. CR: credible interval coverage rate of the true parameter value.

	Mean			Median			75th quantile		
	Bias (SD)	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**
Baseline (X_1 = 0, X_2 = 0)									
N_{total} = 500									
J=5*	0.27(0.57)	0.40	94.9	0.00(0.00)	0.00	100.0	0.16(0.34)	0.14	100.0
J=5	0.24(0.55)	0.36	94.6	0.00(0.00)	0.00	100.0	0.14(0.26)	0.09	100.0
J=10	0.32(0.61)	0.48	95.7	0.01(0.01)	0.01	100.0	0.27(0.48)	0.30	100.0
J=20	0.51(0.83)	0.95	95.1	0.01(0.01)	0.01	100.0	0.63(0.92)	1.24	99.6
N_{total} = 1000									
J=5*	0.12(0.40)	0.18	93.6	0.00(0.00)		100.0	0.03(0.06)	0.01	100.0
J=5	0.12(0.40)	0.18	95.5	0.00(0.00)	0.00	100.0	0.02(0.06)	0.01	100.0
J=10	0.15(0.43)	0.20	96.4	0.00(0.00)	0.00	100.0	0.07(0.16)	0.03	100.0
J=20	0.19(0.53)	0.32	95.4	0.01 (0.01)	0.01	100.0	0.16 (0.31)	0.12	100.0
N_{total} = 2000									
J=5*	0.05(0.28)	0.08	95.6	0.00(0.00)		100.0	0.01(0.01)	0.00	100.0
J=5	0.06(0.28)	0.08	94.0	0.00(0.00)	0.00	100.0	0.01(0.01)	0.00	100.0
J=10	0.09(0.32)	0.11	94.2	0.00(0.00)	0.00	100.0	0.01(0.03)	0.01	100.0
J=20	0.14(0.40)	0.17	93.2	0.00(0.00)	0.00	100.0	0.05(0.14)	0.02	100.0
One unit increase in X_1									
N_{total} = 500									
J=5*	0.34(1.25)	1.69	93.6	0.59(1.49)	2.57	95.3	0.96(2.63)	7.86	95.3
J=5	0.34(1.26)	1.70	94.4	0.58(1.41)	2.31	94.8	0.93(2.66)	7.95	94.4
J=10	0.51(1.25)	1.83	96.1	0.71(1.59)	3.05	94.5	1.13(2.77)	8.96	97.5
J=20	0.69(1.51)	2.75	96.1	1.19(1.98)	5.33	93.9	1.33(3.31)	12.71	96.7
N_{total} = 1000									
J=5*	0.11(0.95)	0.91	93.2	0.28(1.09)	1.27	92.4	0.45(2.21)	5.08	94.0
J=5	0.19(0.91)	0.86	95.9	0.21(1.04)	1.12	94.5	0.59(2.00)	4.34	95.3
J=10	0.19(0.92)	0.89	95.6	0.27(1.15)	1.40	93.0	0.56(2.02)	4.42	96.8
J=20	0.24(1.01)	1.08	96.0	0.40(1.32)	1.91	95.2	0.61(2.27)	5.53	95.6
N_{total} = 2000									
J=5*	0.07(0.67)	0.46	92.4	0.06(0.81)	0.65	94.8	0.26(1.44)	2.16	92.0
J=5	0.04(0.67)	0.45	92.6	0.02(0.81)	0.66	92.8	0.22(1.42)	2.07	93.0
J=10	0.07(0.69)	0.49	95.4	0.11(0.88)	0.79	94.8	0.32(1.55)	2.49	94.4
J=20	0.16(0.79)	0.64	94.4	0.28(1.05)	1.18	95.0	0.51(1.81)	3.55	94.0

