JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2021.1926459

Bayesian auxiliary variable model for birth records data with qualitative and quantitative responses

Xiaoning Kang[a*], Shyam Ranganathan [b*], Lulu Kang [c], Julia Gohlke [d] and Xinwei Deng [b]

[a]International Business College and Institute of Supply Chain Analytics, Dongbei University of Finance and Economics, Dalian, People's Republic of China; [b]Department of Statistics, Virginia Tech, Blacksburg, VA, USA; [c]Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL, USA; [d]Department of Population Health Sciences, Virginia Tech, Blacksburg, VA, USA

ABSTRACT

Many applications involve data with qualitative and quantitative responses. When there is an association between the two responses, a joint model will produce improved results than fitting them separately. In this paper, a Bayesian method is proposed to jointly model such data. The joint model links the qualitative and quantitative responses and can assess their dependency strength via a latent variable. The posterior distributions of parameters are obtained through an efficient MCMC sampling algorithm. The simulation is conducted to show that the proposed method improves the prediction capacity for both responses. Further, the proposed joint model is applied to the birth records data acquired by the Virginia Department of Health for studying the mutual dependence between preterm birth of infants and their birth weights.

ARTICLE HISTORY
Received 16 July 2020
Accepted 2 May 2021

KEYWORDS
Bayesian model; latent variable; MCMC sampling; quantitative and qualitative responses

1. Introduction

In many applications, mutually dependent quantitative and qualitative (QQ) types of outcome data are simultaneously observed. It is important to jointly model them to make accurate estimation and inferences, which provide scientific and meaningful conclusions. In this paper, our application focuses on a birth records study examining the mutual dependency of birth weight and preterm birth. The birth weight of the infant, a quantitative outcome, is an important variable that doctors need to monitor [1]. The average birth weight of healthy infants is about 3.5 kg. Children with low birth weight are more likely to have complications soon after birth and later in life, compared to children with normal birth weights [2,3]. The birth weight is known to be related to another key variable, *preterm birth*, a qualitative outcome whose value is set to be 1 if an infant is born before 36 gestational weeks and is 0 otherwise. Several factors are related to low birth weight and preterm birth. Such factors include the socio-economic and health status of the mother, stresses caused by the environment, etc. [4,5]. Naturally, a preterm born infant is more likely to

CONTACT Lulu Kang lkang2@iit.edu
*Xiaoning Kang and Shyam Ranganathan share the first authorship due to equal contribution.

suffer from low birth weight, and the two are highly correlated. Both preterm birth and low birth weight are rare outcomes in the population, accounting for less than 10% of all live births. Meanwhile, they have a significant effect on the health of the population, as well as the economy in general, due to the expenses spent on caring and monitoring infants [6].

Modernized maternal care is designed to provide personalized health care to mothers and children. It is important to understand how various factors affect both preterm birth and low birth weight. A statistical model that accurately predicts both quantitative and qualitative outcomes may offer useful information to health practitioners and expectant mothers. Many other applications are in need of such a joint model for quantitative and qualitative responses. For example, in [7,8], the total thickness variation (continuous) and the site total indicator reading (binary) are both measured to evaluate the quality of the wafer after the lapping stage in the wafer manufacturing process. In [9], survey data with both quantitative scores and categorical answers are jointly analysed. More applications and methodologies on the mixed types of quantitative and qualitative response data are reviewed in Section 2.

In this article, a Bayesian hierarchical model is developed for the mixed quantitative and qualitative types of responses. A latent variable is introduced to connect the two types of responses, which is similar to the joint model in [10]. This joint model is suitable for the data on the birth records study described previously. It is much simpler than the joint models by Dunson [11] and yet still sufficiently effective. In [10], the joint distribution is factorized into two regression components – the marginal distribution of the continuous outcome and the conditional distribution of the binary outcome conditioned on the continuous outcome. The latter is obtained through the latent variable, which is correlated with the observed continuous outcome. Based on the factorization, the estimation is done in two steps. The first step is to estimate the marginal regression model of the continuous outcome. The second step is to estimate the probit regression model of the binary outcome conditioned on the continuous outcome. The generalized estimating equations approach is used to obtain the estimation. Different from Catalano and Ryan [10], this work incorporates the Bayesian framework, assumes the proper prior and hyperprior distributions, derives the posterior distributions, and then develops the MCMC sampling procedure to obtain the posterior distributions. Compared to the frequentist approach in [10], there are some merits with the Bayesian approach. First, the posterior distribution of the latent variable is available. Second, the Bayesian inference is more accurate since it is not based on the asymptotic distribution as in maximum likelihood estimation. Third, sparsity on both regression models of the two outcomes is induced due to the informative prior distributions, which are assumed for the regression coefficients.

The remainder of the article is outlined as follows. Section 2 provides a literature of recent work on modelling quantitative–qualitative responses. Section 3 introduces the joint quantitative-qualitative model via latent variable within the Bayesian framework, as well as the full-conditional distributions of the parameters and the leave-one-out conditional posterior distribution of the latent variable. Section 4 lays out the MCMC sampling procedures. Numerical study and the case study in birth records are provided in Sections 5 and 6 to illustrate the merits of the proposed model. This article concludes in Section 7.

2. Literature review

Some works in the literature have tackled the issue of mixed quantitative and qualitative types of outcomes. Some of them, such as Wang and Tsung [12], Liu and Huang [13], Cheng et al. [14], Zhou et al. [15] and Shi [16], modelled the two types of responses separately. They overlooked the possible association that may exist between the two types of responses. As a result, if there exists a dependency between the two types of responses, separate modelling could lead to less accurate prediction and misinterpretation compared to the joint models. Most other works are on joint models for mixed types of outcomes, including [7–11,17–21]. Some interesting practical application cases can be found in [15,22–25].

These works can be further categorized into different groups. From the perspective of estimation methods, they can be divided into Bayesian methods, such as [8,11,18], and non-Bayesian methods, for example, [7,9,10,20]. Depending on the form of the joint model, Fitzmaurice and Laird [19], Deng and Jin [7] and Kang et al. [8] considered modelling the quantitative response conditioned on the qualitative response, leading to conditional linear regression models and marginal classification models, whereas other methods such as [9–11,18,20] used a latent variable to link the quantitative and qualitative outcomes.

Some representatives of the latent variable models are highlighted below. Motivated to analyse a toxicity experiment, Catalano and Ryan [10] used a latent variable to obtain a joint distribution of mixed responses. The joint distribution is a product of a linear regression model for the quantitative variable and a probit model for the qualitative variable. Dunson [11] suggested using the generalized linear models to describe the joint distribution of variables and proposed a Markov chain Monte Carlo (MCMC) sampling algorithm for estimating the posterior distributions of the parameters. Dunson [18] extended the previous work to multidimensional longitudinal data. However, such early methods focus on model estimation without investigating a sparse and interpretable model. Different from Catalano and Ryan [10], the latent variables in [11,18] appear in the generalized linear model as the linear coefficients. But in [10], the latent variable is used to define the probit model for the binary outcome.

Deng and Jin [7] proposed the QQ model for joint fitting quantitative and qualitative responses by the maximum likelihood estimation and identified the significant variables by imposing non-negative garrotte constraints on the likelihood function. The likelihood of the joint QQ model is the product of the conditional distribution of the quantitative responses conditioned on the qualitative responses and the marginal distribution of the qualitative responses. The authors also developed an iterative algorithm to solve the constrained optimization problem. Consequently, the classic asymptotical distribution of the maximum likelihood estimation cannot be easily applied, hence, making it difficult to conduct statistical inference. Using the same QQ model in [7] as the sampling distribution of the data, Kang et al. [8] introduced a sparse hierarchical Bayesian framework, which can easily provide statistical inference on the estimated parameters and prediction of the QQ model. However, since they constructed their model by fitting the quantitative response conditioned on the qualitative response, Deng and Jin [7] and Kang et al. [8] appeared to improve the prediction accuracy for the quantitative response, while the model of qualitative response would be similar as it was modelled independently of the quantitative response.

3. Bayesian QQ model with a latent variable

3.1. Sampling distribution

Denote the observed data as $(\boldsymbol{x}_i, y_i, z_i), i = 1, \ldots, n$, where $y_i \in \mathbb{R}$ and $z_i \in \{0, 1\}$ are the continuous and binary observations, respectively. Here the vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$ contains p predictors (intercept is included if needed). To jointly model the mixed types of responses Y and Z given $\boldsymbol{x}$, the key is to describe the association between them. Hence, a latent variable U for the binary response Z is introduced to facilitate this task. Assume the binary response follows the Bernoulli distribution

$$Z = \begin{cases} 1, & \text{if } U \geq 0 \\ 0, & \text{else if } U < 0 \end{cases} \qquad \text{with } U|\boldsymbol{\beta}_1, \boldsymbol{x} \sim N(\boldsymbol{x}'\boldsymbol{\beta}_1, 1). \tag{1}$$

This kind of latent variable approach is also used in cases other than the mixed types of outcomes. For example, Holmes and Held [26] used an auxiliary variable in Bayesian binary and multinomial regression. Regarding the quantitative response Y, its marginal distribution is assumed to be

$$Y|\boldsymbol{\beta}_2, \sigma^2, \boldsymbol{x} \sim N(\boldsymbol{x}'\boldsymbol{\beta}_2, \sigma^2). \tag{2}$$

To link the continuous and binary responses, a joint distribution of (U, Y) is introduced, with an assumption of a bivariate normal distribution for parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2, \rho)$ as follows:

$$\begin{bmatrix} U \\ Y \end{bmatrix} \Bigg| \boldsymbol{\theta}, \boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{with } \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{x}'\boldsymbol{\beta}_1 \\ \boldsymbol{x}'\boldsymbol{\beta}_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix}.$$

If ρ is positive, meaning that Y and the probability of $Z = 1$ is positively correlated, then the larger the value of Y the more likely that Z would be equal to 1. Therefore, to conclude the association between Y and Z, the key is to estimate ρ and make inference on the estimation.

