REVSTAT – Statistical Journal
Volume 0, Number 0, Month 0000, 000-000
https://doi.org/00.00000/revstat.v00i0.000

Bayesian Variable Selection for Zero-inflated Longitudinal Count Data

Authors: Nawar Alsalim
– Department of Statistics, Faculty of Sciences, Al-Baath University, Homs, Syria

– Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran.
Nawjimaa@gmail.com

Taban Baghfalaki ✉
– Inserm, Research Center U1219, Univ. Bordeaux, ISPED, F33076 Bordeaux, France.
taban.baghfalaki@u-bordeaux.fr

Received: Month 0000 Revised: Month 0000 Accepted: Month 0000

Abstract:

- In this paper, we present a Bayesian variable selection method for zero-inflated longitudinal data. For this purpose, we consider a zero-inflated power series random effects model that includes the zero-inflated Poisson and negative binomial random effects models. We propose using continuous spike and Dirac spike priors to simultaneously estimate the regression coefficients and select the important covariate variables. We apply the MCMC method using Gibbs sampling for posterior inference. Some simulation studies are performed to investigate the performance of the proposed approach, and it is also applied to analyze a real dataset from the RAND Health Insurance Experiment.

Keywords:

- *Bayesian variable selection; Continuous spike; Dirac spike; Longitudinal data; Power series family; Random effects models.*

AMS Subject Classification:

- 62F15, 62J99.

✉ Corresponding author

1. INTRODUCTION

In medical research, count variables with many zeros are pervasive. Models that deal with and analyze a high proportion of zeros are known as zero-inflated models. [23] and [24] performed Hurdle models for modeling zero-inflated data. Also, the zero-inflated negative binomial (ZINB) regression is used for count data that exhibit overdispersion and excess zeros. In general, zero-inflated power series distributions are applied to assess the excess of zeros [30]. [20] have studied zero-inflated modified power series distributions along with their applications for simulated data. [33] considered non-zero-inflated modified power series distributions and extended the results of [20]. [35] discussed a Bayesian paradigm for the ZIP and ZINB models for analyzing the data set of a study of psychiatric outpatient services. [38] discussed the application of the ZI and Hurdle models for longitudinal studies concerning vaccination safety. [36] studied different aspects of the zero-inflated power series distributions. [5] considered the ZIP, ZINB, and Hurdle models, to observe whether there is any effect of the proportion of zeros in the performance of the models with the given overall rate of the counts. [12] reviewed the zero-inflated and hurdle models and highlighted their differences in terms of their data-generating processes. [43] surveyed the developments in handling zero inflation for correlated count settings. [3] discussed the approximate Bayesian approach for zero-inflated longitudinal models.

In longitudinal studies, data are collected repeatedly for the same set of units on more than one occasion. Random effect models have often been used in longitudinal data analysis since they allow for association among repeated measurements due to unobserved heterogeneity. To take into account the correlation among repeated measurements for each subject, zero-inflated count models with random effects have been developed. For example, a random effect was used to account for the within-subject dependency in the Poisson part of the ZIP model [21]. [34] proposed a random effect model to analyze the ZI longitudinal count data. [28] incorporated shared subject-specific random effects in each part of the zero-inflated model to account for zero-inflation and overdispersion within longitudinal count measurements.

Variable selection is an essential part of regression modeling for longitudinal data because many variables are measured and it is common in practice to include only a subset of important variables in the model. [44] discussed a variable selection approach for zero-inflated count data analysis based on the adaptive lasso technique. [31] assumed that the regression coefficients were mutually independent with a two-point mixture distribution made up of a uniform flat distribution (the slab) and a degenerate distribution at zero (the spike). [13] used a different prior for the regression coefficients. This involved a scale (variance) mixture of two normal distributions. In particular, the use of a normal prior was instrumental in facilitating efficient Gibbs sampling of the posterior. This made spike and slab variable selection computationally attractive and heavily popularized the method. Normal-scale mixture priors constitute a wide class

of models termed spike and slab models [25]. Spike and slab models are extended to the class of re-scaled spike and slab models [25]. [29] developed a Bayesian variable selection model for multivariate count data with excess zeros that incorporates information on the covariance structure of the outcomes. [14] introduced the basic concepts of the Bayesian approach for variable selection based on model choice.[42] studied some of the classic and contemporary literature on parametric zero-inflated count regression models. [15] studied Bayesian variable selection in high dimensional data sets while simultaneously accounting for the error-prone nature of self-reported outcomes. [26] presented a Bayesian analysis of linear mixed models for quantile regression based on a Cholesky decomposition of the covariance matrix of random effects. [32] considered linear regression models for count data, specifically negative binomial regression models and Dirichlet-multinomial regression models, they also addressed variable selection criteria via the use of spike-and-slab priors on the regression coefficients. [1] discussed new variable selection methods for the power series, specifically ZIP and ZINB transition models using LASSO, MCP, and SCAD penalties for analyzing longitudinal count data with extra zeros.

In this paper, we focus primarily on Bayesian approaches for variable selection that use spike and slab priors in the zero-inflated power series (ZIPS) model. For posterior inference, we apply MCMC methods via Gibbs sampling, and a test for variable selection of regression coefficients is considered by using both a local Bayesian false discovery rate and a Bayes factor procedure. After checking the performance of the proposed model using some simulation studies, we apply the proposed method to analyze the RAND health insurance experiment data.

This paper is organized as follows. Section 2 is a review of ZIPS distributions and the use of these distributions for analyzing zero-inflated longitudinal data. Also, this section includes some notation and definitions of models. Section 3 includes the likelihood functions, the Bayesian variable selection method, using spike and slab priors, and the test for variable selection of the regression coefficients. In Section 4, some simulation studies are performed. In Section 5, after describing the RAND health insurance experiment data, the data is analyzed using the proposed approaches. The last Section includes some conclusions.

2. MATERIALS AND METHODS

2.1. Notation

Let $Y_{ij}, i = 1, \cdots, n$ and $j = 1, \cdots, T$ be the longitudinal measurements for the ith individual at the jth time point. The ZIPS model is given as follows:

$$P(Y_{ij} = y_{ij}|\mu_{ij}, \pi_{ij}) = \begin{cases} (1-\pi_{ij})p(Y_{ij} = y_{ij}|\mu_{ij}) & y_{ij} = 1, 2, \cdots \\ \pi_{ij} + (1-\pi_{ij})p(Y_{ij} = 0|\mu_{ij}) & y_{ij} = 0. \end{cases} \tag{2.1}$$

where $p(Y_{ij} = y_{ij}|\mu_{ij}, \pi_{ij})$ is a member of the power series (PS) family with the general form of

$$\frac{b_{y_{ij}}\mu_{ij}^{y_{ij}}}{f(\mu_{ij})},\ y_{ij} = 0, 1, \cdots;\ i = 1, \cdots, n;\ j = 1, \cdots, T. \tag{2.2}$$

where $b_{y_{ij}} > 0$, μ_{ij} is positive and $f(\mu_{ij}) = \sum_{y_{ij}=0}^{\infty} b_{y_{ij}}\mu_{ij}^{y_{ij}}$ is a finite and differentiable function of μ_{ij}. The Poisson distribution and the negative binomial distribution belong to the PS distributions with $b_{y_{ij}} = \frac{1}{y_{ij}!}$, $f(\mu_{ij}) = \exp(\mu_{ij})$, $b_{y_{ij}} = \frac{\Gamma(y_{ij}+\theta)}{\Gamma(y_{ij}+1)}$, and $f(\mu_{ij}) = \Gamma(\theta)(1-\mu_{ij})^{-\theta}$, respectively, where $\theta > 0$ is an overdispersion parameter for negative binomial model. For considering ZIPS random effects models μ_{ij} and $\pi_{ij}, i = 1, \cdots, n,\ j = 1, \cdots, T$ are considered as follows:

$$\begin{aligned} \log(\mu_{ij}) &= \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}_j + \kappa_{i1}^{'}\boldsymbol{b}_{i1}, \\ \text{logit}(\pi_{\text{ij}}) &= \boldsymbol{z}_{ij}^{'}\boldsymbol{\alpha}_j + \kappa_{i2}^{'}\boldsymbol{b}_{i2}, \end{aligned} \tag{2.3}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots \beta_p)'$ and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_q)'$ are the outcome-specific vectors of fixed-effect regression coefficients. Also, $\boldsymbol{\kappa}_1$ and $\boldsymbol{\kappa}_2$ are, respectively, q_1-dimensional and q_2-dimensional explanatory variables. The random effects $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}, \boldsymbol{b}_{i2})', i = 1, \cdots, n$ characterize the unobserved characteristics that are associated with the mean count for time j of subject i such that $\boldsymbol{b}_i \sim N_{q_1+q_2}(\boldsymbol{0}, \boldsymbol{D})$, where $N_K(\boldsymbol{0}, \boldsymbol{D})$ denotes a K-variate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{D}$.