	Mean			**Median**			**75th quantile**		
	Bias (SD)	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**
One unit increase in X_2									
$N_{total} = 500$									
J=5[*]	0.18(1.36)	1.89	95.5	1.13(1.12)	2.54	97.7	0.72(2.88)	8.81	95.7
J=5	0.15(1.36)	1.87	95.8	1.08(1.05)	2.28	97.7	0.62(2.93)	8.95	95.8
J=10	0.35(1.51)	2.40	92.5	1.38(1.32)	3.66	97.8	0.88(3.32)	11.77	94.7
J=20	0.44(1.62)	2.81	96.5	1.84(1.67)	6.20	96.7	0.89(3.66)	14.16	96.5
$N_{total} = 1000$									
J=5[*]	0.07(1.01)	1.02	94.4	0.79(0.72)	1.15	97.2	0.45(2.21)	5.08	94.0
J=5	0.09(0.98)	0.97	95.5	0.69(0.69)	0.95	98.2	0.43(2.15)	4.82	94.1
J=10	0.19(1.06)	1.15	93.4	0.93(0.91)	1.40	96.0	0.62(2.32)	5.79	95.0
J=20	0.13(1.16)	1.36	94.0	1.15(1.12)	2.57	96.0	0.38(2.72)	7.55	93.4
$N_{total} = 2000$									
J=5[*]	0.06(0.70)	0.49	95.4	0.53(0.51)	0.55	97.2	0.26(1.56)	2.50	96.0
J=5	0.01(0.68)	0.46	94.4	0.50(0.49)	0.49	97.0	0.19(1.48)	2.22	94.8
J=10	0.07(0.77)	0.59	94.0	0.64(0.60)	0.77	96.4	0.33(1.73)	3.11	93.0
J=20	0.11(0.84)	0.72	95.0	0.85(0.79)	1.34	97.4	0.46(2.03)	4.32	94.6

[*] Only for the rows with “*”, the correlation between the part I and part II random intercept is 0.6. For other rows without “*”, the correlations are fixed at 0.2.

TABLE 2

Cross-sectional simulation results with 500 replications. Mean, median and 75th quantile of the distribution of Z at the baseline $X_1 = 0$, $X_2 = 0$, and the covariate effects comparing to baseline. True value for baseline, X_1 effect and X_2 effect are: 1.58, 11.69 and 8.86 for the mean; 0.00, 5.54 and 0.00 for the median; 0.00, 22.21 and 17.42 for 75th quantile, respectively. SD: standard deviation. MSE: mean square error. CR: credible interval coverage rate of the true parameter value.

	Mean			**Median**			**75th quantile**		
	Bias (SD)	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**
Baseline ($X_1 = 0, X_2 = 0$)									
$N_{total} = 500$	0.07(0.33)	0.11	94.2	0.00(0.00)	0.00	100.0	0.01(0.04)	0.01	100.0
$N_{total} = 1000$	0.04(0.22)	0.05	96.2	0.00(0.00)	0.00	100.0	0.01(0.01)	0.01	100.0
$N_{total} = 2000$	0.02(0.16)	0.03	95.4	0.00(0.00)	0.00	100.0	0.01(0.01)	0.01	100.0
$N_{total} = 4000$	0.01(0.12)	0.01	94.0	0.00(0.00)	0.00	100.0	0.00(0.00)	0.00	100.0
One unit increase in X_1									
$N_{total} = 500$	0.36(1.48)	2.32	95.2	0.37(2.69)	7.35	94.4	0.46(2.60)	6.99	95.2
$N_{total} = 1000$	0.12(1.05)	1.12	95.0	0.09(2.00)	4.00	96.2	0.08(1.84)	3.40	95.6
$N_{total} = 2000$	0.13(0.75)	0.58	95.2	0.15(1.49)	2.24	96.0	0.16(1.32)	1.76	95.0
$N_{total} = 4000$	0.05(0.51)	0.26	95.2	−0.03(1.02)	1.03	94.8	0.05(0.91)	0.83	94.8
One unit increase in X_2									
$N_{total} = 500$	0.06(1.17)	1.36	95.8	1.37(1.25)	3.43	98.2	0.06(2.36)	5.59	94.8
$N_{total} = 1000$	−0.05(0.84)	0.72	93.8	0.94(0.92)	1.74	97.6	−0.15(1.70)	2.93	93.0
$N_{total} = 2000$	0.03(0.58)	0.34	94.2	0.71(0.67)	0.95	97.2	0.05(1.17)	1.37	95.0
$N_{total} = 4000$	−0.01(0.39)	0.16	95.2	0.48(0.47)	0.45	96.8	−0.04(0.82)	0.68	96.0

TABLE 3

Spline misspecification simulation results with 500 replications for fixed total cross-sectional sample size of 1000. True value for baseline, X_1 effect and X_2 effect are: 1.58, 11.69 and 8.86 for the mean; 0.00, 5.54 and 0.00 for the median; 0.00, 22.21 and 17.42 for 75th quantile, respectively. SD: standard deviation. MSE: mean square error. CR: credible interval coverage^2 rate of the true parameter value.