3.2. Full-conditional distributions

In this part, the joint posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{X})$ is derived, where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{z} = (z_1, \ldots, z_n)'$ and $\boldsymbol{X}$ is the model matrix of the regression with each row as $\boldsymbol{x}_i'$. Based on the model assumption in Section 3.1, the joint distribution of (Y, Z, U) can be directly written as follows, given a single point of input $\boldsymbol{x}$:

$$\begin{aligned} p(z = 1, y, u|\boldsymbol{\theta}, \boldsymbol{x}) &= \Pr(Z = 1|U = u)p(u, y|\boldsymbol{\theta}, \boldsymbol{x}) = I(u \geq 0)p(u, y|\boldsymbol{\theta}, \boldsymbol{x}), \\ p(z = 0, y, u|\boldsymbol{\theta}, \boldsymbol{x}) &= \Pr(Z = 0|U = u)p(u, y|\boldsymbol{\theta}, \boldsymbol{x}) = I(u < 0)p(u, y|\boldsymbol{\theta}, \boldsymbol{x}), \\ p(z, y, u|\boldsymbol{\theta}, \boldsymbol{x}) &= [zI(u \geq 0) + (1 - z)I(u < 0)]p(u, y|\boldsymbol{\theta}, \boldsymbol{x}). \end{aligned}$$

The joint sampling distribution of the two response variables is

$$p(z, y|\boldsymbol{\theta}, \boldsymbol{x}) = (1 - z)p(y|\boldsymbol{\theta}, \boldsymbol{x}) + (2z - 1) \int I(u \geq 0)p(u, y|\boldsymbol{\theta}, \boldsymbol{x}) \, \mathrm{d}u.$$

To obtain the exact form of $\int I(u \geq 0)p(u, y|\boldsymbol{\theta}, \boldsymbol{x}) \, \mathrm{d}u$, the probability density $p(u, y|\boldsymbol{\theta}, \boldsymbol{x})$ is rewritten into $p(u|y, \boldsymbol{\theta}, \boldsymbol{x})p(y|\boldsymbol{\theta}, \boldsymbol{x})$. Based on the bivariate normal distribution of (U, Y),

the distribution $U|Y = y$ is

$$U|y, \boldsymbol{\theta}, \boldsymbol{x} \sim N\left(\boldsymbol{x}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma}(y - \boldsymbol{x}'\boldsymbol{\beta}_2), (1-\rho^2)\right).$$

Hence, it is easy to obtain the following

$$\begin{aligned}\int I(u \geq 0)p(u, y|\boldsymbol{\theta}, \boldsymbol{x})\,\mathrm{d}u &= \int I(u \geq 0)p(u|y, \boldsymbol{\theta}, \boldsymbol{x})p(y|\boldsymbol{\theta}, \boldsymbol{x})\,\mathrm{d}u \\ &= p(y|\boldsymbol{\theta}, \boldsymbol{x})\Phi\left(\frac{\boldsymbol{x}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma}(y - \boldsymbol{x}'\boldsymbol{\beta}_2)}{\sqrt{(1-\rho^2)}}\middle| y, \boldsymbol{\theta}, \boldsymbol{x}\right),\end{aligned}$$

where $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal random variable. To simplify the notation, define

$$s(y|\boldsymbol{\theta}, \boldsymbol{x}) = \frac{\boldsymbol{x}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma}(y - \boldsymbol{x}'\boldsymbol{\beta}_2)}{\sqrt{(1-\rho^2)}}.$$

Consequently, the joint distribution of (Z, Y) can be written as

$$p(z, y|\boldsymbol{\theta}, \boldsymbol{x}) = p(y|\boldsymbol{\theta}, \boldsymbol{x})\left[(1 - \Phi(s(y)|\boldsymbol{\theta}, \boldsymbol{x})) + z(2\Phi(s(y)|\boldsymbol{\theta}, \boldsymbol{x}) - 1)\right],$$

or more explicitly

$$p(z = 1, y|\boldsymbol{\theta}, \boldsymbol{x}) = p(y|\boldsymbol{\theta}, \boldsymbol{x})\Phi(s(y)|\boldsymbol{\theta}, \boldsymbol{x}), \quad p(z = 0, y|\boldsymbol{\theta}, \boldsymbol{x}) = p(y|\boldsymbol{\theta}, \boldsymbol{x})(1 - \Phi(s(y)|\boldsymbol{\theta}, \boldsymbol{x})).$$

The conditional distribution of the latent variable $U|z, y, \boldsymbol{\theta}, \boldsymbol{x}$ is

$$\begin{aligned}p(u|z, y, \boldsymbol{\theta}, \boldsymbol{x}) &= \frac{p(z, y, u|\boldsymbol{\theta}, \boldsymbol{x})}{p(z, y|\boldsymbol{\theta}, \boldsymbol{x})} = \frac{[zI(u \geq 0) + (1-z)I(u < 0)]p(u, y|\boldsymbol{\theta}, \boldsymbol{x})}{[(1-\Phi) + z(2\Phi - 1)]p(y|\boldsymbol{\theta}, \boldsymbol{x})} \\ &= p(u|y, \theta, \boldsymbol{x})\frac{(1 - I(u \geq 0)) + z(2I(u \geq 0) - 1)}{(1-\Phi) + z(2\Phi - 1)}.\end{aligned}$$

In the above equation, Φ stands for $\Phi(s(y)|\boldsymbol{\theta}, \boldsymbol{x})$. Write the conditional distribution separately and obtain

$$\begin{aligned}p(u|y, z = 1, \boldsymbol{\theta}, \boldsymbol{x}) &= N\left(u|\boldsymbol{x}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma}(y - \boldsymbol{x}'\boldsymbol{\beta}_2), (1-\rho^2)\right)\frac{I(u \geq 0)}{\Phi(s(y)|\boldsymbol{\theta}, \boldsymbol{x})}, \\ p(u|y, z = 0, \boldsymbol{\theta}, \boldsymbol{x}) &= N\left(u|\boldsymbol{x}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma}(y - \boldsymbol{x}'\boldsymbol{\beta}_2), (1-\rho^2)\right)\frac{1 - I(u \geq 0)}{1 - \Phi(s(y)|\boldsymbol{\theta}, \boldsymbol{x})}.\end{aligned}$$

Clearly, the latent variable U, given the two response variables and the parameters, follows two different truncated normal distributions.

Since the outputs of different experimental runs are independent of each other, the sampling distribution for all the data $\{\boldsymbol{x}_i, z_i, y_i\}_{i=1}^n$ is

$$\begin{aligned} p(\boldsymbol{z}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X}) &= \prod_{i=1}^{n} p(z_i, y_i|\boldsymbol{\theta}, \boldsymbol{x}_i) \\ &= \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}, \boldsymbol{x}_i)[(1 - \Phi(s(y_i)|\boldsymbol{\theta}, \boldsymbol{x}_i)) + z_i(2\Phi(s(y_i)|\boldsymbol{\theta}, \boldsymbol{x}_i) - 1)] \\ &= N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}_2, \sigma^2\boldsymbol{I}_n) \prod_{z_i=1} \Phi(s(y_i)|\boldsymbol{\theta}, \boldsymbol{x}_i) \prod_{z_i=0} \left(1 - \Phi(s(y_i)|\boldsymbol{\theta}, \boldsymbol{x}_i)\right). \end{aligned}$$

The conditional distribution for $\boldsymbol{u} = (u_1, \ldots, u_n)'$ is

$$\begin{aligned} &p(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{X}) \\ &\quad = N\left(\boldsymbol{u}|\boldsymbol{X}\boldsymbol{\beta}_1 + \frac{\rho}{\sigma}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_2), (1 - \rho^2)\boldsymbol{I}_n\right) \prod_{z_i=1} \frac{I(u_i \geq 0)}{\Phi(s(y_i)|\boldsymbol{\theta}, \boldsymbol{x}_i)} \prod_{z_i=0} \frac{1 - I(u_i \geq 0)}{1 - \Phi(s(y_i)|\boldsymbol{\theta}, \boldsymbol{x}_i)}. \end{aligned}$$

To obtain the joint posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{X})$, we first derive the conditional distribution $p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, , \sigma^2, \rho, \boldsymbol{X})$. To simplify the notation, the matrix $\boldsymbol{X}$ in all the conditional side of the distribution is omitted, as $\boldsymbol{X}$ is always considered to be known in all Bayesian regression modelling. The previous regression model for $(\boldsymbol{u}, \boldsymbol{y})$ is written via the following matrix form

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix}_{2n\times 1} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where } \mathbb{X} = \boldsymbol{I}_2 \otimes \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X} \end{bmatrix}_{2n\times 2p} \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}_{2p\times 1}.$$

Here $\boldsymbol{I}_2$ is a 2×2 identity matrix and the symbol $\otimes$ stands for the Kronecker product between two matrices. The covariance matrix of the noise $\boldsymbol{\epsilon}$ as well as the vector $(\boldsymbol{u}', \boldsymbol{y}')'$ is

$$\boldsymbol{\Sigma}_\epsilon = \text{cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n = \begin{pmatrix} \boldsymbol{I}_n & \rho\sigma\boldsymbol{I}_n \\ \rho\sigma\boldsymbol{I}_n & \sigma^2\boldsymbol{I}_n \end{pmatrix},$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Denote the conjugate prior for $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_0), \quad \text{where } \boldsymbol{\Sigma}_0 = \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_2 \end{bmatrix}. \tag{3}$$

Such prior covariance matrix assumes that $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are independent. Consequently, the full-conditional distribution of $\boldsymbol{\beta}$ is as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \Bigg| \boldsymbol{y}, \boldsymbol{u}, \sigma^2, \rho \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \tag{4}$$

where the covariance is

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (\boldsymbol{\Sigma}_0^{-1} + \mathbb{X}'\boldsymbol{\Sigma}_\epsilon^{-1}\mathbb{X})^{-1} = (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{5}$$

The inverse matrix $\boldsymbol{\Sigma}^{-1}$ is easily computed by

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix}^{-1} = \frac{1}{(1-\rho^2)\sigma^2} \begin{bmatrix} \sigma^2, & -\rho\sigma, \\ -\rho\sigma, & 1 \end{bmatrix}.$$

The mean of the full conditional is

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\beta}} &= (\boldsymbol{\Sigma}_0^{-1} + \mathbb{X}'\boldsymbol{\Sigma}_\epsilon^{-1}\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\Sigma}_\epsilon^{-1} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix} = (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}') \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix} \\ &= \frac{1}{(1-\rho^2)\sigma^2}(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}'\boldsymbol{X})^{-1} \begin{bmatrix} \sigma^2\boldsymbol{X}'\boldsymbol{u} - \rho\sigma\boldsymbol{X}'\boldsymbol{y} \\ -\rho\sigma\boldsymbol{X}'\boldsymbol{u} + \boldsymbol{X}'\boldsymbol{y} \end{bmatrix}. \end{aligned} \tag{6}$$

As regards the priors for the parameters σ^2 and ρ, a weekly informative prior for σ^2 is adopted as $\sigma^2 \sim \text{Inv}{-}\chi^2(0.001, 0.001)$, and the uniform prior for ρ is employed, i.e. $p(\rho) \sim \text{Unif}(-1, 1)$. Let $\eta_i = u_i - \boldsymbol{x}_i'\boldsymbol{\beta}_1$ and $\varphi_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_2$, then the posterior distributions of σ^2 and ρ are easily derived as

$$\begin{aligned} p(\sigma^2|\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\beta}, \rho) &\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\eta_i, \varphi_i)\boldsymbol{\Sigma}^{-1}(\eta_i, \varphi_i)'\right\} \frac{1}{(\sigma^2)^{\frac{2+0.001}{2}}} \exp\left\{-\frac{10^{-6}}{2\sigma^2}\right\} \\ &\propto \frac{1}{(\sigma^2)^{\frac{n+2+0.001}{2}}} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{1}{1-\rho^2}\sum_{i=1}^{n}(\sigma^2\eta_i^2 - 2\rho\sigma\varphi_i\eta_i + \varphi_i^2) + 10^{-6}\right]\right\}, \end{aligned} \tag{7}$$

and

$$\begin{aligned} p(\rho|\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\beta}, \sigma) &\propto \frac{1}{(1-\rho^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\eta_i, \varphi_i)\boldsymbol{\Sigma}^{-1}(\eta_i, \varphi_i)'\right\} \\ &\propto \frac{1}{(1-\rho^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2(1-\rho^2)}\sum_{i=1}^{n}(\sigma^2\eta_i^2 - 2\rho\sigma\varphi_i\eta_i + \varphi_i^2)\right\}. \end{aligned} \tag{8}$$

Since their posteriors are not from any known distributions, the Metropolis-Hasting (MH) algorithm is used to draw the samples of σ^2 and ρ.