2.2. Bayesian variable selection for ZIPS random effects model

We complete the Bayesian formulation of the proposed framework by specifying prior distributions for the unknown parameters. To facilitate outcome-specific variable selection, we adopt spike and slab priors for the regression parameters of equation (2.3). A spike and slab prior is a mixture of spike and slab distributions, where the spike is a distribution with its mass concentrated around zero and the slab is a flat distribution spread over the parameter space. The spike component, representing a null effect, can be either a positive mass at zero (Dirac spike, DS,[29, 41]) or a normal distribution with mean zero and a small variance (continuous spike, CS,[13]). In DS, a point mass at zero represents the prior belief that the coefficient in the regression equation is zero and the corresponding predictor has no relevance to the outcome, but in CS, each predictor's coefficient is modeled as coming from a mixture of two normal distributions with different variances: one with a density concentrated around zero, the other with a density spread out over large plausible values. Thus, unlike DS, it allows for 'almost zero' regression coefficients which is a much more realistic assumption than assuming that a predictor has absolutely no effect on the outcome. The slab component represents a non-null effect. [31] introduced this type before facilitating variable selection by constraining regression coefficients to be zero or not. Such a prior has

been widely used in the context of Bayesian stochastic search variable selection [13]. Figure 1 shows graphical examples of the CS and DS priors; slab densities are colored red and spike densities are colored blue.

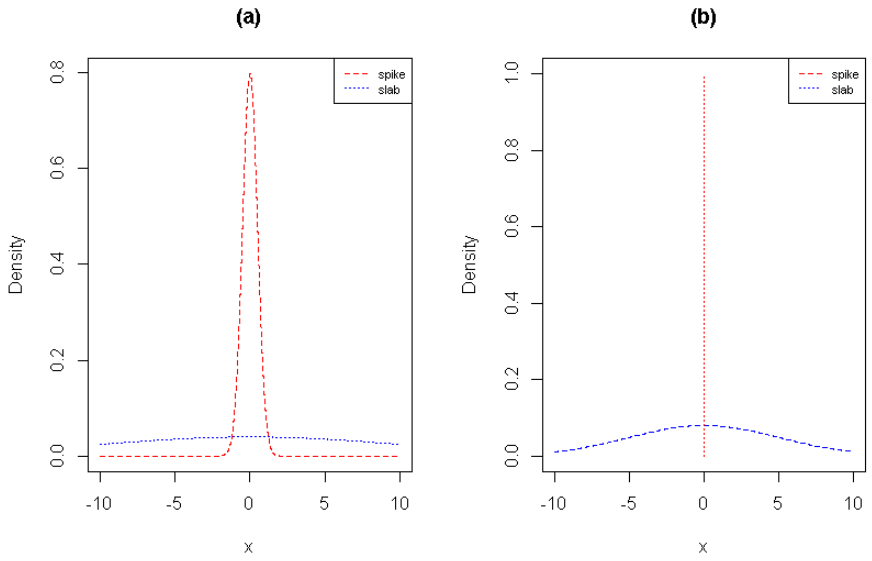

Figure 1: Example of the continuous spike (a) and the Dirac spike priors (b).

2.2.1. Continuous spike

The hierarchical setup of the zero-inflated random effects power series with CS prior [13, 3] for the regression coefficients of rate and probability models is

given by:

$$
\begin{aligned}
&Y_{ij}|\boldsymbol{b}_i \sim ZIPS(\pi_{ij},\ \mu_{ij}, \boldsymbol{a}),\\
&\boldsymbol{b}_i\ \sim\ N_2(\boldsymbol{0},\ \boldsymbol{D}),\\
&\beta_k|\zeta_k, \sigma^2_{\beta_k} \sim \zeta_k N(0,\ \sigma^2_{\beta_k}) + (1-\zeta_k)N(0,\ \tau^2_{\beta_k}),\ k=1,\ \cdots p,\\
&\tau^2_{\beta_k}|c_{1k}, c_{2k} \sim IG(c_{1k},\ c_{2k}),\\
&\zeta_k|\lambda_{\beta_k}\ \sim\ Ber(\lambda_{\beta_k}),\\
&\lambda_{\beta_k}|f_{1k}, f_{2k} \sim Beta(f_{1k},\ f_{2k}),\\
&\alpha_l|\omega_l, \sigma^2_{\alpha_l} \sim \omega_l N(0, \sigma^2_{\alpha_l}) + (1-\omega_l)N(0, \tau^2_{\alpha_l}),\ l=1,\ \cdots q,\\
&\tau^2_{\alpha_l}|d_{1l}, d_{2l} \sim IG(d_{1l}, d_{2l}),\\
&\omega_l|\lambda_{\alpha_l}\ \sim\ Ber(\lambda_l),\\
&\lambda_{\alpha_l}|m_{1l}, m_{2l} \sim Beta(m_{1l},\ m_{2l}),\\
&\boldsymbol{D} \sim IWishart(r,\ \boldsymbol{\Psi}),\\
&\boldsymbol{a}\ \sim\ \pi(\boldsymbol{a}),
\end{aligned}
\tag{2.4}
$$

where $\boldsymbol{\zeta} = (\zeta_1,\ \cdots\ \zeta_p)',\ \boldsymbol{\omega} = (\omega_1,\ \cdots\ \omega_q)'$ are vectors of binary latent variables indicating the membership of each regression coefficient to one of the mixture components, such that if $\zeta_k = 1$, then $\beta_k \sim\ N(0,\ \sigma^2_{\beta_k})$, otherwise, $\beta_k \sim\ N(0,\ \tau^2_{\beta_k})$ and its components can be considered non-zero values for large values of $\tau^2_{\beta_k}$ and $\boldsymbol{a}$ are the other parameters. The values of $\sigma^2_{\beta_k}$ and $\sigma^2_{\alpha_l}$ should be small (e.g. -3 or -4). By these conditions, $N(0, \sigma^2_{\beta_k})$ and $N(0, \sigma^2_{\alpha_l})$, leads us to a spike prior. The values of c_{1k} and c_{2k} are considered such that τ_{β_k} is large enough to yield a slab prior. $ZIPS(.,.,.)$ is used to denote a zero-inflated power series random effects model, $IG(.,.)$ denotes an inverse gamma distribution, $Beta(.,.)$ denotes a beta distribution, $Ber(.)$ is used to denote a Bernoulli distribution, $\pi(\boldsymbol{a})$ denotes the prior of $\boldsymbol{a}$; for example, $a > 0$ is an overdispersion parameter in the ZINB model, and the $\pi(a) = \Gamma(r_1, r_2)$ where $\Gamma(.,.)$ denotes a gamma distribution and $IWishart(r, \boldsymbol{\Psi})$ denotes an inverse Wishart distribution with parameters degrees of freedom r and scale matrix $\boldsymbol{\Psi}$. Note that, the natural conjugate prior to the multivariate normal distribution is the inverse Wishart distribution [4]. Due to its conjugacy, this is the most common prior implemented in the Bayesian paradigm. However, this prior has issues: the uncertainty for all variances is controlled by a single degree of freedom parameter (r) [16], the marginal distribution for the variances has a low density in a region near zero [17], and there is a prior dependence between correlations and variances [40]. These characteristics of the prior can impact posterior inferences about the covariance matrix. Here, the hyperparameters are chosen to be uninformative in the simulation study and application sections chosen to be uninformative.

2.2.2. Dirac spike

The hierarchical setup of zero-inflated random effects power series with a DS prior [29, 41] for the regression coefficients of rate and probability models is given by:

$$
\begin{aligned}
&Y_{ij}|\boldsymbol{b}_i \sim ZIPS(\pi_{ij}, \mu_{ij}, \boldsymbol{a}),\\
&\boldsymbol{b}_i \sim\ N_2(\boldsymbol{0}, \boldsymbol{D}),\\
&\beta_k|\gamma_k, \sigma^2_{\beta_k} \sim \gamma_k\delta_0(\beta_k) + (1-\gamma_k)N(0, \sigma^2_{\beta_k}), k = 1, \cdots, p,\\
&\sigma^2_{\beta_k} \sim IG(c_{1k}, c_{2k}),\\
&\gamma_k \sim Beta(f_{1k}, f_{2k})\\
&\alpha_l|\nu_l, \sigma^2_{\alpha_l} \sim \nu_l\delta_0(\alpha_l) + (1-\nu_l)N(0, \sigma^2_{\alpha_l}), l = 1, \cdots, q,\\
&\sigma^2_{\alpha_l} \sim IG(d_{1l}, d_{2l}),\\
&\nu_l \sim Beta(m_{1l}, m_{2l}),\\
&\boldsymbol{D} \sim IWishart(r, \boldsymbol{\Psi}),\\
&\boldsymbol{a} \sim \pi(\boldsymbol{a}),
\end{aligned}
\tag{2.5}
$$

where, $\delta_0(.)$ denotes a Dirac mass at 0, such that $\delta_0(\beta_k) = 1$ if $\beta_k = 0$ and $\delta_0(\beta_k) = 0$ if $\beta_k \neq 0$ and the other notations are the same as those described for CS.

3. STATISTICAL INFERENCE

Our inference is based on Metropolis-Hastings within Gibbs samplers because the full conditional distributions of the regression coefficients and the random effects do not have closed forms. The full conditional posterior distributions of all parameters and for all models are presented in supplementary materials A and B for CS and DS, respectively. A local Bayesian false discovery rate and a Bayes factor are proposed to perform the test of checking the significance of the regression coefficients [11].