	Mean			**Median**			**75th quantile**		
	Bias (SD)	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**
Baseline ($X_1 = 0$, $X_2 = 0$)									
$K = 5$[*]	0.04(0.22)	0.05	96.2	0.00(0.00)	0.00	100.0	0.01(0.01)	0.01	100.0
$K = 7$	−0.03(0.25)	0.06	92.6	0.00(0.00)	0.00	100.0	0.01(0.01)	0.01	100.0
$K = 10$	−0.04(0.26)	0.07	92.4	0.00(0.00)	0.00	100.0	0.01(0.01)	0.01	100.0
$K = 10$[†]	−0.05(0.26)	0.07	91.2	0.00(0.00)	0.00	100.0	0.01(0.01)	0.01	100.0
$K = 15$	−0.01(0.23)	0.06	93.6	0.00(0.00)	0.00	100.0	0.01(0.01)	0.01	100.0
$K = 10$, $J = 5$	−0.42(0.48)	0.42	87.4	0.00(0.00)	0.00	100.0	0.03(0.09)	0.01	100.0
$K = 10$, $J = 10$	0.40(0.53)	0.44	88.8	0.00(0.00)	0.00	100.0	0.08(0.19)	0.04	100.0
$K = 10$, $J = 20$	0.50(0.64)	0.66	91.2	0.00(0.00)	0.00	100.0	0.24(0.44)	0.25	100.0
One unit increase in X_1									
$K = 5$[*]	0.12(1.05)	1.12	95.0	0.09(2.00)	4.00	96.2	0.08(1.84)	3.40	95.6
$K = 7$	0.48(1.06)	1.37	93.2	0.32(1.99)	4.07	93.2	0.15(2.06)	4.25	95.0
$K = 10$	0.61(1.12)	1.63	90.6	0.27(2.09)	4.43	93.2	0.04(2.13)	4.55	94.4
$K = 10$[†]	−0.65(1.12)	1.64	90.6	0.24(2.00)	4.05	94.8	0.10(2.10)	4.42	93.6
$K = 15$	0.38(1.16)	1.47	92.0	0.37(1.67)	2.92	95.4	−0.22(2.13)	4.61	90.8
$K = 10$, $J = 5$	1.65(2.02)	6.80	87.2	0.96(2.81)	8.81	92.4	1.41(3.74)	15.95	92.6
$K = 10$, $J = 10$	1.50(2.18)	7.00	89.0	0.84(3.16)	10.69	91.8	1.07(4.17)	18.54	94.0
$K = 10$, $J = 20$	1.64(2.47)	8.79	91.2	1.04(3.49)	13.29	96.0	1.25(4.65)	23.22	94.2
One unit increase in X_2									
$K = 5$[*]	−0.05(0.84)	0.72	93.8	0.94(0.92)	1.74	97.6	−0.15(1.70)	2.93	93.0
$K = 7$	0.19(0.88)	0.81	93.6	1.06(0.97)	2.06	98.2	−0.18(1.74)	3.05	93.4
$K = 10$	0.39(0.89)	0.95	91.4	1.19(1.06)	2.56	97.0	−0.16(1.70)	2.91	95.0
$K = 10$[†]	0.40(0.88)	0.93	93.0	1.08(1.00)	2.06	98.4	−0.18(1.69)	2.89	94.8
$K = 15$	0.24(0.89)	0.85	94.4	1.43(1.18)	3.42	97.8	−0.35(1.82)	3.42	94.0
$K = 10$, $J = 5$	1.10(1.54)	3.58	90.4	1.98(1.64)	6.60	98.6	0.78(2.96)	9.40	94.6
$K = 10$, $J = 10$	1.12(1.84)	4.61	90.8	2.28(2.07)	9.47	97.4	0.62(3.69)	13.99	94.8
$K = 10$, $J = 20$	1.18(2.02)	5.48	92.6	2.82(2.36)	13.55	97.6	0.58(4.13)	17.42	95.4

[*] The correct model (same number of B-splines and knot positions as the data generating model).

[†] Knot placements are determined by quantiles of the sampled simulationdata. All other knot placements are equally spaced.