3.3. Leave-one-out sampling of $\boldsymbol{u}$

One might think that the simplest way to sample from $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z})$ is to use Gibbs sampling that draws $\boldsymbol{\theta}$ and $\boldsymbol{u}$ iteratively in the following steps:

(1) $\boldsymbol{u}_j \leftarrow p(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_{j-1}, \sigma_{j-1}^2, \rho_{j-1})$,
(2) $\boldsymbol{\beta}_j \leftarrow p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{u}_j, \sigma_{j-1}^2, \rho_{j-1})$,
(3) $\sigma_j^2 \leftarrow p(\sigma^2|\boldsymbol{y}, \boldsymbol{u}_j, \boldsymbol{\beta}_j, \rho_{j-1})$
(4) $\rho_j \leftarrow p(\rho|\boldsymbol{y}, \boldsymbol{u}_j, \boldsymbol{\beta}_j, \sigma_j^2)$.

However, as discussed in [26], a potential problem lurks in the strong posterior correlation between $\boldsymbol{\beta}_1$ and $\boldsymbol{u}$, as assumed in the model $\boldsymbol{u}|\boldsymbol{\theta} \sim N(\boldsymbol{X}\boldsymbol{\beta}_1, \boldsymbol{I}_n)$. This strong correlation would cause slow mixing in the MCMC chain and thus lead to large computation. Instead, we follow the approach suggested by Holmes and Held [26] and update $\boldsymbol{\beta}$ and $\boldsymbol{u}$ jointly by making the factorization

$$p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{y}, \boldsymbol{z}, \sigma^2, \rho) = p(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{z}, \sigma^2, \rho)p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{u}, \sigma^2, \rho).$$

The distribution of $\boldsymbol{\beta}|(\boldsymbol{y}, \boldsymbol{u}, \sigma^2, \rho)$ is the normal distribution in (4). The distribution of $\boldsymbol{u}|(\boldsymbol{y}, \boldsymbol{z}, \sigma^2, \rho)$ can be obtained by integrating $p(\boldsymbol{\beta})p(\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}, \sigma^2, \rho)$ with respect to $\boldsymbol{\beta}$. Given the prior of $\boldsymbol{\beta}$ in (3), one can obtain

$$\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{z}, \sigma^2, \rho \sim \boldsymbol{N}\left(\frac{\rho}{\sigma}\boldsymbol{y}, (1-\rho^2)\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{V}_1\boldsymbol{X}' + \frac{\rho^2}{\sigma^2}\boldsymbol{X}\boldsymbol{V}_2\boldsymbol{X}'\right) Ind(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}),$$

where $Ind(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u})$ is an indicator function that truncates the multivariate normal distribution into the appropriate region. It is well known that directly sampling from a truncated multivariate normal distribution is difficult, as pointed out by Holmes and Held [26]. Hence, a more straightforward Gibbs sampling method is used as

$$u_i|\boldsymbol{u}_{-i}, \boldsymbol{y}, z_i, \sigma^2, \rho \sim \begin{cases} N(m_i, v_i)I(u_i \geq 0), & \text{if } z_i = 1, \\ N(m_i, v_i)I(u_i < 0), & \text{if } z_i = 0, \end{cases}$$

where $\boldsymbol{u}_{-i}$ denotes all the latent variables $\boldsymbol{u}$ without u_i. The mean m_i and variance v_i for $i = 1, \ldots, n$ are obtained from the leave-one-out marginal predictive distributions,

$$m_i = \frac{\rho}{\sigma}y_i + \left[\boldsymbol{x}_i', -\frac{\rho}{\sigma}\boldsymbol{x}_i'\right]\boldsymbol{\mu}_{\boldsymbol{\beta},-i},$$

$$v_i = \left[\boldsymbol{x}_i', -\frac{\rho}{\sigma}\boldsymbol{x}_i'\right]\boldsymbol{\Sigma}_{\boldsymbol{\beta},-i}\begin{bmatrix} \boldsymbol{x}_i, \\ -\frac{\rho}{\sigma}\boldsymbol{x}_i \end{bmatrix} + (1-\rho^2).$$

Its detailed derivation is provided in Appendix 1. The notations $\boldsymbol{\mu}_{\boldsymbol{\beta},-i}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta},-i}$ are the mean and covariance matrices of the distribution of $\boldsymbol{\beta}|(\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho)$. Since these two need to be calculated frequently, a shortcut formula is derived to facilitate the computation in Appendix 2.

4. MCMC sampling

In this section, the prior distribution is specified for the parameter $\boldsymbol{\beta}$ as well as hyperprior distributions for the hyperparameters r_1, r_2, τ_1^2, τ_2^2. Then the corresponding posteriors of these parameters are obtained. The Gibbs sampling algorithm is laid out to sample the posterior distributions.

4.1. Prior and hyperprior distributions

The marginal prior components for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are

$$\boldsymbol{\beta}_i \sim N(\boldsymbol{0}, \tau_i^2\boldsymbol{R}_i) \quad \text{for } i = 1, 2. \tag{9}$$

The correlation matrices in (9) in the marginal prior components for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are assumed to be diagonal, which means that the coefficients are independent of each other.

This assumption is reasonable if the orthogonal polynomial basis of $\boldsymbol{x}$ is used, consisting of the intercept, the linear effects, the quadratic effects, and the interactions, etc., up to a user-specified order. If the controllable variable settings are from a full factorial design or an orthogonal design, the full or near orthogonality between the bases can be achieved. For the bases involving covariates, it is not likely to achieve full or near orthogonality. But we still assume independence for simplicity and leave the data to correct it in the posterior distribution. Let $\boldsymbol{R}_i = \text{diag}\{1, r_i, \ldots, r_i, r_i^2, \ldots, r_i^2, \ldots\}$ for $i = 1, 2, 3$, where $r_i \in (0, 1)$ is a user-specified tuning parameter. The power index of r_i is the same as the order of the corresponding polynomial term. For example, if the polynomial regression terms of $\boldsymbol{x} \in \mathbb{R}^2$ is a full quadratic model and contains the term $\{1, x_1, x_2, x_1^2, x_2^2, x_1x_2\}$, the corresponding prior correlation matrix should specified as $\boldsymbol{R} = \text{diag}\{1, r, r, r^2, r^2, r^2\}$. In this way, the prior variance of the effect is decreasing exponentially as the order of effect increases, following the hierarchy ordering principle defined in [27]. The hierarchy ordering principle can reduce the size of the model and avoid including higher order and less significant model terms. Such prior distribution was firstly proposed by Joseph [28], and later used in [8,29,30].

Additionally, the hyperprior distributions for the hyperparameters $\tau_1^2, \tau_2^2 \sim_{iid} \text{Inv}{-}\chi^2(\nu, \delta^2)$ and $r_1, r_2 \sim_{iid} \text{Beta}(a, b)$ are used in this work, where $\text{Inv}{-}\chi^2(\nu, \delta^2)$ stands for the scaled inverse-chi-square distribution with ν degrees of freedom and scale δ^2. Beta distribution is a reasonable prior for r_i since $r_i \in (0, 1)$. Accordingly, it is not difficult to derive the posterior distributions for r_1, r_2, τ_1^2 and τ_2^2 listed below

$$\tau_1^2|\text{rest parameters}, \boldsymbol{y}, \boldsymbol{z} \sim \text{Inv}{-}\chi^2\left(\nu + p, \frac{1}{\nu + p}[\boldsymbol{\beta}_1'\boldsymbol{R}_1^{-1}\boldsymbol{\beta}_1 + \nu\delta^2]\right), \tag{10}$$

$$\tau_2^2|\text{rest parameters}, \boldsymbol{y}, \boldsymbol{z} \sim \text{Inv}{-}\chi^2\left(\nu + p, \frac{1}{\nu + p}[\boldsymbol{\beta}_2'\boldsymbol{R}_2^{-1}\boldsymbol{\beta}_2 + \nu\delta^2]\right), \tag{11}$$

$$p(r_1|\text{rest parameters}, \boldsymbol{y}, \boldsymbol{z}) \propto |\boldsymbol{R}_1|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\tau_1^2}\boldsymbol{\beta}_1'\boldsymbol{R}_1^{-1}\boldsymbol{\beta}_1\right\} r_1^{a-1}(1 - r_1)^{b-1}, \tag{12}$$

$$p(r_2|\text{rest parameters}, \boldsymbol{y}, \boldsymbol{z}) \propto |\boldsymbol{R}_2|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\tau_2^2}\boldsymbol{\beta}_2'\boldsymbol{R}_2^{-1}\boldsymbol{\beta}_2\right\} r_2^{a-1}(1 - r_2)^{b-1}. \tag{13}$$

The posterior samples of τ_1^2 and τ_2^2 are drawn directly from their respective scaled inverse-chi-square distributions, and the Metropolis–Hastings (MH) algorithm is applied to sample r_1 and r_2 from (12) and (13).

4.2. Gibbs sampling algorithm

The following Gibbs sampling algorithm is employed to generate the posterior distributions for the (hyper)parameters and the latent variable.

Step 0 Set up the initial values for the parameters and the latent variable. Set the counter $j = 0$. For the counter $j = 1, 2, \ldots, B$.