3.1. Bayesian Implementation

3.1.1. Continuous spike

Let $\boldsymbol{\theta} = (\boldsymbol{b}, \boldsymbol{D}, \boldsymbol{a}, \{\beta_k, \sigma^{2r}_{\beta_k}, \zeta_k, \lambda_{\beta_k}\}_{k=1}^{p}, \{\alpha_l, \sigma_{\alpha_l}, \omega_l, \lambda_{\alpha_l}\}_{l=1}^{q})$ be the vector of all the unknown parameters in the model, $\boldsymbol{y} = (\boldsymbol{y}_i, \cdots, \boldsymbol{y}_n)'$, $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)', \boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)'$, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)'$, $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iT})'$, $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})'$ and

$\boldsymbol{z}_i = (z_{i1},\ \cdots\ ,\ z_{iq})'$.
The likelihood function of the model can be written as:

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\pi},\boldsymbol{\mu}) &= \prod_{i=1}^{n}\prod_{j=1}^{T}(1-\pi_{ij})p(Y_{ij}=y_{ij}|\mu_{ij})^{1-I(y_{ij})} \\ &\times\ \{\pi_{ij}+\ (1-\pi_{ij})p(Y_{ij}=0|\mu_{ij})\}^{I(y_{ij})}, \end{aligned} \tag{3.1}$$

where, [left=I(Y$_{ij}$) =]$align at = 20 \qquad y_{ij} = 1, 2, \cdots$
1 $\qquad y_{ij} = 0$. The joint posterior distribution of the unknown parameters is as follows:

$$\begin{aligned} &\pi(\theta|\boldsymbol{y},\boldsymbol{x}) \propto L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\pi},\boldsymbol{\mu}) \times p(\boldsymbol{\beta}|\boldsymbol{\tau}_1^2,\boldsymbol{\zeta},\boldsymbol{\lambda}_1) \times p(\boldsymbol{\tau}_1^2) \times p(\boldsymbol{\alpha}|\boldsymbol{\tau}_2^2,\boldsymbol{\omega},\boldsymbol{\lambda}_2) \times p(\boldsymbol{\tau}_2^2) \times p(\boldsymbol{b}|\boldsymbol{D}) \\ &\times p(\boldsymbol{D}) \times p(\boldsymbol{\zeta}|\boldsymbol{\lambda}_1) \times p(\boldsymbol{\omega}|\boldsymbol{\lambda}_2) \times p(\boldsymbol{\lambda}_1) \times p(\boldsymbol{\lambda}_2) \times p(\boldsymbol{a}) \\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{T}\{(1-\pi_{ij})p(Y_{ij}=y_{ij}|\mu_{ij})\}^{1-I(y_{ij})}\{\pi_{ij}+(1-\pi_{ij})p(Y_{ij}=0|\mu_{ij})\}^{I(y_{ij})} \\ &\times \phi(\boldsymbol{b}_i,\boldsymbol{0},\boldsymbol{D})|\boldsymbol{D}|^{\frac{-(r+2+1)}{2}}\exp(-\frac{\boldsymbol{D}^{-1}\boldsymbol{\Psi}}{2}) \\ &\times \prod_{k=1}^{p}\zeta_k\phi(\beta_k,0,\sigma_{\beta_k}^2)+(1-\zeta_k)\phi(\beta_k,0,\tau_{\beta_k}^2)\times\tau_{\beta_k}^{2c_{1k}-1}\exp(-\frac{c_{2k}}{\tau_{\beta_k}^2}) \\ &\times \prod_{l=1}^{q}\omega_l\phi(\alpha_l,0,\sigma_\alpha^2)+(1-\omega_l)\phi(\alpha_l,0,\tau_\alpha^2)\tau_{\alpha l}^{2d_{1l}-1}\exp(-\frac{d_{2l}}{\tau_{\alpha_l}^2}) \\ &\times \prod_{l=1}^{q}(\lambda_{\alpha_l})^{\omega_l}(1-\lambda_{\alpha_l})^{1-\omega_l}\prod_{k=1}^{p}(\lambda_{\beta_k})^{\zeta_l}(1-\lambda_{\beta_k})^{1-\zeta_l} \\ &\times \prod_{k=1}^{p}(\lambda_{\beta_k})^{f_{1k}-1}(1-\lambda_{\beta_k})^{f_{2k}-1}\prod_{l=1}^{q}(\lambda_{\alpha_l})^{m_{1l}-1}(1-\lambda_{\alpha_l})\pi(\boldsymbol{a}). \end{aligned}$$

Where $\boldsymbol{\lambda}_1 = (\lambda_{\beta_1},\ \cdots\ ,\ \lambda_{\beta_p})'$ and $\boldsymbol{\lambda}_2 = (\lambda_{\alpha_1},\ \cdots\ ,\ \lambda_{\alpha_q})'$. For applying MCMC methods, the full conditional posterior distributions of all the unknown parameters for this model are computed and presented in supplementary material A.

3.1.2. Dirac spike

Let $\boldsymbol{\theta} = (\boldsymbol{b},\boldsymbol{D},\boldsymbol{a},\ \{\beta_k,\ \sigma_{\beta_k}^2,\ \gamma_k\}_{k=1}^{p},\ \{\alpha_l,\ \sigma_{\alpha_l}^2,\ \nu_l\}_{l=1}^{q})$ be the vector of all the unknown parameters in the model, $\boldsymbol{y} = (\boldsymbol{y}_i,\cdots,\boldsymbol{y}_n)'$, $\boldsymbol{x} = (\boldsymbol{x}_1,\ \cdots\ ,\boldsymbol{x}_n)'$, $\boldsymbol{\pi} = (\pi_1,\ \cdots\ ,\ \pi_n)'$, $\boldsymbol{\mu} = (\mu_1,\ \cdots\ ,\ \mu_n)'$, $\boldsymbol{y}_i = (y_{i1},\ \cdots\ ,\ y_{iT})'$, $\boldsymbol{x}_i = (\boldsymbol{x}_{i1},\ \cdots\ ,\boldsymbol{x}_{ip})'$ and $\boldsymbol{z}_i = (\boldsymbol{z}_{i1},\ \cdots\ ,\boldsymbol{z}_{iq})'$. The joint posterior distribution of all unknown parameters,

given data, is as follows:

$$\begin{aligned}
&\pi(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x}) \propto L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\pi},\ \boldsymbol{\mu}) \times p(\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{\sigma}_1^2) \times p(\boldsymbol{\sigma}_1^2) \times p(\boldsymbol{\alpha}|\boldsymbol{\nu},\boldsymbol{\sigma}_2^2) \times p(\boldsymbol{\sigma}_2^2) p(\boldsymbol{b}_i|\boldsymbol{D}) \\
&\times p(\boldsymbol{D}) \times p(\boldsymbol{\gamma}) \times p(\boldsymbol{\nu}) \times p(\boldsymbol{a}) \\
&\propto \prod_{i=1}^{n}\prod_{j=1}^{T}\{(1-\pi_{ij})p(Y_{ij}=y_{ij}|\mu_{ij})\}^{1-I(y_{ij})}\{\pi_{ij}+(1-\pi_{ij})p(Y_{ij}=0|\mu_{ij})\}^{I(y_{ij})} \\
&\times \phi(\boldsymbol{b}_i,\ \boldsymbol{0},\ \boldsymbol{D})|\boldsymbol{D}|^{\frac{-(r+2+1)}{2}} \exp(-\frac{\boldsymbol{D}^{-1}\boldsymbol{\Psi}}{2}) \\
&\times \prod_{k=1}^{p} \gamma_k \phi(\beta_k,\ 0,\ \sigma_{\beta_k}^2) + (1-\gamma_k)\delta_0(\beta_k) \times \sigma_{\beta_k}^{2c_{1k}-1} \exp(-\frac{c_{2k}}{\sigma_{\beta_k}^2}) \\
&\times \prod_{l=1}^{q} \nu_l \phi(\alpha_l, 0,\ \sigma_{\alpha_l}^2) + (1-\nu_l)\delta_0(\alpha)\sigma_{\alpha\iota}^{2d_{1l}-1} \exp(-\frac{d_{2l}}{\sigma_{\alpha_l}^2}) \\
&\times \prod_{k=1}^{p} \gamma_k^{f_{1k}-1}(1-\gamma_k)^{f_{2k}-1} \prod_{l=1}^{q} \nu_l^{m_{1l}-1}(1-\nu_l)^{m_{2l}-1}\pi(\boldsymbol{a}),
\end{aligned}$$

where $\boldsymbol{\sigma}_1^2 = (\sigma_{\beta_1}^2,\ \cdots,\ \sigma_{\beta_p}^2)'$, $\boldsymbol{\sigma}_2^2 = (\sigma_{\alpha_1}^2,\ \cdots,\ \sigma_{\alpha_q}^2)'$, $\boldsymbol{\gamma} = (\gamma_1,\ \cdots,\ \gamma_p)'$ and $\boldsymbol{\nu} = (\nu_1,\ \cdots,\ \nu_q)'$. For applying MCMC methods, the full conditional posterior distributions of all the unknown parameters for this model are computed and presented in supplementary material B.

3.2. Variable selection

In the following, we propose some strategies to select variables in the rate and probability models.