TABLE 4

Model comparison results with 500 replications for fixed total cross-sectional sample size of 2000. GG:gamma data, gamma model; GB: gamma data, B-spline model, ***J*** indicates the number of observations per subject for the longitudinal data; BG: B-spline data, gamma model; BB: B-spline data, B-spline model with correct number of B-spline densities and knot placement; SD: standard deviation. MSE: mean square error. CR: credible interval coverage rate of the true parameter value.

	Mean			**Median**			**75th quantile**		
	Bias (SD)	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**
Baseline (X_1 = 0, X_2 = 0)									
GG	0.00(0.04)	0.00	95.6	-	-	-	-	-	-
GB10*	0.05(0.05)	0.01	81.0	0.05(0.05)	0.01	74.6	0.01(0.08)	0.01	91.0
GB20*	0.01(0.05)	0.00	95.8	0.06(0.05)	0.01	76.0	0.00(0.08)	0.01	91.6
GB30*	0.00(0.05)	0.00	94.2	0.06(0.05)	0.01	74.1	0.01(0.08)	0.01	92.4
GB15†	0.03(0.06)	0.01	93.8	0.01(0.06)	0.00	92.2	0.02(0.10)	0.01	93.4
GB15	−0.02(0.05)	0.00	87.6	0.04(0.05)	0.01	82.2	−0.02(0.08)	0.01	89.8
GB20	−0.01(0.05)	0.00	91.2	0.05(0.05)	0.01	79.6	−0.02(0.07)	0.01	90.8
GB30	0.00(0.05)	0.00	94.6	0.05(0.05)	0.01	78.6	−0.01(0.08)	0.01	92.6
GG, J=5	0.00(0.09)	0.01	93.4	-	-	-	-	-	-
GB30, J=5	−0.05(0.09)	0.01	87.6	−0.09(0.08)	0.01	70.2	−0.02(0.14)	0.02	88.4
GB30, J=10	−0.03(0.11)	0.01	89.6	−0.07(0.11)	0.02	79.8	0.02(0.18)	0.03	91.4
GB30, J=20	−0.01(0.14)	0.02	92.8	−0.05(0.13)	0.02	91.2	0.06(0.21)	0.05	94.0
BG	−1.74(2.58)	9.71	15.6	-	-	-	-	-	-
BB	−0.01(0.14)	0.02	94.0	0.33(0.55)	0.41	95.6	−0.03(0.11)	0.01	93.2
One unit increase in X_1									
GG	−0.01(0.08)	0.01	95.8	-	-	-	-	-	-
GB10*	0.03(0.09)	0.01	95.6	−0.13(0.11)	0.03	80.4	0.33(0.14)	0.13	38.2
GB20*	−0.04(0.09)	0.01	94.0	−0.06(0.11)	0.02	89.0	−0.09(0.16)	0.04	93.2
GB30*	−0.07(0.09)	0.01	89.0	−0.06(0.11)	0.02	89.0	−0.08(0.17)	0.03	91.4
GB15†	0.04(0.15)	0.02	95.2	−0.01(0.17)	0.03	93.8	0.01(0.26)	0.07	95.2
GB15	0.04(0.09)	0.01	92.6	0.02(0.12)	0.01	93.4	0.01(0.17)	0.03	92.4
GB20	0.03(0.09)	0.01	93.4	0.01(0.11)	0.01	94.2	−0.02(0.17)	0.03	94.4
GB30	0.01(0.09)	0.01	95.8	−0.04(0.11)	0.01	93.6	−0.06(0.16)	0.03	93.0
GG, J=5	0.00(0.17)	0.03	93.0	-	-	-	-	-	-
GB30, J=5	−0.09(0.20)	0.05	90.4	−0.10(0.22)	0.06	87.8	−0.09(0.32)	0.11	87.4
GB30, J=10	−0.11(0.24)	0.07	90.2	−0.12(0.23)	0.07	86.6	−0.19(0.34)	0.15	87.6
GB30, J=20	−0.07(0.27)	0.08	93.6	−0.08(0.25)	0.07	93.4	−0.14(0.36)	0.14	91.0
BG	−3.37(0.32)	11.5	1.23	-	-	-	-	-	-
BB	0.01(0.12)	0.02	94.0	−0.35(0.53)	0.40	92.2	0.02(0.13)	0.01	90.4
One unit increase in X_2									
GG	0.00(0.04)	0.00	95.2	-	-	-	-	-	-