Step 1 Sample $\boldsymbol{u}_j$ from $p(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{z}, \sigma_{j-1}^2, \rho_{j-1})$ by drawing $u_{i,j}$ from the leave-one-out marginal distribution $p(u_i|\boldsymbol{u}_{-i,j-1}, \boldsymbol{y}, z_i, \sigma_{j-1}^2, \rho_{j-1})$ for $i = 1, \ldots, n$.

Step 2 Sample $\boldsymbol{\beta}_j$ from $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{u}_j, \sigma_{j-1}^2, \rho_{j-1})$ according to (5) and (6).

Step 3 Sample σ_j^2 and ρ_j from (7) and (8) by the MH algorithm.

Step 4 Sample $\tau_{1,j}^2$ and $\tau_{2,j}^2$ from (10) and (11).
Step 5 Sample $r_{1,j}$ and $r_{2,j}$ by the MH algorithm from distributions (12) and (13).
Step 6 Do Step 1–Step 5 until the MCMC chain converges.

The initial values of $\boldsymbol{\beta}_2$ are set to be the least square estimate from $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2$, and the initial σ^2 value is the mean squared error of the linear regression model. The initial values of $\boldsymbol{\beta}_1$ are the MLE of the probit regression of $\boldsymbol{z}$ with the same model matrix $\boldsymbol{X}$. The estimated link function values of the probit regression can be the initial values of $\boldsymbol{u}$. The initial value of ρ is calculated from the sample correlation between $\boldsymbol{u}$ and $\boldsymbol{y}$.

In Step 1, given the current $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, the short-cut formula in Appendix 2 is used to calculate $\boldsymbol{\Sigma}_{\boldsymbol{\beta},-i}$ and $\boldsymbol{\mu}_{\boldsymbol{\beta},-i}$. After each u_i is updated, the vector of $\boldsymbol{u}_{j-1}$ is updated to be $(u_{1,j}, \ldots, u_{i,j}, u_{i+1,j-1}, \ldots, u_{n,j-1})$. The covariance $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ remains the same for all i, but $\boldsymbol{\mu}_{\boldsymbol{\beta},-i}$ needs to be updated using $(u_{1,j}, \ldots, u_{i,j}, u_{i+1,j-1}, \ldots, u_{n,j-1})$.

5. Numerical study

In this section, the performance of the proposed model is examined and compared with two approaches SM(F) and SM(B), where the qualitative variable Z and quantitative variable Y are modelled separately. Hence, both SM(F) and SM(B) ignore the association between variables Z and Y. SM(F) employs a logistic model for the variable Z, and a linear regression model to fit Y. The LASSO regularization is applied for both logistic and linear regression models to select the significant variables. SM(B) denotes the separate modelling of Z using probit regression and of Y using linear regression under the Bayesian framework. SM(B) sets the marginal normal priors for the parameters in both linear and probit models.

Since the parameter ρ reflects the strength and direction of the relationship between the value of Y and the probability $Z = 1$, five different cases are considered: (1) $\rho = 0$; (2) $\rho = 0.3$; (3) $\rho = 0.85$; (4) $\rho = -0.3$; (5) $\rho = -0.85$. In each case, we generate $n = 100$ training data points and $n = 100$ testing data points based on models (1) and (2). All data are independently and identically distributed from normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_x = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$. The variance σ^2 in model (2) is set to be 2. To further examine the performance of the proposed model, different settings of model size $p \in \{10, 30\}$ and different proportions of sparsity $s \in \{20\%, 50\%\}$ are considered, where the value of s represents the proportion of nonzero entries in the parameter vector $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Overall, the full combinations have $5 \times 2 \times 2 = 20$ settings.

For the true values of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, firstly their zeroes are randomly placed. Then the values of non-zeroes are generated from $N(3, 1)$ independently, with positive signs and negative signs randomly assigned to the non-zeroes elements of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. To evaluate the estimation accuracy of each method with respect to $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, the following loss measures are used

$$L_2(\hat{\boldsymbol{\beta}}_1) = \|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_2^2 \quad \text{and} \quad L_2(\hat{\boldsymbol{\beta}}_2) = \|\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\|_2^2,$$

where $\|\cdot\|_2$ denotes the vector L_2 norm. Additionally, to gauge the performance of variable selection for $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, false-positive (FP) and false-negative (FN) cases are considered. An FP occurs if a nonsignificant predictor in the true model is incorrectly identified as a significant one. Similarly, an FN occurs if a significant predictor in the true model

Table 1. The averages and standard errors (in parenthesis) of loss measures when $p = 10$.

		BLQQ		SM(F)		SM(B)	
ρ		$s = 0.2$	$s = 0.5$	$s = 0.2$	$s = 0.5$	$s = 0.2$	$s = 0.5$
0	*RMSE*	0.315 (0.021)	0.469 (0.016)	0.327 (0.016)	0.459 (0.014)	0.425 (0.014)	0.473 (0.013)
	ME	0.044 (0.005)	0.044 (0.004)	0.054 (0.004)	0.059 (0.005)	0.082 (0.005)	0.103 (0.005)
	FSL	0.700 (0.115)	0.500 (0.112)	7.160 (0.365)	5.640 (0.209)	0.540 (0.087)	0.420 (0.099)
	$L_2(\hat{\boldsymbol{\beta}}_1)$	9.061 (2.372)	13.19 (1.580)	18.65 (7.801)	24.53 (8.101)	8.961 (2.963)	13.34 (2.664)
	$L_2(\hat{\boldsymbol{\beta}}_2)$	0.140 (0.019)	0.247 (0.021)	0.224 (0.025)	0.333 (0.028)	0.301 (0.020)	0.349 (0.022)
	$\hat{\rho}$	0.047 (0.038)	−0.017 (0.030)	–	–	–	–
0.3	*RMSE*	0.322 (0.019)	0.414 (0.020)	0.367 (0.014)	0.432 (0.016)	0.453 (0.014)	0.453 (0.017)
	ME	0.048 (0.004)	0.038 (0.003)	0.071 (0.004)	0.063 (0.004)	0.099 (0.005)	0.085 (0.004)
	FSL	0.760 (0.150)	0.460 (0.108)	6.840 (0.341)	5.360 (0.215)	0.380 (0.090)	0.540 (0.108)
	$L_2(\hat{\boldsymbol{\beta}}_1)$	9.411 (1.361)	8.229 (0.704)	9.937 (1.955)	24.56 (8.897)	15.10 (0.297)	17.58 (0.335)
	$L_2(\hat{\boldsymbol{\beta}}_2)$	0.139 (0.016)	0.239 (0.029)	0.239 (0.024)	0.345 (0.035)	0.329 (0.020)	0.362 (0.034)
	$\hat{\rho}$	0.296 (0.034)	0.261 (0.034)	–	–	–	–
0.85	*RMSE*	0.315 (0.018)	0.424 (0.020)	0.363 (0.022)	0.473 (0.019)	0.456 (0.015)	0.481 (0.019)
	ME	0.065 (0.004)	0.061 (0.004)	0.086 (0.004)	0.074 (0.004)	0.110 (0.005)	0.096 (0.005)
	FSL	0.680 (0.138)	0.740 (0.106)	5.540 (0.389)	5.680 (0.195)	0.300 (0.071)	1.160 (0.096)
	$L_2(\hat{\boldsymbol{\beta}}_1)$	3.789 (0.808)	7.425 (1.013)	37.68 (21.65)	28.64 (9.023)	10.74 (1.707)	15.72 (3.115)
	$L_2(\hat{\boldsymbol{\beta}}_2)$	0.130 (0.016)	0.267 (0.047)	0.197 (0.022)	0.378 (0.032)	0.324 (0.020)	0.395 (0.032)
	$\hat{\rho}$	0.749 (0.015)	0.785 (0.017)	–	–	–	–
−0.3	*RMSE*	0.370 (0.023)	0.425 (0.016)	0.410 (0.019)	0.437 (0.014)	0.460 (0.017)	0.466 (0.012)
	ME	0.057 (0.005)	0.055 (0.005)	0.099 (0.004)	0.076 (0.005)	0.125 (0.005)	0.101 (0.004)
	FSL	0.800 (0.125)	0.660 (0.133)	6.620 (0.362)	5.560 (0.227)	0.620 (0.114)	0.680 (0.119)
	$L_2(\hat{\boldsymbol{\beta}}_1)$	4.430 (0.487)	15.38 (1.267)	4.808 (0.540)	15.49 (2.665)	7.574 (0.266)	35.57 (0.369)
	$L_2(\hat{\boldsymbol{\beta}}_2)$	0.188 (0.025)	0.239 (0.021)	0.262 (0.030)	0.330 (0.022)	0.352 (0.029)	0.342 (0.021)
	$\hat{\rho}$	−0.341 (0.026)	−0.322 (0.035)	–	–	–	–
−0.85	*RMSE*	0.364 (0.023)	0.370 (0.016)	0.370 (0.017)	0.417 (0.015)	0.453 (0.015)	0.444 (0.012)
	ME	0.069 (0.004)	0.060 (0.003)	0.086 (0.005)	0.085 (0.004)	0.105 (0.004)	0.104 (0.005)
	FSL	0.600 (0.164)	0.300 (0.071)	6.020 (0.335)	5.620 (0.228)	0.560 (0.149)	0.360 (0.085)
	$L_2(\hat{\boldsymbol{\beta}}_1)$	3.744 (0.391)	11.18 (0.875)	4.931 (0.204)	28.12 (11.949)	4.464 (0.234)	24.91 (0.361)
	$L_2(\hat{\boldsymbol{\beta}}_2)$	0.182 (0.027)	0.204 (0.020)	0.254 (0.028)	0.305 (0.021)	0.336 (0.028)	0.313 (0.021)
	$\hat{\rho}$	−0.799 (0.011)	−0.693 (0.021)	–	–	–	–

is incorrectly estimated as a nonsignificant one. The loss $FSL = FP + FN$, which is the total number of FP and FN cases, is reported as the performance measure of variable selection. In the SM(F) method, the significant predictors are selected by the LASSO. For the proposed model and SM(B), the variable selection is conducted based on the 95% credible intervals constructed from the MCMC samples after the burn-in period. Furthermore, the model's prediction capacity is evaluated using the root-mean-square error $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ for the quantitative variable Y, where $\hat{y}_i$ is the predicted value for y_i in the testing data set. The misclassification error $ME = \frac{1}{n}\sum_{i=1}^{n} I_{(z_i \neq \hat{z}_i)}$ is used to measure the model's prediction performance on the qualitative variable Z, where $I_{(\cdot)}$ stands for the indicator function and $\hat{z}_i$ is the predicted value for z_i. For the proposed model, set $(\nu, \delta^2, a, b) = (2, 2, 0.1, 0.1)$ and initial values $(\tau_{1,0}^2, \tau_{2,0}^2, r_{1,0}, r_{2,0}) = (0.5, 0.5, 0.3, 0.3)$. The length of the MCMC chain is 10,000 with the first 1000 as the burn-in period. Tables 1 and 2 report the simulation results for each loss measure of estimates obtained from each approach over 50 replicates. Only the proposed approach (BLQQ column) shows the average and standard error (in the parenthesis) of the 50 replicates of the estimated $\hat{\rho}$.