3.2.1. Continuous spike

Let $\boldsymbol{\theta}^{(r)},\ r = 1, \cdots, M$ be M generated samples from the full conditional distributions of CS using MCMC. We first define a test for checking the significance of the parameters, as follows:

$$\begin{aligned}
&H_{0k} : \zeta_k = 0 \ \ versus \ \ H_{1k} : \zeta_k \neq 0;\ k = 1, \cdots, p, \\
&H_{0l} : \omega_l = 0 \ \ versus \ \ H_{1l} : \omega_l \neq 0;\ l = 1, \cdots, q.
\end{aligned} \tag{3.2}$$

Both a local Bayesian false discovery rate and a Bayes factor procedure are applied to perform this test.

Local Bayesian false discovery rate

Let $p(H_{0k}|\beta_k, \sigma^2_{\beta_k})$ be the posterior probability of the null hypothesis, i.e., the probability of making a false discovery for a non-null effect, which is called the local Bayesian false discovery rate for data (denoted by LBFDR). The phrase "local" comes from a single point 0 as the domain of the null hypothesis test [11]. Note that small values of LBFDR show strong evidence for the existence of a substantive effect. For computing LBFDR for each of the regression coefficients, we have

$$\begin{aligned} LBFDR_k &= P(H_{0k}|\beta_k, \sigma_{\beta_k^2}) = P(\zeta_k = 0|\beta_k, \sigma^2_{\beta_k}) \\ &\approx \frac{1}{M}\sum_{r=1}^{M} P(\zeta_k = 0|\beta_k^{(r)}, \sigma^{2(r)}_{\beta_k}), k = 1, \cdots, p, \\ LBFDR_l &= P(H_{0l}|\alpha_l, \sigma^2_{\alpha_l}) = P(\omega_l = 0|\alpha_l, \sigma^2_{\alpha_l}) \\ &\approx \frac{1}{M}\sum_{r=1}^{M} P(\omega_l = 0|\alpha_l^{(r)}, \sigma^{2(r)}_{\alpha_l}),\ l = 1, \cdots, q. \end{aligned} \tag{3.3}$$

where $\left(\beta_k{}^{(r)}, \sigma^{2(r)}_{\beta_k}, \alpha_l^{(r)}, \sigma^{2(r)}_{\alpha_l}\right)$ denotes the r^{th} generated sample of $\left(\beta_k, \sigma^2_{\beta_k}, \alpha_l, \sigma^2_{\alpha_l}\right)$ using the MCMC for $r = 1, \cdots, M$.

Bayes factor

The Bayes factor for each of the regression coefficients for testing (3.2) is defined as

$$\begin{aligned} BF_k &= \frac{P(H_{1k}|\beta_k, \sigma^2_{\beta_k})/P(H_{1k})}{P(H_{0k}|\beta_k, \sigma^2_{\beta_k})/P(H_{0k})}, \\ BF_l &= \frac{P(H_{1l}|\alpha_l, \sigma^2_{\alpha_l})/P(H_{1l})}{P(H_{0l}|\alpha_l, \sigma^2_{\alpha_l})/P(H_{0l})}. \end{aligned} \tag{3.4}$$

which describes the evidence of $H_{1k}(H_{1l})$ against $H_{0k}(H_{0l})$. Note that, $P(\zeta_k) = E(\lambda_{\beta_k}) = \frac{f_{1k}}{f_{1k}+f_{2k}}$ and $P(\omega_l) = E(\lambda_{\alpha_l}) = \frac{m_{1l}}{m_{1l}+m_{2l}}$. Also, since $LBFDR_k = P(H_{0k}|\beta_k, \sigma^2_{\beta_k})$ and $LBFDR_l = P(H_{0l}|\alpha_l, \sigma^2_{\alpha_l})$,

$$\begin{aligned} BF_k &= \frac{1 - LBFDR_k}{LBFDR_k} \times \frac{f_{1k}}{f_{1k} + f_{2k}}, \\ BF_l &= \frac{1 - LBFDR_l}{LBFDR_l} \times \frac{m_{1l}}{m_{1l} + m_{2l}}. \end{aligned} \tag{3.5}$$

Unlike, LBFDR, a large value of BF indicates strong evidence in favor of $H_{1k}(H_{1l})$.

3.2.2. Dirac spike

The same as those discussed for CS, let $\boldsymbol{\theta}^{(r)}, r = 1, \cdots, M$ be M generated samples from the full conditional distributions of DS using MCMC. The global test for checking DS is defined by:

$$\begin{aligned} H_{0k} : \beta_k = 0,\ versus\ H_{1k} : \beta_k \neq 0;\ k = 1, \cdots, p, \\ H_{0l} : \alpha_l = 0,\ versus\ H_{1l} : \alpha_l \neq 0;\ l = 1, \cdots, q. \end{aligned} \tag{3.6}$$

Local Bayesian false discovery rate

In this status, the following proposition gives insight into simplifying equation (3.6). Define indicator variables: $I_{1k},\ I_{2l},\ k = 1, 2, \cdots, p,\ l = 1, 2, \cdots, q$, such that

$$\begin{aligned} I_{1k} &= \begin{cases} 1 & \beta_k \neq 0 \\ 0 & \beta_k = 0, \end{cases} \\ I_{1l} &= \begin{cases} 1 & \alpha_l \neq 0 \\ 0 & \alpha_l = 0. \end{cases} \end{aligned} \tag{3.7}$$

Also, consider the hierarchical model (2.5), thus,

$$\begin{aligned} LBFDR_k &= P(\beta_k = 0 | \gamma_k, \sigma^2_{\beta_k}) \\ &\approx \frac{1}{M} \sum_{r=1}^{M} P(\beta_k = 0 | \gamma_k^{(r)}, \sigma^{2(r)}_{\beta_k}), \\ LBFDR_l &= P(\alpha_l = 0 | \nu_l, \sigma^2_{\alpha_k}) \\ &\approx \frac{1}{M} \sum_{r=1}^{M} P(\alpha_l = 0 |, \nu_l^{(r)}, \sigma^{2(r)}_{\alpha_l}). \end{aligned} \tag{3.8}$$

Bayes factor

The Bayes factor for each of the regression coefficients for DS is given by

$$\begin{aligned} BF_k &= \frac{P(\beta_k \neq 0 | \gamma_k, \sigma^2_{\beta_k}) / P(\beta_k \neq 0)}{P(\beta_k = 0 | \gamma_k, \sigma^2_{\beta_k}) / P(\beta_k = 0)}, \\ BF_l &= \frac{P(\alpha_l \neq 0 | \nu_l, \sigma^2_{\alpha_l}) / P(\alpha_l \neq 0)}{P(\alpha_l = 0 | \nu_l, \sigma^2_{\alpha_l}) / P(\alpha_l = 0)}, \end{aligned} \tag{3.9}$$

we have $P(\beta_k) = E(\gamma_{\beta_k}) = \frac{f_{1k}}{f_{1k}+f_{2k}}$ and $P(\alpha_l) = E(\nu_{\alpha_l}) = \frac{m_{1l}}{m_{1l}+m_{2l}}$. Thus the Bayes factor for DS is the same as that of equation (3.5).

4. SIMULATION STUDIES

In this section, some simulation studies are performed to investigate the performance of the proposed methods. For this purpose, the data is generated from random effects models under ZIP and ZINB. The sample sizes $n = 500$ and 1000 with $T = 6$ repeated measurements, $p = 15$, and $q = 10$ predictors are considered. Also, we consider 40000 MCMC iterations, including 20000 pre-convergence burn-in. The convergence of the chains is checked using Brooks-Gelman-Rubin (BGR) diagnostics [6, 19]. Also, some figures are given in supplementary material E for checking the convergence of the proposed model visually. The simulation studies are performed for $M = 100$ replications. For comparison of the results, relative bias (Rbias) and the root of the mean squared error (RMSE) are computed, these are defined as $Rbias(\theta) = \frac{\bar{\hat{\theta}}}{\theta} - 1$, $RMSE(\theta) = \sqrt{\frac{\sum_{r=1}^{M}(\hat{\theta}_r - \theta)^2}{M}}$, where $\hat{\theta}_r$ is the estimated value of parameter θ for the $r-$th simulation run, M is the number of simulation runs, and $\bar{\hat{\theta}} = \frac{\sum_{r=1}^{M}\hat{\theta}_r}{M}$.
To investigate the performance of the proposed approaches in variable selection, we consider the true positive rate (TPR), the false positive rate (FPR), and the Matthews correlation coefficient (MCC) criteria [**?**]. The latter is defined as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{4.1}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. The MCC and TPR are expected to reach 1, and the FPR is expected to be near zero for a good performance.