	Mean			Median			75th quantile		
	Bias (SD)	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**	**Bias (SD)**	**MSE**	**CR**
GB10[*]	−0.04(0.06)	0.01	89.2	−0.01(0.06)	0.01	93.8	−0.03(0.09)	0.01	90.4
GB20[*]	0.00(0.05)	0.01	94.0	−0.02(0.06)	0.01	95.0	−0.03(0.09)	0.01	91.8
GB30[*]	0.01(0.05)	0.01	93.2	−0.04(0.06)	0.01	87.4	−0.04(0.09)	0.01	90.2
GB15[+]	0.03(0.06)	0.00	91.6	0.03(0.07)	0.01	92.0	0.06(0.11)	0.02	90.4
GB15	0.03(0.05)	0.00	86.4	−0.02(0.06)	0.01	92.6	0.02(0.08)	0.01	90.8
GB20	−0.03(0.05)	0.00	90.4	−0.01(0.06)	0.01	93.4	0.01(0.08)	0.01	93.6
GB30	0.02(0.05)	0.00	93.2	−0.01(0.06)	0.01	94.4	−0.01(0.09)	0.01	92.6
GG, J=5	0.00(0.07)	0.01	93.2	-	-	-	-	-	-
GB30, J=5	0.14(0.10)	0.03	67.0	0.13(0.08)	0.02	57.0	0.11(0.17)	0.04	74.0
GB30, J=10	0.09(0.10)	0.02	78.4	0.10(0.10)	0.02	75.2	0.01(0.18)	0.03	88.8
GB30, J=20	0.07(0.11)	0.02	89.2	0.10(0.11)	0.02	85.2	−0.02(0.19)	0.04	93.8
BG	1.54(1.73)	5.37	9.57	-	-	-	-	-	-
BB	0.01(0.16)	0.03	93.0	−0.32(0.55)	0.41	95.4	0.05(0.13)	0.02	92.2

[*] Knots determined by quantiles

[+] Equqally spaced knots with multinomial logistic regression for part II

TABLE 5

Distribution mean, median and 75th quantile of the number of binge drinking episodes at baseline or comparing to baseline at different age and sex. CI: 95% credible interval.

Covariate	Mean		Median		75th quantile	
	Estimate	CI	Estimate	CI	Estimate	CI
Baseline *	0.71	(0.41, 1.06)	0.00	(0.00,0.00)	0.00	(0.00, 0.00)
Effect of age **	4.78	(4.09, 5.52)	2.33	(0.59, 3.75)	8.80	(7.30, 10.06)
Effect of sex **	0.02	(−0.40, 0.45)	0.00	(0.00,0.00)	0.00	(0.00, 0.00)
Effect of education **	0.62	(0.28, 1.03)	0.00	(0.00,0.00)	0.00	(0.00, 0.00)

*Standardized age=0, sex=Female, standardized education=0

** One unit increase in standardized age or standardized education level or switching to male, comparing to baseline.

TABLE 6

Posterior mean estimate and credible interval of model parameters for part I fixed effect coefficients ($\boldsymbol{\gamma}$), part II fixed effect coefficients ($\boldsymbol{\alpha}$) and threshold parameters (λ) of the NCANDA data model. Part I random intercept posterior mean estimate is 3.66 (CI: 2.87, 4.58) and part II random intercept estimate is 3.81 (CI: 2.44, 5.83).

Covariate	$\boldsymbol{\gamma}$	$\boldsymbol{\alpha}$	$\boldsymbol{\lambda}_2$	$\boldsymbol{\lambda}_3$	$\boldsymbol{\lambda}_4$
Intercept	−1.39 (−1.66, −1.15)	−3.71 (−4.50, −3.00)	0.23 (0.03, 0.53)	1.80 (1.34, 2.27)	2.51 (1.96, 3.41)
Age	1.72 (1.55, 1.89)	1.18 (0.89, 1.49)	-	-	-
Sex	0.02 (−0.30, 0.35)	0.20 (−0.28, 0.68)	-	-	-
Education	0.37 (0.21, 0.54)	0.22 (0.48, −0.03)	-	-	-