From Tables 1 to 2, it is clear to see the following results.

Table 2. The averages and standard errors (in parenthesis) of loss measures when $p = 30$.

		BLQQ		SM(F)		SM(B)	
ρ		$s = 0.2$	$s = 0.5$	$s = 0.2$	$s = 0.5$	$s = 0.2$	$s = 0.5$
0	*RMSE*	0.647 (0.030)	0.804 (0.023)	0.633 (0.023)	0.842 (0.021)	0.896 (0.025)	0.936 (0.021)
	ME	0.087 (0.005)	0.161 (0.008)	0.094 (0.006)	0.148 (0.007)	0.130 (0.005)	0.144 (0.006)
	FSL	2.800 (0.206)	5.640 (0.298)	19.60 (0.648)	16.98 (0.483)	2.720 (0.216)	5.220 (0.332)
	$L_2(\hat{\boldsymbol{\beta}}_1)$	11.59 (0.847)	67.07 (4.089)	17.03 (3.243)	81.64 (3.192)	20.79 (0.389)	89.91 (0.911)
	$L_2(\hat{\boldsymbol{\beta}}_2)$	0.605 (0.055)	0.874 (0.052)	0.600 (0.046)	1.075 (0.059)	1.376 (0.068)	1.368 (0.060)
	$\hat{\rho}$	−0.027 (0.040)	0.017 (0.041)	–	–	–	–
0.3	*RMSE*	0.639 (0.031)	0.819 (0.025)	0.700 (0.024)	0.863 (0.024)	0.880 (0.018)	0.947 (0.021)
	ME	0.121 (0.006)	0.163 (0.008)	0.125 (0.005)	0.165 (0.007)	0.126 (0.006)	0.168 (0.005)
	FSL	3.720 (0.256)	5.130 (0.309)	17.28 (0.682)	16.72 (0.463)	4.480 (0.259)	9.087 (0.379)
	$L_2(\hat{\boldsymbol{\beta}}_1)$	19.39 (1.718)	44.02 (4.091)	19.93 (0.930)	54.50 (2.342)	19.26 (0.363)	56.38 (0.642)
	$L_2(\hat{\boldsymbol{\beta}}_2)$	0.527 (0.054)	0.965 (0.058)	0.742 (0.063)	1.331 (0.055)	1.343 (0.065)	1.438 (0.053)
	$\hat{\rho}$	0.356 (0.034)	0.359 (0.036)	–	–	–	–
0.85	*RMSE*	0.671 (0.028)	0.761 (0.019)	0.702 (0.031)	0.865 (0.019)	0.926 (0.025)	0.906 (0.020)
	ME	0.091 (0.006)	0.141 (0.006)	0.115 (0.005)	0.181 (0.005)	0.135 (0.005)	0.180 (0.005)
	FSL	2.100 (0.210)	5.020 (0.302)	17.86 (0.794)	16.14 (0.472)	3.320 (0.247)	10.12 (0.309)
	$L_2(\hat{\boldsymbol{\beta}}_1)$	24.35 (1.555)	64.11 (2.646)	26.51 (1.617)	79.71 (3.457)	43.79 (0.470)	90.35 (0.620)
	$L_2(\hat{\boldsymbol{\beta}}_2)$	0.583 (0.049)	0.768 (0.050)	0.769 (0.072)	1.160 (0.057)	1.507 (0.082)	1.278 (0.060)
	$\hat{\rho}$	0.801 (0.017)	0.737 (0.018)	–	–	–	–
−0.3	*RMSE*	0.654 (0.029)	0.833 (0.022)	0.640 (0.020)	0.851 (0.020)	0.894 (0.018)	0.936 (0.020)
	ME	0.098 (0.006)	0.166 (0.007)	0.108 (0.006)	0.168 (0.007)	0.146 (0.005)	0.163 (0.005)
	FSL	1.980 (0.205)	7.489 (0.287)	18.30 (0.680)	15.66 (0.448)	2.520 (0.216)	11.36 (0.269)
	$L_2(\hat{\boldsymbol{\beta}}_1)$	20.00 (4.174)	84.23 (3.303)	20.69 (2.890)	99.35 (3.508)	27.31 (0.482)	102.1 (0.861)
	$L_2(\hat{\boldsymbol{\beta}}_2)$	0.534 (0.048)	0.954 (0.054)	0.741 (0.053)	1.263 (0.061)	1.316 (0.053)	1.431 (0.062)
	$\hat{\rho}$	−0.281 (0.033)	−0.228 (0.044)	–	–	–	–
−0.85	*RMSE*	0.635 (0.029)	0.779 (0.029)	0.639 (0.019)	0.829 (0.020)	0.858 (0.021)	0.919 (0.021)
	ME	0.108 (0.007)	0.157 (0.007)	0.122 (0.006)	0.160 (0.006)	0.136 (0.005)	0.161 (0.005)
	FSL	2.280 (0.239)	7.000 (0.310)	18.70 (0.687)	17.28 (0.447)	3.380 (0.214)	10.64 (0.318)
	$L_2(\hat{\boldsymbol{\beta}}_1)$	20.61 (1.468)	65.37 (2.751)	25.34 (2.245)	83.97 (2.848)	28.64 (0.369)	81.76 (0.726)
	$L_2(\hat{\boldsymbol{\beta}}_2)$	0.549 (0.059)	0.991 (0.078)	0.795 (0.056)	1.345 (0.078)	1.269 (0.060)	1.469 (0.077)
	$\hat{\rho}$	−0.754 (0.018)	−0.765 (0.028)	–	–	–	–

- In the case of $\rho = 0$, the proposed method is comparable to SM(F) and slightly better than SM(B) in terms of *RMSE*. Regarding the loss *ME*, the proposed method shows better performance when $p = 10$ and a comparable, sometimes even worse performance when $p = 30$. The proposed model is always inferior to SM(B) with respect to *FSL*. Additionally, the proposed method performs the best under $L_2(\hat{\boldsymbol{\beta}}_2)$. However, for $L_2(\hat{\boldsymbol{\beta}}_1)$, it is worse than SM(B) when $p = 10$ and better than SM(B) in the case of $p = 30$. Overall, the proposed method performs comparably when $\rho = 0$. This is expected since there is no association between the variables Y and Z. Hence, the proposed joint model does not show its advantages.
- When $\rho = 0.3$, the proposed method remarkably outperforms the other two approaches, since SM(F) and SM(B) ignore the dependency between variables Y and Z in this case. Specifically, the proposed method gives superior performance over SM(F) regarding every criterion, especially in terms of *FSL*. Compared with SM(B), although the proposed method is comparable or even inferior under *FSL* when the model is sparse as $s = 0.2$, it is better when the true model becomes denser as $s = 0.5$. For other comparison criteria, the proposed method greatly outperforms SM(B). The results from this case demonstrate the advantages of the proposed joint model over the separate models.
- When the variables Y and $P(Z = 1)$ are negatively correlated as $\rho = -0.3$, the con-

clusions are very similar to those for $\rho = 0.3$. The proposed method consistently outperforms SM(F) and SM(B) which ignore the association between Y and Z. We also observe the same results when $\rho = \pm 0.85$ that the proposed method gives superior performance over other compared methods by taking advantage of the dependency between two responses.

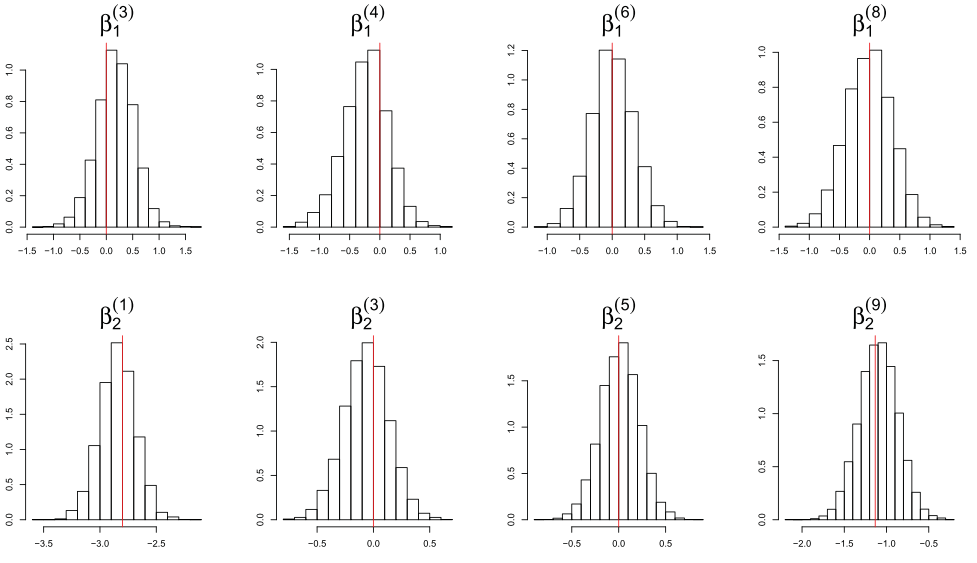

Figure 1. Histograms for the selected parameters of one replicate from $\rho = 0.3$ when $p = 10$ and $s = 0.5$.

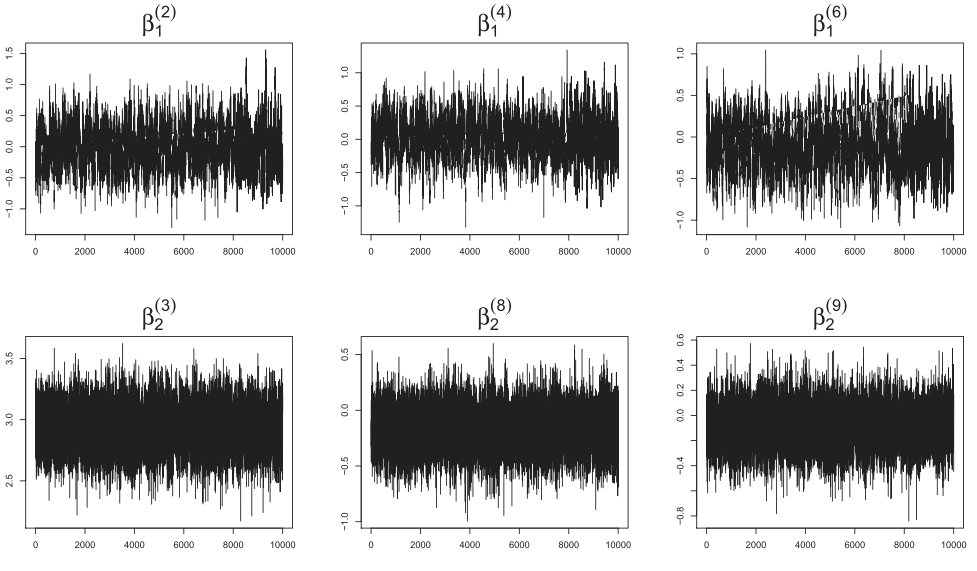

Figure 2. Trace plots for the selected parameters of one replicate from $\rho = 0.85$ when $p = 10$ and $s = 0.2$.