4.1. Zero-inflated Poisson random effects model

We consider a ZIP random effects model, such that, $Y_{ij}|\mu_{ij}, \pi_{ij} \sim ZIP(\mu_{ij},\ \pi_{ij})$ is used to denote it, where μ_{ij} and π_{ij} are considered the same as equations (2.3). For this simulation study, the explanatory variables $\boldsymbol{x}$ and $\boldsymbol{z}$ are randomly drawn from multivariate normal distributions $N_{15}(\mathbf{0}, \boldsymbol{I})$ and $N_{10}(\mathbf{0}, \boldsymbol{I})$, respectively. Also, four scenarios are considered for the real values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ these scenarios are different for the real values of the regression coefficients:
Scenario 1:

$$\boldsymbol{\beta} = (\underbrace{1, \cdots, 1}_{5}, \underbrace{0, \cdots, 0}_{10})',$$
$$\boldsymbol{\alpha} = (\underbrace{1, \cdots, 1}_{5}, \underbrace{0, \cdots, 0}_{5})'.$$

Scenario 2:

$$\boldsymbol{\beta} = (\underbrace{0.5, \cdots, 0.5}_{5}, \underbrace{0, \cdots, 0}_{10})',$$
$$\boldsymbol{\alpha} = (\underbrace{0.5, \cdots, 0.5}_{5}, \underbrace{0, \cdots, 0}_{5})'.$$

Scenario 3:

$$\boldsymbol{\beta} = (\underbrace{0.5, \cdots, 0.5}_{15})',$$
$$\boldsymbol{\alpha} = (\underbrace{0.5, \cdots, 0.5}_{10})'.$$

Scenario 4:

$$\boldsymbol{\beta} = (\underbrace{0, \cdots, 0}_{15})',$$
$$\boldsymbol{\alpha} = (\underbrace{0, \cdots, 0}_{10})'.$$

In scenario 2, the values of the regression coefficients are reduced to check if the Bayesian approach for variable selection has a good performance after the reduction of the signals. Also, scenario 3 was simulated to represent cases in which all the covariates had non-zero effects, and scenario 4 was simulated to represent cases in which all the covariates had zero effects. Also, the real value for the covariance of the random effects is as follows:

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}.$$

For DS and CS, we set the hyperparameters to $c_{1k} = 0.01$, $c_{2k} = 0.01$, $d_{1l} = 0.01$, $d_{2l} = 0.01$, $f_{1k} = 1$, $f_{2k} = 1$, $m_{1l} = 1$, $m_{2l} = 1$, $k = 1, \cdots, p$, $l = 1, \cdots, q$, also for CS, we set $\sigma^2_{\beta_k} = \sigma^2_{\alpha_l} = 10^{-4}$ and the hyperparameters of inverse Wishart distribution are considered as $r = 3$ and $\boldsymbol{\Psi} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, so that the priors are low-informative [15, 17]. The results of this simulation study are reported as follows:

Continuous spike The results of this simulation study are reported in Tables 1 and 2 for scenario 1 and Tables C.1 and C.2 for scenario 2 of supplementary material C. Table 1 includes estimates of posterior mean, standard errors, posterior median, RMSE, and Rbias. The values of Rbiases and RMSEs for all the parameters are small, and by increasing the sample size from n = 500 to n = 1000, the accuracy and efficiency of the estimations are increased. Also, Table 2 reports TPR, FPR, and MCC, where the results are based on threshold 1 of BF and threshold 0.05 of LBFDR. The values of the MCC as a balanced measure between TPR and FPR, show that the performances of BF and LBFDR are similar. Overall, the results

show that all the parameters are well estimated, and the values of the MCC show the good performance of the method in variable selection. The values of the regression coefficients are reduced in scenario 2, and the results of this simulation study are summarized in Tables C.1 and C.2. The results of these tables show similar results as those of scenario 1, even with smaller values of the signals.

Dirac spike The results of these simulation studies are reported in Tables C.3 and C.4 for scenario 1. In Table C.4, the value of the criteria for median thresholding is also given. Table C.3 shows that all the parameters are well estimated; also, the values of the MCC are given in Table C.4 and confirm that the performances of BF and LBFDR are similar, and they are better than median thresholding. The values of the regression coefficients are reduced in scenario 2, and the results of this simulation study are summarized in Tables C.5 and C.6. The results of these tables show the same results as in scenario 1.

4.2. Zero-inflated negative binomial random effects model

In this simulation study, we simulate data from a ZINB random effects model as follows:

$$Y_{ij}|\mu_{ij},\pi_{ij} \sim ZINB(\phi,\ \frac{\phi}{\phi+\mu_{ij}},\ \pi_{ij}),$$

where μ_{ij} and π_{ij} are considered the same as equation (2.3). The parameterization and the real values of parameters μ_{ij} and π_{ij} are the same as the set of real values and those described for two scenarios in the previous subsection; also, we set $\phi = 2,\ 0.25$. The results of this simulation study are reported as follows:

Continuous spike The results of this simulation study for $\phi = 2$ are reported in Tables 2 and C.7 for scenario 1 and Tables C.2 and C.8 for scenario 2. Also, the results of this simulation study for $\phi = 0.25$ are given in Tables E.1 and E.2 for scenario 1 and in Tables E.3 and E.4 for scenario 2 of supplementary material E. The performance of the proposed model is good in both parameter estimation and variable selection, and this is in agreement with different values of ϕ.

Dirac spike The results of this simulation studies for $\phi = 2$ are reported in Tables C.4 and C.9 for scenario 1 and in Tables C.6 and C.10 for scenario 2. Also, the results of this simulation study are given in E.5 and E.6 for scenario 1 and in Tables E.7 and E.8 for scenario 2 when $\phi = 0.25$. The results of these tables, the same as those in CS, show that all the parameters are estimated well and the values of the MCC show the performance of the method for variable selection as well.

A comparison between CS and DS for variable selection in the proposed model for both ZIP and ZINB models shows that DS performs better than CS based on values of MCC. Also, the parameter estimates by DS are closer to the true values of the parameters than those obtained by CS.

4.3. Results of simulation studies for scenarios 3 and 4

A sample size of $n = 500$ is considered for scenarios 3 and 4, which are whole non-zero and whole zero signals, respectively. The results of these two scenarios, which are the same as the previous scenarios, include estimates, standard errors, posterior median, RMSE, and Rbias. It is not possible to check the performance of the variable selection of the proposed model by TPR, FPR, and MCC when all of the signals are significant or all of them are non-significant. Therefore, instead, the mean and standard deviation of LBFDR and BF are given in the tables of results for these two scenarios.

4.3.1. ZIP model

The results of this simulation study for CS prior are reported in Tables D.1 for scenario 3 and D.2 for scenario 4 of supplementary material D. The results show that all the parameters are well estimated, and based on the mean of BF and LBFDR, all the variables are selected for scenario 3, but none of them are selected for scenario 4.
The results of this simulation study for DS prior are reported in Table D.3 and Table D.4 for scenarios 3 and 4, respectively. The results of these tables also confirm the good performance of the proposed model.

4.3.2. ZINB model

The results of this simulation study for CS prior are given in Table D.5 for scenario 3 and in Table D.6 for scenario 4. As with our results for the ZIP model, the results show that the performance of the proposed model is good in parameter estimation and variable selection. Also, for the DS prior, the results are shown in Table D.7 for scenario 3 and in Table D.8 for scenario 4. The results of these tables also confirm the good performance of the model in both parameter estimation and variable selection.

Table 1: Results of the simulation study of CS for generated data under the ZIP random effects model for scenario 1. The posterior mean, the standard deviation of estimators, the posterior median, the root of the mean squared error (RMSE), and relative bias (Rbias) for each of the parameter estimates for M= 100 simulated data with sample sizes of 500 and 1000. The generated data are analyzed with the ZIP model (*: the relative bias cannot be calculated since the real value of the related parameter is zero).

Parameter	True	mean	sd	median	RMSE	Rbias
			n=500			
β_1	1.000	0.991	0.044	0.993	0.004	-0.014
β_2	1.000	0.959	0.046	0.959	0.003	-0.118
β_3	1.000	0.966	0.043	0.968	0.004	-0.064
β_4	1.000	0.923	0.044	0.925	0.003	-0.040
β_5	1.000	0.989	0.044	0.990	0.000	-0.020
β_6	0.000	-0.005	0.002	-0.002	0.000	*
β_7	0.000	-0.003	0.002	-0.001	0.000	*
β_8	0.000	-0.043	0.003	-0.044	0.002	*
β_9	0.000	0.003	0.002	0.001	0.000	*
β_{10}	0.000	0.011	0.004	0.003	0.000	*
β_{11}	0.000	-0.005	0.003	-0.002	0.000	*
β_{12}	0.000	0.020	0.004	0.018	0.000	*
β_{13}	0.000	-0.031	0.004	-0.038	0.001	*
β_{14}	0.000	-0.003	0.003	-0.001	0.000	*
β_{15}	0.000	-0.002	0.002	-0.001	0.000	*
α_1	1.000	0.955	0.063	0.454	0.002	-0.092
α_2	1.000	0.918	0.065	0.919	0.000	-0.038
α_3	1.000	0.943	0.065	0.942	0.002	-0.084
α_4	1.000	0.992	0.065	0.992	0.000	-0.016
α_5	1.000	0.977	0.065	0.976	0.001	-0.048
α_6	0.000	0.000	0.005	0.000	0.000	*
α_7	0.000	-0.005	0.005	-0.001	0.000	*
α_8	0.000	0.002	0.006	0.000	0.000	*
α_9	0.000	0.000	0.006	0.000	0.000	*
α_{10}	0.000	0.016	0.009	0.002	0.000	*
			n=1000			
β_1	1.000	0.997	0.043	0.996	0.002	-0.008
β_2	1.000	0.906	0.045	0.906	0.002	-0.012
β_3	1.000	0.973	0.043	0.974	0.003	-0.052
β_4	1.000	0.995	0.042	0.994	0.003	-0.052
β_5	1.000	0.994	0.043	0.995	0.001	-0.010
β_6	0.000	-0.007	0.002	-0.003	0.000	*
β_7	0.000	-0.003	0.002	-0.001	0.000	*
β_8	0.000	0.001	0.002	0.001	0.000	*
β_9	0.000	-0.020	0.003	-0.022	0.000	*
β_{10}	0.000	-0.001	0.002	0.000	0.000	*
β_{11}	0.000	0.002	0.003	0.001	0.000	*
β_{12}	0.000	0.001	0.002	0.001	0.000	*
β_{13}	0.000	0.002	0.003	0.001	0.000	*
β_{14}	0.000	0.000	0.002	0.000	0.000	*
β_{15}	0.000	0.000	0.002	0.000	0.000	*
α_1	1.000	0.987	0.061	0.986	0.000	-0.028
α_2	1.000	0.907	0.063	0.906	0.000	-0.012
α_3	1.000	0.928	0.062	0.927	0.001	-0.054
α_4	1.000	0.908	0.063	0.908	0.000	-0.016
α_5	1.000	0.970	0.063	0.969	0.001	-0.062
α_6	0.000	0.003	0.005	0.001	0.000	*
α_7	0.000	0.018	0.006	0.005	0.000	*
α_8	0.000	-0.004	0.006	-0.001	0.000	*
α_9	0.000	-0.001	0.005	0.000	0.000	*
α_{10}	0.000	0.002	0.007	0.000	0.000	*