- The proposed method is able to provide an estimate of ρ, while the other two approaches cannot. This correlation indicates both the strength and direction of the association between Y and the probability of $Z = 1$. Hence, the estimated $\hat{\rho}$ provides us with more insight to understand data.

For illustration, based on a single simulation, Figure 1 displays the histograms for the posterior samples of some randomly selected parameters after the burn-in period with their true values indicated by the solid vertical lines. Such histogram and posterior distributions can be used for inferences. Figure 2 depicts the trace plots of posterior draws for some parameters. It is clear to see that the plots fluctuate around the mean values, indicating the MCMC chains converge. To further examine the convergence property of our interests $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and ρ, the Gelman–Rubin diagnostic is employed, which evaluates MCMC convergence by analysing the difference between multiple Markov chains. The convergence is assessed by comparing the estimated between-chains and within-chain variances. The averaged values of the potential scale reduction factor of setting $\rho = 0.85$, $p = 10$ and $s = 0.2$ over 50 replicates are 1.198, 1.002 and 1.088 for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and ρ, respectively, with their standard errors 0.095, 0.0002 and 0.032, further confirming the convergence of their MCMC chains.

6. Birth records case study

In this section, the proposed method is applied to evaluate its utility in evaluating factors associated with preterm birth and birth weight, as described in Section 1. The birth record dataset was acquired from the Virginia Department of Health via a Data Sharing Agreement and this application is approved by the Virginia Department of Health Institutional Review Board (IRB) (Protocol #40221) and Virginia Tech IRB (Protocol # 16-898). The full dataset includes over three million observations for more than two decades. Only a subset of the data was used for this study with a total of 1000 observations. In the original dataset, the binary outcome variable 'preterm birth' is extremely skewed as preterm births, in general, account for less than 10% of all live births. Hence, a random sample of $n = 1000$ is chosen such that it is more balanced with an equal number of preterm births and non-preterm births. This balancing is done for computational reasons. Further enhancements to the model to handle unbalanced data are feasible due to the Bayesian specification.

There are 9 covariates contained in this dataset, along with the two outcome variables of interest 'preterm birth' or PTB, which is dichotomous, and 'Birth Weight', which is continuous (measured in grams). The covariates include the age of the mother, day of birth, day of the week (previous research has shown seasonal as well as weekly patterns for preterm birth, e.g.[31,32], parity number (whether this is the first pregnancy carried to 24 weeks gestation or not), college education of mother (a proxy for socio-economic status of the mother), etc. A more detailed description is given in Table 3. Intuitively, the two outcome variables are negatively correlated as children who experience preterm births are also more likely to have lower birth weight.

The number of MCMC iterations is set to be 10,000 with the burn-in period of 2000. Let $(\nu, \delta^2, a, b) = (2, 2, 0.1, 0.1)$ and initial values $(\tau_{1,0}^2, \tau_{2,0}^2, r_{1,0}, r_{2,0}) = (1.5, 3, 0.3, 0.3)$ for the proposed Bayesian model. To evaluate its performance, the whole data set is randomly split into a training set with 100 observations and a testing set with 900 observations. Such partitions are repeated 50 times. For each random split, four compared methods are applied

Table 3. The variables used in the birth records case study.

Variable name	Variable description	Type of variable
z: preterm Birth	Indicator variable for whether the child was born preterm (defined as born before 36 gestational weeks)	Dichotomous dependent variable (1 = preterm, 0 = non preterm)
y: Birth Weight	Weight of the infant at birth in grams	Quantitative dependent variable
x_1: Day of birth	Day of the year (1–366) the infant was born	Quantitative independent variable
x_2: Day of week	Whether the infant was born on a weekend or a weekday	Dichotomous independent variable (1 = weekend, 0 = weekday)
x_3: Age of mother	Age of the mother in years	Quantitative independent variable
x_4: Race	Race reported on the birth record collapsed to whether the infant is identified as African-American or not	Dichotomous independent variable (1 = African-American, 0 = Not African-American)
x_5: Ethnicity	Whether the infant is identified as Hispanic or not	Dichotomous independent variable (1 = Hispanic, 0 = Not Hispanic)
x_6: Mother's Education	Whether the mother completed at least high school or not	Dichotomous independent variable (1 = More than High School, 0 = High School or less)
x_7: Marriage status	Whether the mother was married at the time of birth or not	Dichotomous independent variable (1 = Married, 0 = Not Married)
x_8: Sex of child	The sex of the infant	Dichotomous independent variable (1 = Male infant, 0 = Female infant)
x_9: Parity	Number of pregnancies carried to 24 weeks gestation collapsed to whether this is the first such pregnancy or not	Dichotomous independent variable (1 = First pregnancy, 0 = Not first pregnancy)

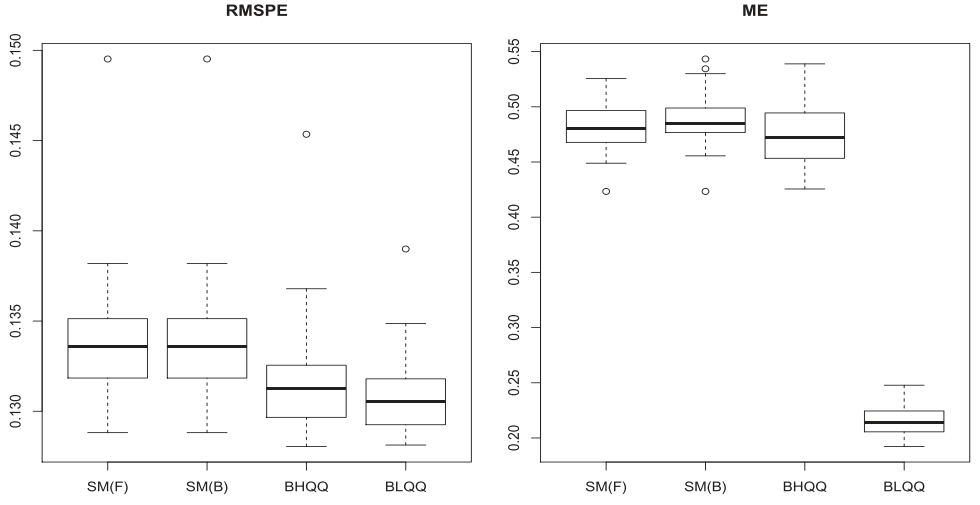

Figure 3. Boxplots of RMSPE and mis-classification error for preterm birth data for each approach.

to fit the training data, including SM(F), SM(B), the Bayesian Hierarchical QQ Model by Kang et al. [8] (BHQQ for short) and the proposed Bayesian Latent QQ model (BLQQ). Then the predictions of two responses are made on the testing data.

Figure 3 shows the root-mean-square prediction error (RMSPE) and misclassification error (ME) for each method. The separate models, SM(F) and SM(B), perform similarly to each other, while the proposed method shows better performance than both of them because of the dependency of two outcome variables. The proposed model gives a significantly lower ME, indicating that it can distinguish preterm births from non-preterm births much more accurately. The proposed model is also better in predicting the birth weight as

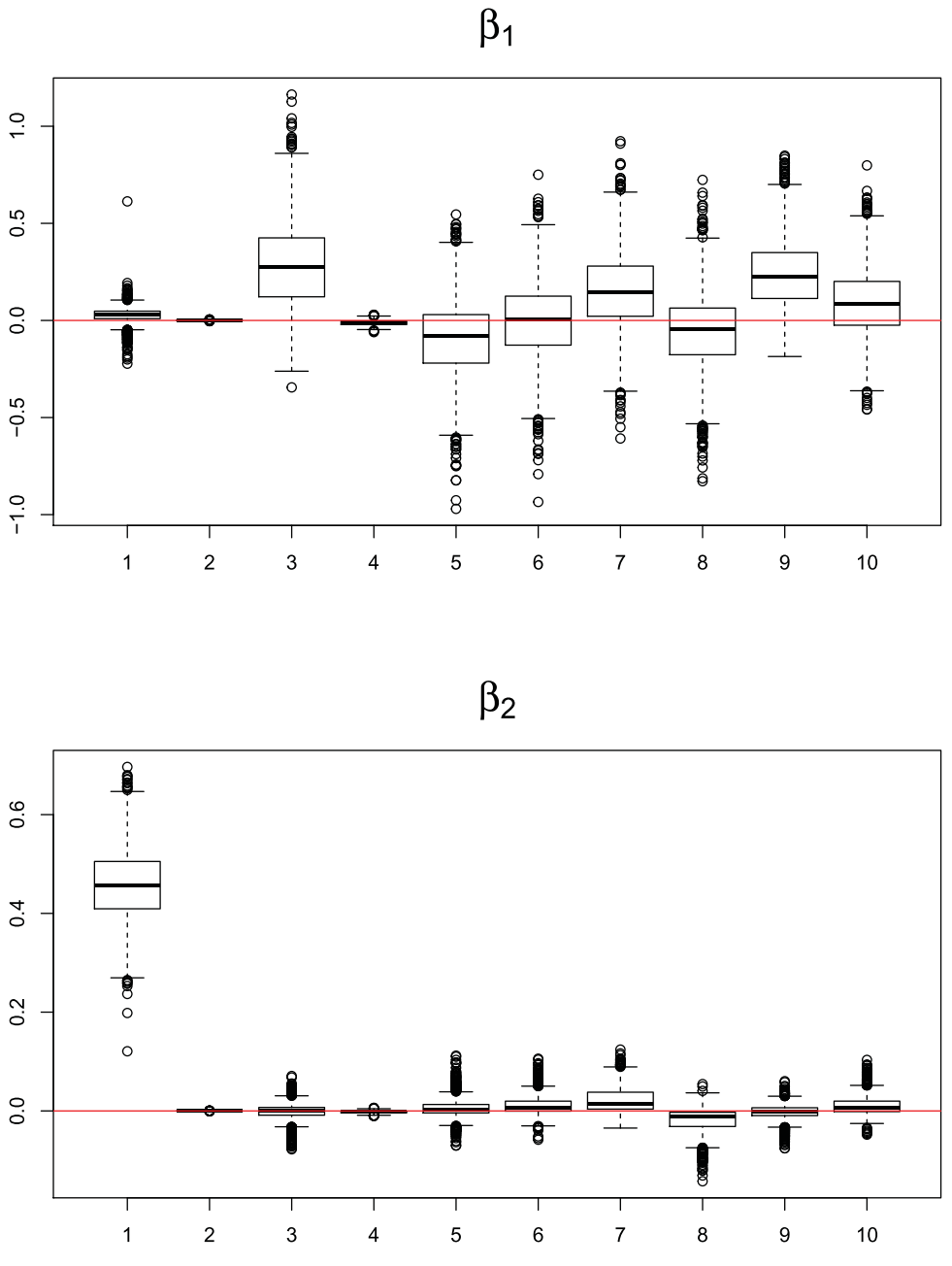

Figure 4. Regression coefficient distributions for the explanatory variables (1 indicates the regression constant) for the quantitative and qualitative responses across 50 replications.

shown in the boxplot of RMSPE. Besides, the proposed method can account for the correlation between birth weight and the probability of PTB. The average of the estimated correlation over 50 splits is -0.772 with a standard error of 0.063. We also note that for each split of the data set, the estimated correlation is negative. It means the smaller the value of the birth weight variable, the more likely the corresponding birth is preterm. Note

that the latest method BHQQ is comparable with the proposed method in terms of prediction accuracy for the continuous outcome, but is much worse regarding ME. This is expected as it has been explained in Section 1. BHQQ uses the marginal logistic regression model for the binary outcome and thus cannot improve the prediction accuracy for the binary outcome.

Next, we investigate the analysis results based on one random split of the training and testing data sets. There are 500 observations with PTB $=$ 1 and 500 observations with PTB $=$ 0 in the testing set. The estimate of the correlation is -0.85. The trace plots indicate that Gibbs sampling iterations converge and the ACF plots show that the autocorrelation dies off. These plots are omitted in the paper. Figure 4 depicts the boxplots of the regression coefficients across 50 replications (the x-axis is numbered from 1 to 10 to indicate regression constant and the slopes corresponding to the 9 explanatory variables). The first subplot corresponds to the regression coefficients for the qualitative response (preterm birth) and the second subplot corresponds to the regression coefficients for the quantitative response (birth weight). Given the complex biological and physiological causes of preterm births and birth weights of children, it is not surprising that the regression coefficients are not statistically significant at the default 0.05 level.

7. Discussion

In this article, we propose a Bayesian latent variable model to jointly fit data with qualitative and quantitative (QQ) outcomes. The work is motivated by a birth records study involving two responses: birth weight (quantitative variable) and preterm birth (qualitative variable). The proposed model uses a latent variable to link the quantitative and qualitative responses, improving the prediction accuracy for both variables, while some existing works without using a latent variable fit one response conditional on the other response, hence improving the prediction accuracy for only one response. Moreover, the proposed model can capture the correlation between the quantitative response and the latent variable, which is an indicator of the dependency strength for the quantitative and qualitative responses. Besides, the proposed Bayesian framework is more convenient to provide statistical inference for the parameters than the frequentist analysis based on the asymptotic distribution of the estimator, which is complicated and difficult to derive. The merits of the proposed Bayesian latent variable model is demonstrated by the numerical study and a birth records data set.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Xiaoning Kang's work was supported by the Ministry of Education of China [20YJC910007]. Lulu Kang's work was supported by the National Science Foundation (Division of Mathematical Sciences) [DMS-1916467]. Xinwei Deng's work was supported by the National Science Foundation CISE Expedition (Division of Computing and Communication Foundations) [CCF-1918770]. The work of Shyam Ranganathan and Julia Gohlke was partially supported by the National Institute of Environmental Health Sciences [R21ES028396].

ORCID

Shyam Ranganathan http://orcid.org/0000-0002-1337-5173
Lulu Kang http://orcid.org/0000-0002-6000-3436
Julia Gohlke http://orcid.org/0000-0002-6984-2893
Xinwei Deng http://orcid.org/0000-0002-1560-2405

References

[1] Horbar JD, Badger GJ, Carpenter JH, et al. Trends in mortality and morbidity for very low birth weight infants, 1991–1999. Pediatrics. 2002;110(1):143–151.
[2] Hack M, Klein NK, Taylor HG. Long-term developmental outcomes of low birth weight infants. Future Child. 1995;5(1):176–196.
[3] Shah PS, Balkhair T, Ohlsson A, et al. Intention to become pregnant and low birth weight and preterm birth: a systematic review. Matern Child Health J. 2011;15(2):205–216.
[4] Goldenberg RL, Culhane JF, Iams JD, et al. Epidemiology and causes of preterm birth. Lancet. 2008;371(9606):75–84.
[5] Saigal S, Doyle LW. An overview of mortality and sequelae of preterm birth from infancy to adulthood. Lancet. 2008;371(9608):261–269.
[6] Russell RB, Green NS, Steiner CA, et al. Cost of hospitalization for preterm and low birth weight infants in the united states. Pediatrics. 2007;120(1):e1–e9.
[7] Deng X, Jin R. Qq models: joint modeling for quantitative and qualitative quality responses in manufacturing systems. Technometrics. 2015;57(3):320–331.
[8] Kang L, Kang X, Deng X, et al. A Bayesian hierarchical model for quantitative and qualitative responses. J Qual Technol. 2018;50(3):290–308.
[9] Moustaki I, Knott M. Generalized latent trait models. Psychometrika. 2000;65(3):391–411.
[10] Catalano PJ, Ryan LM. Bivariate latent variable models for clustered discrete and continuous outcomes. J Am Statist Assoc. 1992;87(419):651–658.
[11] Dunson DB. Bayesian latent variable models for clustered mixed outcomes. J R Statist Soc Ser B (Statist Methodol). 2000;62(2):355–366.
[12] Wang K, Tsung F. Run-to-run process adjustment using categorical observations. J Qual Technol. 2007;39(4):312–325.
[13] Liu K, Huang S. Integration of data fusion methodology and degradation modeling process to improve prognostics. IEEE Trans Automat Sci Eng. 2014;13(1):344–354.
[14] Cheng C, Sa-Ngasoongsong A, Beyca O, et al. Time series forecasting for nonlinear and non-stationary processes: a review and comparative study. IIE Trans. 2015;47(10):1053–1071.
[15] Zhou Y, Whitehead J, Bonvini E, et al. Bayesian decision procedures for binary and continuous bivariate dose-escalation studies. Pharm Statist J Appl Statist Pharm Ind. 2006;5(2):125–133.
[16] Shi J. Stream of variation modeling and analysis for multistage manufacturing processes. Boca Raton: CRC Press; 2006.
[17] Cox DR, Wermuth N. Response models for mixed binary and quantitative variables. Biometrika. 1992;79(3):441–461.
[18] Dunson DB. Dynamic latent trait models for multidimensional longitudinal data. J Am Statist Assoc. 2003;98(463):555–563.
[19] Fitzmaurice GM, Laird NM. Regression models for a bivariate discrete and continuous outcome with clustering. J Am Statist Assoc. 1995;90(431):845–852.
[20] Gueorguieva RV, Agresti A. A correlated probit model for joint modeling of clustered binary and continuous responses. J Am Statist Assoc. 2001;96(455):1102–1112.
[21] Olkin I, Tate RF. Multivariate correlation models with mixed discrete and continuous variables. Ann Math Statist. 1961;32(2):448–465.
[22] McCulloch C. Joint modelling of mixed outcome types using latent variables. Statist Methods Med Res. 2008;17(1):53–73.
[23] Hwang BS, Pennell ML. Semiparametric Bayesian joint modeling of a binary and continuous outcome with applications in toxicological risk assessment. Stat Med. 2014;33(7):1162–1175.

[24] Yeung WY, Whitehead J, Reigner B, et al. Bayesian adaptive dose-escalation procedures for binary and continuous responses utilizing a gain function. Pharm Stat. 2015;14(6):479–487.
[25] Sun H, Rao PK, Kong ZJ, et al. Functional quantitative and qualitative models for quality modeling in a fused deposition modeling process. IEEE Trans Automat Sci Eng. 2017;15(1):393–403.
[26] Holmes CC, Held L. Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Anal. 2006;1(1):145–168.
[27] Wu CJ, Hamada MS. Experiments: planning, analysis, and optimization. Vol. 552. Hoboken, NJ: John Wiley & Sons; 2011.
[28] Joseph VR. A Bayesian approach to the design and analysis of fractionated experiments. Technometrics. 2006;48(2):219–229.
[29] Ai M, Kang L, Joseph VR. Bayesian optimal blocking of factorial designs. J Statist Plan Inference. 2009;139(9):3319–3328.
[30] Kang L, Joseph VR. Bayesian optimal single arrays for robust parameter design. Technometrics. 2009;51(3):250–261.
[31] Darrow LA, Strickland MJ, Klein M, et al. Seasonality of birth and implications for temporal studies of preterm birth. Epidemiol. (Cambridge, MA.) 2009;20(5):699.
[32] Palmer WL, Bottle A, Aylin P. Association between day of delivery and obstetric outcomes: observational study. BMJ. 2015;351:h5774.