Table 2: Mean (SD) of true/false positive rate (TPR/FPR) and Matthews correlation coefficient (MCC) of BF and LBFDR for ZINB and ZIP random effects models of Scenario 1 for CS with M = 100 simulated data with sample sizes of 500 and 1000.

		ZIP		ZINB	
n		BF	LBFDR	BF	LBFDR
500	TPR	0.925(0.096)	0.925(0.096)	0.966(0.070)	0.975(0.061)
	FPR	0.033(0.038)	0.000(0.000)	0.053(0.021)	0.000(0.000)
	MCC	0.897 (0.106)	0.941(0.077)	0.838(0.088)	0.938(0.079)
1000	TPR	1.000(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)
	FPR	0.025(0.011)	0.000(0.000)	0.000(0.000)	0.000(0.000)
	MCC	0.912(0.011)	1.000(0.000)	1.000(0.000)	1.000(0.000)

5. APPLICATION: THE RAND HEALTH INSURANCE EXPERIMENT

In this section, we shall analyze the RAND Health Insurance Experiment (RAND HIE) data from [8]. The data investigate how medical care utilization, measured by the number of visits to a medical doctor (MD), is affected by health insurance plans, demographic characteristics, and the health status of patients. This particular data set consists of 5792 participants with 20,190 observations in total. The vast majority of participants are observed either three or five times, and each observation corresponds to data collected for the participant in a given year. The response variable MD is the yearly count of outpatient visits to physicians, which represents the health care utilization for the experimental subject for a specific year. Over 30% of the observations are zeros, motivating the use of the proposed approach. We use a simple zero-score test for checking the zero-inflation in the data, and not having zero-inflation in the data is rejected by a p-value of 0.000. The bar plot of this variable is presented in Figure 2 where the zero-inflation in the data set is also implied. The insurance variables were randomly assigned and included, an indicator variable for plans with a deductible (IDP), aparticipation-incentive payment function (LPI), a maximum dollar-expenditure function (FMDE), and other covariates including factors representing the demographic information including a log of annual family income (LINC), gender (FEMALE), race (BLACK), education of the head of household in years (EDUCDEC), age, an indicator for age less than 18 (CHILD), log of the family size (LFAM), a coinsurance rate (LC), health status including an indicator for physical limitations (PHYSLIM), index of chronic diseases (NDISEASE), fair self-rated health (HLTHF), good self-rated health (HLTHG) and poor self-rated health (HLTHP). For detailed variable definitions and summary statistics of each variable, see Table F.4 of supplementary material D. For checking the effect of time, we let time be modeled as a square polynomial in the rate model [2]. The

proposed variable selection models are applied to analyze the data such that:

$$\begin{aligned}(5.1)\quad \log(\mu_{ij}) &= \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 \text{IDP}_i + \beta_4 \text{LPI}_i + \beta_5 \text{FMDE}_i \\ &+ \beta_6 \text{LINC}_i + \beta_7 \text{FEMALE}_i + \beta_8 \text{PHYSLM}_i + \beta_9 \text{BLACK}_i \\ &+ \beta_{10} \text{EDUCDEC}_i + \beta_{11} \text{NDISEASE}_i + \beta_{12} \text{HLTHF}_i + \beta_{13} \text{HLTHG}_i \\ &+ \beta_{14} \text{HLTHP}_i + \beta_{15} \text{AGE}_i + \beta_{16} \text{CHILD}_{ij} + \beta_{17} \text{LFAM}_i \text{LC}_{ij} \\ &+ \beta_{18} \text{FCHILD}_{ij} + b_{1i}\end{aligned}$$

and

$$\begin{aligned}(5.2)\quad \text{logit}(\pi_{ij}) &= \alpha_0 + \alpha_1 t_{ij} + \alpha_2 \text{IDP}_i + \alpha_3 \text{LPI}_i + \alpha_4 \text{FMDE}_i \\ &+ \alpha_5 \text{LINC}_i + \alpha_6 \text{FEMALE}_i + \alpha_7 \text{PHYSLM}_i + \alpha_8 \text{BLACK}_i \\ &+ \alpha_9 \text{EDUCDEC}_i + \alpha_{10} \text{NDISEASE}_i + \alpha_{11} \text{HLTHF}_i \\ &+ \alpha_{12} \text{HLTHG}_i + \alpha_{13} \text{HLTHP}_i + \alpha_{14} \text{AGE}_i + \alpha_{15} \text{CHILD}_{ij} \\ &+ \alpha_{16} \text{LFAM}_i + \alpha_{17} \text{LC}_{ij} + \alpha_{18} \text{FCHILD}_{ij} + b_{2i}.\end{aligned}$$

where $\boldsymbol{b}_i \sim N_2(\boldsymbol{0}, \boldsymbol{D})$. The prior distributions for the unknown parameters of the ZINB and ZIP random effects models are the same as those of the simulation study section and are given by:
$\beta_k \sim \gamma_k \delta_0(\beta_k) + (1-\gamma_k) N(0, \sigma^2_{\beta_k}),\ \gamma_k \sim Beta(0.1, 0.1),\ \sigma^2_{\beta_k} \sim IG(0.1, 0.1),\ k = 1, \cdots, 19,\ \alpha_l \sim \nu_l \delta_0(\alpha_l) + (1-\nu_l) N(0, \sigma^2_{\alpha_l}), \nu_l \sim Beta(0.1, 0.1),\ \sigma^2_{\alpha_l} \sim IG(0.1, 0.1),\ l = 1, \cdots, 18,\ \boldsymbol{D} \sim IWishart(2, \Psi)$,
$\Psi = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. In the Bayesian method, two parallel MCMC chains are run with different initial values for 40,000 iterations each. Then, we discarded the first 20,000 iterations as pre-convergence burn-in and retained 20,000 for the posterior inference. For checking the convergence of the MCMC chains, we have used the Gelman–Rubin diagnostic test. The results, including parameter estimates, standard deviations, 95% credible intervals, LBFDR, and Gelman–Rubin statistics for analyzing the data using ZINB and ZIP random effects models, are presented in Tables 3 and F.1. The negative binomial random effects model (NB) and the Poisson random effects model (P) are also applied to analyze the data. The prior distributions for the unknown parameters of the NB and P random effects models are given by:
$\beta_k \sim \gamma_k \delta_0(\beta_k) + (1-\gamma_k) N(0, \sigma^2_{\beta_k}),\ \gamma_k \sim Beta(0.1, 0.1),\ \sigma^2_{\beta_k} \sim IG(0.1, 0.1),\ k = 1, \cdots, 19,\ b_i \sim N(0, \sigma^2_b),\ \sigma^2_b \sim IG(0.1, 0.1),\ i = 1, \cdots, n$. Tables F.2 and F.3 show the results for the NB and P random effects models, respectively. Based on the values of DIC, the performance of the ZINB random effects model is better than that of the ZIP, P, and NB random effects models. Based on the results of Table 3, IDP, FMDE, LINC, FEMALE, PHYSLIM, BLACK, NDISEASE, HLTHF, HLTHP, CHILD, LFAM and FEMCHILD are selected for the rate model (they have LBFDR < 0.05), that is, these variables are significant predictors such that increasing in IDP leads to decreasing the medical doctor visit (MD) and LINC is positively significant meaning the more the natural logarithm of income (LINC), the more visits to a MD. FMDE is a negatively significant predictor, such that by increasing it, the probability of zero decreases.