Appendices

Appendix 1

The leave-one-out predictive distribution for $u_i|\boldsymbol{u}_{-i}, \boldsymbol{y}, \boldsymbol{z}, \sigma^2, \rho$ can be obtained through

$$p(u_i|\boldsymbol{u}_{-i}, \boldsymbol{y}, \boldsymbol{z}, \sigma^2, \rho) = \int p(u_i|y_i, z_i, \boldsymbol{\beta}, \sigma^2, \rho) p(\boldsymbol{\beta}|\boldsymbol{u}_{-i}, \boldsymbol{y}, \boldsymbol{z}, \sigma^2, \rho)\, \mathrm{d}\boldsymbol{\beta},$$

where $p(\boldsymbol{\beta}|\boldsymbol{u}_{-i}, \boldsymbol{y}, \boldsymbol{z}, \sigma^2, \rho)$ can be derived in the same way as we did for (4). The sampling distribution of $(\boldsymbol{u}_{-i}, \boldsymbol{y})$ is directly obtain as

$$\left[\begin{array}{c} \boldsymbol{u}_{-i} \\ \boldsymbol{y} \end{array}\right] \Bigg| \boldsymbol{\theta} \sim N\left(\mathbb{X}_{-i}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\epsilon,-i}\right),$$

where $\mathbb{X}_{-i}$ is the matrix $\mathbb{X}$ with its ith row removed, i.e.

$$\mathbb{X}_{-i} = \left[\begin{array}{cc} \boldsymbol{X}_{-i}, & \boldsymbol{0}_{(n-1)\times p} \\ \boldsymbol{0}_{n\times p}, & \boldsymbol{X} \end{array}\right].$$

Here $\boldsymbol{X}_{-i}$ is $\boldsymbol{X}$ without its ith row. The covariance matrix $\boldsymbol{\Sigma}_{\epsilon,-i}$ is $\boldsymbol{\Sigma}_{\epsilon}$ with the ith row and ith column removed. For convenience, permute the rows and columns of $\boldsymbol{\Sigma}_{\epsilon}$ so that the ith row and column are the last,

$$\boldsymbol{\Sigma}_{\epsilon} = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{\epsilon,-i}, & \boldsymbol{l} \\ \boldsymbol{l}', & 1 \end{array}\right],$$

where $\boldsymbol{l} = [\boldsymbol{0}_{1\times(n-1)}, 0, \ldots, 0, \rho\sigma, 0, \ldots, 0]$. So all the elements of $\boldsymbol{l}$ are zeroes except the $(n-1+i)$th element is $\rho\sigma$. Since the prior of $\boldsymbol{\beta}$ is $N(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$, the full-conditional distribution of $\boldsymbol{\beta}$ conditioned on $(\boldsymbol{u}_{-i}, \boldsymbol{y})$ is

$$\boldsymbol{\beta}|\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta},-i}, \boldsymbol{\Sigma}_{\boldsymbol{\beta},-i}).$$

Through direct calculation,

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta},-i} = \left(\boldsymbol{\Sigma}_0^{-1} + \mathbb{X}'_{-i}(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}\mathbb{X}_{-i}\right)^{-1},$$

$$\boldsymbol{\mu}_{\boldsymbol{\beta},-i} = \boldsymbol{\Sigma}_{\boldsymbol{\beta},-i}\mathbb{X}'_{-i}(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}\left[\begin{array}{c} \boldsymbol{u}_{-i} \\ \boldsymbol{y} \end{array}\right].$$

Previously, it has been shown that

$$u_i|y_i, z_i, \boldsymbol{\theta} \sim \begin{cases} N\left(\boldsymbol{x}_i'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_2), (1-\rho^2)\right) \frac{I(u_i \geq 0)}{\Phi(s(y_i)|\boldsymbol{\theta})}, & \text{if } z_i = 1, \\ N\left(\boldsymbol{x}_i'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_2), (1-\rho^2)\right) \frac{I(u_i < 0)}{1 - \Phi(s(y_i)|\boldsymbol{\theta})}, & \text{if } z_i = 0, \end{cases}$$

and

$$u_i|y_i, \boldsymbol{\theta} \sim N\left(\boldsymbol{x}_i'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_2), (1-\rho^2)\right).$$

Hence, the distribution for $u_i|\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho$ should also be a normal distribution. Its mean and variance are

$$\begin{aligned} m_i &= E(u_i|\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho) \\ &= E_{\boldsymbol{\beta}}\left(E_{u_i}\left(u_i|y_i, \boldsymbol{\beta}, \sigma^2, \rho\right)|\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho\right) \\ &= E_{\boldsymbol{\beta}}\left(\boldsymbol{x}_i'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_2)|\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho\right) \\ &= \frac{\rho}{\sigma}y_i + \left[\boldsymbol{x}_i', -\frac{\rho}{\sigma}\boldsymbol{x}_i'\right]\boldsymbol{\mu}_{\boldsymbol{\beta},-i} \end{aligned}$$

and

$$\begin{aligned} v_i &= \text{var}\left(u_i|\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho\right) \\ &= \text{var}_{\boldsymbol{\beta}}\left(E_{u_i}\left(u_i|y_i, \boldsymbol{\beta}, \sigma^2, \rho\right)|\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho\right) + E_{\boldsymbol{\beta}}\left(\text{var}\left(u_i|y_i, \boldsymbol{\beta}, \sigma^2, \rho\right)|\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho\right) \\ &= \text{var}_{\boldsymbol{\beta}}\left(\boldsymbol{x}_i'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_2)|\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho\right) + E_{\boldsymbol{\beta}}\left((1-\rho^2)1|\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho\right) \\ &= \left[\boldsymbol{x}_i', -\frac{\rho}{\sigma}\boldsymbol{x}_i'\right]\boldsymbol{\Sigma}_{\boldsymbol{\beta},-i}\begin{bmatrix} \boldsymbol{x}_i, \\ -\frac{\rho}{\sigma}\boldsymbol{x}_i \end{bmatrix} + (1-\rho^2). \end{aligned}$$

Therefore, the leave-one-out distribution for $u_i|\boldsymbol{u}_{-i}, \boldsymbol{y}, \sigma^2, \rho$ is $N(m_i, v_i)$. Adding $\boldsymbol{z}$, one can obtain

$$p(u_i|\boldsymbol{u}_{-i}, \boldsymbol{y}, z_i, \sigma^2, \rho) \propto \begin{cases} N(u_i|m_i, v_i)I(u_i \geq 0), \text{if } z_i = 1, \\ N(u_i|m_i, v_i)I(u_i < 0), \text{if } z_i = 0. \end{cases}$$

Appendix 2

Since the values of $\boldsymbol{\Sigma}_{\epsilon,-i}$ and $\boldsymbol{\mu}_{\boldsymbol{\beta},-i}$ have to be computed for every u_i in each sampling of $\boldsymbol{u}$, it is thus necessary to find a quick way to compute both. Suppose $(\boldsymbol{\Sigma}_\epsilon)^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ have already been computed. It can be shown that

$$\boldsymbol{\Sigma}_\epsilon^{-1} = \begin{bmatrix} (\boldsymbol{\Sigma}_{\epsilon,-i})^{-1} + c\left((\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}\boldsymbol{l}\boldsymbol{l}'(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}\right), & -c(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}\boldsymbol{l} \\ -c\boldsymbol{l}'(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}, & c \end{bmatrix},$$

where c is the diagonal entry of $\boldsymbol{\Sigma}_\epsilon^{-1}$ and $c = (\sigma_1^2 - \boldsymbol{l}'\boldsymbol{\Sigma}_{\epsilon,-i}^{-1}\boldsymbol{l})^{-1}$. As a result, it is easy to obtain

$$(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1} = \left(\boldsymbol{\Sigma}_\epsilon^{-1}\right)_{-i,-i} - c\left((\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}\boldsymbol{l}\boldsymbol{l}'(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}\right) = \left(\boldsymbol{\Sigma}_\epsilon^{-1}\right)_{-i,-i} - c^{-1}\left(\boldsymbol{\Sigma}_\epsilon^{-1}\right)_{-i,i}\left(\boldsymbol{\Sigma}_\epsilon^{-1}\right)_{i,-i}.$$

Here $(\boldsymbol{\Sigma}_\epsilon^{-1})_{-i,i}$ is the ith column of matrix $\boldsymbol{\Sigma}_\epsilon^{-1}$ without the ith diagonal entry $(\boldsymbol{\Sigma}_\epsilon^{-1})_{ii}$, the notation $(\boldsymbol{\Sigma}_\epsilon^{-1})_{i,-i}$ is the ith row of $\boldsymbol{\Sigma}_\epsilon^{-1}$ without the ith diagonal entry, and $(\boldsymbol{\Sigma}_\epsilon)_{-i,-i}^{-1}$ is the matrix $\boldsymbol{\Sigma}_\epsilon^{-1}$ with the ith row and ith column removed. Define

$$\boldsymbol{b} = \mathbb{X}_i - \mathbb{X}_{-i}'(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}\boldsymbol{l} = \mathbb{X}_i + c^{-1}\mathbb{X}_{-i}'\left(\boldsymbol{\Sigma}_\epsilon^{-1}\right)_{-i,i}.$$

The column vector $\mathbb{X}_i$ is the transpose of the ith row of $\mathbb{X}$. In addition, we have

$$\begin{aligned}\mathbb{X}'\boldsymbol{\Sigma}_\epsilon^{-1}\mathbb{X} &= \mathbb{X}'_{-i}(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}\mathbb{X}_{-i} + c\boldsymbol{b}\boldsymbol{b}',\\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= \left(\boldsymbol{\Sigma}_0^{-1} + \mathbb{X}'\boldsymbol{\Sigma}_\epsilon^{-1}\mathbb{X}\right)^{-1}\\ &= \left(\boldsymbol{\Sigma}_0^{-1} + \mathbb{X}'_{-i}(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}\mathbb{X}_{-i} + c\boldsymbol{b}\boldsymbol{b}'\right)^{-1}\\ &= \left((\boldsymbol{\Sigma}_{\boldsymbol{\beta},-i})^{-1} + c\boldsymbol{b}\boldsymbol{b}'\right)^{-1}.\end{aligned}$$

Thus,

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\beta},-i} &= \left((\boldsymbol{\Sigma}_{\boldsymbol{\beta}})^{-1} - c\boldsymbol{b}\boldsymbol{b}'\right)^{-1}\\ &= \boldsymbol{\Sigma}_{\boldsymbol{\beta}} + \frac{c}{1 - c\boldsymbol{b}'\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{b}}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{\Sigma}_{\boldsymbol{\beta}}.\end{aligned}$$

The vector $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{b}$ can be obtained from an intermediate calculation of $\boldsymbol{\mu}_{\boldsymbol{\beta}}$.

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{b} = c^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbb{X}'\boldsymbol{\Sigma}_\epsilon^{-1}\right)_{.,i}.$$

Here $(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbb{X}'\boldsymbol{\Sigma}_\epsilon^{-1})_{.,i}$ is the ith column of matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbb{X}'\boldsymbol{\Sigma}_\epsilon^{-1}$ of size $2p \times 2n$. The mean $\boldsymbol{\mu}_{\boldsymbol{\beta},-i}$ is

$$\boldsymbol{\mu}_{\boldsymbol{\beta},-i} = \boldsymbol{\Sigma}_{\boldsymbol{\beta},-i}\mathbb{X}'_{-i}(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1}\begin{bmatrix}\boldsymbol{u}_{-i}\\ \boldsymbol{y}\end{bmatrix},$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta},-i}$ can be obtained as above, and $(\boldsymbol{\Sigma}_{\epsilon,-i})^{-1} = (\boldsymbol{\Sigma}_\epsilon^{-1})_{-i,-i} - c^{-1}(\boldsymbol{\Sigma}_\epsilon^{-1})_{-i,i}(\boldsymbol{\Sigma}_\epsilon^{-1})_{i,-i}$.