The greater the physical limitations (PHYSLIM), the larger the values of the estimated probability of nonzero. BLACK is positively significant, which means the number of visits to a medical doctor of black patients is higher than that of white patients, and increasing the NDISEASE leads to higher MD numbers. Also, HLTHF and HLTHP are factors that motivate patients to visit the doctor. CHILD is positively significant, which means the patient who is under 18 years old has more MD than others. The effect of FEMALE on MD number depends on the level of CHILD, and the effect of CHILD on MD number depends on the level of FEMALE. These estimates indicate females visit the doctor more than males. Also, LPI, EDUCDEC, HLTHG, AGE, and LC have a LBFDR > 0.05 which means they are not significant predictors. Time and time^2 have LBFDR > 0.05, i.e., time is not a significant predictor. Also, in the probability model, IDP, LPI, FMDE, FEMALE, BLACK, NDISEASE, CHILD, and FEMCHILD are significant predictors, such that increasing IDP leads to decreasing medical doctor visits, with increasing participation incentive payment (LPI) the probability of nonzeros decreases. FEMALE is positively significant, which means the probability of zero for men is the largest. BLACK is negatively significant, which means the white patient visits a MD less than the black patient. Also, increasing NDISEASE leads to a larger probability of nonzeros.

As mentioned before, the zero-inflated regression model assumes that the count numbers arise from a two-component mixture of a standard count distribution and a degenerated distribution at zero. Under such models, a zero can belong to either the degenerate state or the count distribution, but it is typically impossible to with certainty to determine to which state it belongs [27]. As two examples of the data, consider the 6th and 12th patients with $\boldsymbol{y}_6 = (1, 0, 0, 0, 0)$ and $\boldsymbol{y}_{12} = (1, 0, 7)$. The 6th patient is a white 16-year-old girl who has neither physical limitations nor chronic diseases. Also, the 12th patient is a black 61-year-old man who has no physical limitations but has chronic diseases. The other characteristics of these two patients are as follows:

	6th patient	12th patient
IDP	1	0
LPI	0.22	0.48
FMDE	1.16	1.29
LINC	0.40	0.52
EDUCDEC	8	18
HLTHF	0	0
HLTHG	0	1
HLTHP	0	0
LFAM	4	2
LC	0.45	0.52

The probability of being zero for the 6th patient at different time points can be estimated by considering ZINB as described in (5.1) and (5.2) and it is given by $\boldsymbol{\pi}_6 = (0.08, 0.64, 0.52, 0.72, 0.61)$; that is, for example, at the first time the probability of coming from a degenerated distribution is 0.08 while for the second time, it is 0.64. Also, this probability for the 12th patient is $\boldsymbol{\pi}_{12} = (0.09, 0.61, 0.02)$.

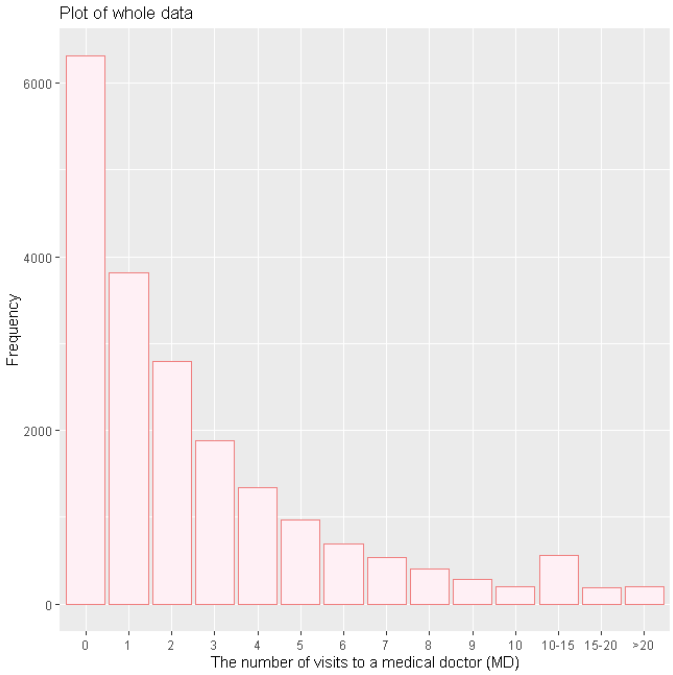

Figure 2: Barplot of the number of visits to a medical doctor of rand health Insurance data.

6. CONCLUSION

In this paper, we have discussed Bayesian variable selection methods for the zero-inflated power series distribution, specifically ZIP and ZINB random effects models that have been used via spike and slab priors for analyzing longitudinal count data with extra zeros.

We have evaluated the selection accuracy of DS and CS approaches through some simulation studies. Also, we have defined a test for checking the significance of the parameters, both a local Bayesian false discovery rate with a threshold of 0.05 and a Bayes factor procedure with a threshold of 1 are applied to perform this test. The other thresholds can also be considered to investigate H_0. The simulation studies show that applying DS has better performance than applying CS. A real data set from the RAND health insurance experiment has been analyzed as an illustrative example. The proposed variable selection models by DS spike are applied to select the important variables in this data set, where the non-significant variables shrink to zero and those estimated are considered significant variables. To the best of our knowledge, ZINB and ZIP are the best models for analyzing zero-inflated count data, but if the range of the zero-inflated data is restricted to a special range such as $0, 1, \cdots, K$, the zero-inflated binomial model is a more appropriate model than the ZINB and ZIP models. The ZINB regression model allows for over-dispersion in the model and can be used to quantify various parameters more effectively. We have used DIC to select among different

Table 3: Parameter estimates (Est.), standard deviation (SD.), 2.5%: lower bound of 95% credible interval, 97.5%: upper bound of 95% credible interval, local Bayesian false discovery rate (LBFDR), and Gelman-Rubin statistics ($\hat{R}$) for analyzing RAND data using the ZINB model.

	Est.	SD.	2.5%	97.5%	LBFDR	$\hat{R}$
	Model of the rate (μ_{ij})					
Intercept (β_0)	0.824	0.044	0.738	0.908	0.000	1.013
time (β_1)	0.000	0.001	0.000	0.000	0.992	1.029
time^2 (β_2)	0.000	0.000	0.000	0.000	1.000	1.000
IDP (β_3)	-0.169	0.037	-0.241	-0.099	0.000	1.004
LPI (β_4)	0.001	0.005	0.000	0.011	0.968	1.016
FMDE (β_5)	-0.129	0.015	-0.157	-0.099	0.000	1.006
LINC (β_6)	0.107	0.019	0.070	0.145	0.000	1.005
FEMALE (β_7)	0.258	0.039	0.184	0.339	0.000	1.006
PHYSLM (β_8)	0.270	0.044	0.183	0.356	0.000	1.009
BLACK (β_9)	0.353	0.056	0.241	0.463	0.000	1.015
EDUCDEC (β_{10})	0.000	0.014	0.000	0.000	0.959	1.017
NDISEASE (β_{11})	0.175	0.015	0.146	0.207	0.000	1.016
HLTHF (β_{12})	0.206	0.058	0.087	0.315	0.011	1.101
HLTHG (β_{13})	0.000	0.004	0.000	0.000	0.986	1.004
HLTHP (β_{14})	0.426	0.107	0.211	0.626	0.001	1.000
AGE (β_{15})	0.000	0.000	0.000	0.000	0.999	1.002
CHILD (β_{16})	0.174	0.042	0.093	0.256	0.000	1.001
LFAM (β_{17})	-0.144	0.027	-0.199	-0.093	0.000	1.002
LC (β_{18})	-0.002	0.016	-0.026	0.000	0.959	1.106
FEMCHILD (β_{19})	0.217	0.058	0.105	0.330	0.000	1.006
	Model of the probability (π_{ij})					
Intercept (α_0)	5.390	0.741	4.356	7.166	0.000	1.050
time (α_1)	0.002	0.016	0.000	0.032	0.948	1.062
IDP (α_2)	-1.026	0.341	-1.799	-0.416	0.003	1.074
LPI (α_3)	-0.663	0.160	-0.991	-0.357	0.000	1.022
FMDE (α_4)	-1.044	0.175	-1.395	-0.714	0.000	1.007
LINC (α_5)	0.151	0.133	0.000	0.383	0.357	1.015
FEMALE (α_6)	-2.475	0.445	-3.450	-1.662	0.000	1.018
PHYSLM (α_7)	-0.005	0.122	-0.334	0.265	0.857	1.014
BLACK (α_8)	-4.915	0.605	-6.255	-4.011	0.000	1.032
EDUCDEC (α_9)	-0.028	0.151	-0.512	0.159	0.825	1.035
NDISEASE (α_{10})	0.452	0.152	0.164	0.762	0.010	1.045
HLTHF (α_{11})	-0.039	0.165	-0.577	0.096	0.840	1.006
HLTHG (α_{12})	0.004	0.087	-0.165	0.246	0.869	1.054
HLTHP (α_{13})	0.178	0.459	-0.270	1.526	0.690	1.030
AGE (α_{14})	0.000	0.000	0.000	0.000	1.000	1.000
CHILD (α_{15})	-1.458	0.367	-2.226	-0.791	0.000	1.028
LFAM (α_{16})	0.005	0.064	0.000	0.153	0.932	1.044
LC (α_{17})	-0.118	0.296	-1.051	0.054	0.740	1.032
FEMCHILD (α_{18})	-2.995	0.628	-4.256	-1.807	0.000	1.015
D_{11}	0.493	0.228	0.226	1.007	-	1.010
$D_{12}(D_{21})$	0.925	0.128	0.516	1.803	-	1.101
D_{22}	1.021	0.020	0.561	2.907	-	1.002
ϕ	3.817	0.140	3.553	4.098	-	1.002
DIC	183127.5					

models, and we have concluded that the zero-inflated negative binomial random effects model is a flexible model to be assumed for analyzing this data.

As a future work, the proposed method can be applied to semi-parametric modeling data sets by considering spline. For this purpose, equation (2.3) can be

improved to be

$$\begin{aligned}\log(\mu_{ij}) &= \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}_j + g_1(t_{ij}) + b_{i1},\\ \text{logit}(\pi_{\text{ij}}) &= \boldsymbol{z}_{ij}^{'}\boldsymbol{\alpha}_j + g_2(t_{ij}) + b_{i2},\ i = 1,\cdots,n,\ j = 1,\cdots,T,\end{aligned}$$

where $g_1(.)$ and $g_2(.)$ are unknown smooth functions of time. For example, they can be considered as follows:

$$g_k(t_{ij}) = \alpha_{k0} + \alpha_{k1}t_{ij} + \ldots + \alpha_{kd_k}t_{ij}^{d_k} + \sum_{l=1}^{K_k}\alpha_{k,d_k+l}(t_{ij} - \kappa_i^l)_+^{d_k},\quad k = 1,2,$$

where d is the degree of the polynomial component, K_k is the number of interior knots, κ_i^l is referred to as knots of the i^{th} subject, $(a)_+ = max(0,a)$, and $\boldsymbol{\alpha}_k = (\alpha_{k0},\cdots,\alpha_{kd_k},\alpha_{kd_k+1}\cdots,\alpha_{kK_k})$ is the vector of spline coefficients. By considering these functions, all the approaches in this paper can be applied to this model, too.

As future work, we can consider marginalized zero-inflated negative binomial (MZINB) and marginalized zero-inflated Poisson (MZIP) models to model the population means count directly, allowing straightforward inference for overall exposure effects that account for both excess zeros and overdispersion [37]. Also, the model can be extended to analyze data in the presence of missing values. For this purpose, a non-ignorable missing mechanism should be considered. Also, Bayesian variable selection by using global-local shrinkage [22] priors can be applied in future works. The Wishart-gamma and half-Cauchy priors can also be considered for the random effects covariance matrix and variance components, respectively.

REFERENCES

[1] Alsalim, N. and Baghfalaki, T. (2021). Variable selection for longitudinal zero-inflated power series transition model. *Journal of Biopharmaceutical Statistics* , (**31**) 668–685.

[2] Baghfalaki T. and Ganjali, M. (2021). Approximate Bayesian inference for joint linear and partially linear modeling of longitudinal zero-inflated count and time to event data. *Statistical Methods in Medical Research*, (**30**), 1484-1501.

[3] Baghfalaki T., Sugier PE., Truong T., Pettitt AN., Mengersen K., Liquet B. (2021). Bayesian meta-analysis models for cross-cancer genomic investigation of pleiotropic effects using group structure. *Statistics in Medicine* 15, 40(**6**), 1498-1518.

[4] Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage, *Statistica Sinica.* **10**, 1281–1312.

[5] Bekalo, D.B., Kebede, D.T. (2021). Zero-Inflated Models for Count Data: An Application to Number of Antenatal Care Service Visits, 1–26. *Ann. Data. Sci*

[6] Brooks, S. P., and A. Gelman. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics.* **7**, 434–455.

[7] Crainiceanu, C. M., Ruppert, D. and Wand, M. P. (2005). Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *Journal of Statistical Software.* 14 (**14**), 1–24.

[8] Deb, P., Trivedi, P., (2002). The Structure of Demand for Health Care: Latent Class Versus Two-Part Models. *Journal of Health Economics.* 21 (**4**), 601–625.

[9] Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002). Analysis of Longitudinal Data. Oxford: Oxford University Press.

[10] DiGiulio DB., Callahan, BJ., McMurdie, PJ., Costello, EK., Lyell, DJ., Robaczewska A., Sun, CL., Goltsman, DS., Wong, RJ., Shaw, G., Stevenson, DK. (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences.* 112(**35**), 11060–5.

[11] Efron, B. (2012) Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. Cambridge University Press.

[12] Feng, C.X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *J Stat Distrib App.* **8**, 8.

[13] George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *J.Amer.Stat. Assoc.* **88**, 881–889.

[14] García-Donato, G., Castellanos, M.E., Quirós. (2021). A Bayesian Variable Selection with Applications in Health Sciences. *Mathematics.* **9**, 218.

[15] Gu, X., Tadesse, M.G., Foulkes, A.S. (2020). Bayesian variable selection for high dimensional predictors and self-reported outcomes. *BMC Med Inform Decis Mak.* **20**, 212.

[16] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*, CRC press.

[17] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis.* **1**, 515–533.

[18] Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press.

[19] Gelman, A., and D. B. Rubin. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science.* **7**, 457–472.

[20] Gupta, P. L., R. C. Gupta, and R. C. Tripathi. (1996). Analysis of zero-adjusted count data. *Comput. Stat. Data Anal.* **23**, 207–218.

[21] Hall, D.B. (2000). Zero-Inflated Poisson and Binomial Regression with Random Effects: *A Case Study Biometrics.* **56**, 1030–1039.

[22] Hamura, Y., Irie, K., Sugasawa, S. (2021). On global-local shrinkage priors for count data. *Bayesian Analysis*, 1(**1**), 1-20.

[23] Heilbron, D.C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal.* **36**, 347–531.

[24] Hu, M.C., Pavlicova, M. and nunes, E.V. (2011). Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. *The American Journal of Drug and Alcohol Abuse.* **37**, 367–375.

[25] Ishwaran, H. and Rao, J.S. (2005). Spike and slab variable selection. frequentist and Bayesian strategies. *Ann.Statist.* **33**, 730–773.

[26] Ji Y, Shi H. (2020). Bayesian variable selection in linear quantile mixed models for longitudinal data with application to macular degeneration. *PLoS One*. 15(**10**).

[27] Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*. 34 (**1**), 1–13.

[28] Lee, AH., Wang, K., Scott, JA., YauK, K., McLachlan, GJ. (2006). Multilevelzero-inflatedPoisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research*. 15(**1**), 47–61.

[29] Lee, K. HA., Coull, B. A., Moscicki, A. B., Paster, B. J. and Starr, J. R. (2020). Bayesian variable selection for multivariate zero-inflated models: Application to microbiome count data. *Biostatistics*. **21**. 499–517.

[30] Luna, PN., Mansbach, JM., Shaw, CA. (2020). A joint modeling approach for longitudinal microbiome data improvesability to detect microbiome associations with disease. *PLOS Computational Biology*. 16 (**12**), e1008473.

[31] Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. *Amer. Stat. Assoc.* **83**, 1023–1036.

[32] Miao, Y., Kook, J. H., Lu, Y., Guindani, M. and Vannucci, M. (2020). Scalable Bayesian variable selection regression models for count data. In Flexible Bayesian Regression Modelling, Yanan F., Smith M., Nott D. and Dortet-Bernadet J.-L.(Eds). Elsevier, 187–219.

[33] Murat, M. and Szynal, D. (1998). non-zero inflated modified power series distributions. *Communications in Statistics*. **12**, 3047–3064.

[34] Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modeling*. **5**, 1–19.

[35] Neelon, B.H., OMalley, A.J. and normand, S.L. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use Statistical Modelling. **10**, 421–439.

[36] Patil, M. K. and Shirke, D. T. (2007). Testing parameter of the power series distribution of a zero inflated power series model. *Stat. Methodol.* **16**, 393–406.

[37] Preisser, J. S., Das, K., Long, D. L., and Divaris, K. (2016). Marginalized zero-inflated negative binomial regression with application to dental caries. *Statistics in medicine*, 35(**10**), 1722–1735.

[38] Rose, C. E., Martin, S.W., Wannemuehler, K. A. and Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics*. **16**, 463–481.

[39] Song, P.X.K. (2007). Correlated Data Analysis, *Springer-Verlag,* New York.

[40] Tokuda, T., Goodrich, B., Van Mechelen, I., and Gelman, A. (2011). Visualizing Distributions of Covariance Matrices. *Columbia University, New York, USA,* Technical Report. 18–18.

[41] Xu, X. and Ghosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Anal*. **10**(4), 909-936.

[42] Young, D. S., Roemmele, E. S., and Yeh, P. (2021). Zero-inflated modeling part I: Traditional zero-inflated count regression models, their applications, and computational tools. *WIREs Computational Statistics*.

[43] Young, D. S., Roemmele, E. S., and Shi, X. (2021). Zero-inflated modeling part II : Zero-inflated models for complex data structures. *WIREs Computational Statistics.*

[44] Zeng, P., Wei, Y., and Zhao,Y. (2014). Variable Selection Approach for Zero-Inflated Count Data via Adaptive Lasso. *Journal of Applied Statistics.* **4**, 879–894.

[45] Zhang , X., Pei, YF., Zhang, L., Guo, B., Pendegraft, AH., Zhuang, W., Yi, N. (2018). Negative Binomial mixed models for analyzing longitudinal microbiome data. *Frontiersin microbiology.* **26**, 9–1683.