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

Robust Model Fitting for Two-Part Model Within Bayesian Semiparametric Framework: Variational Approach

Authors: Jinye Chen
– College of Economics and Management, Nanjing Forestry University, Nanjing, China
chenjy@njfu.edu.cn

Qi Zhang
– College of Science, Nanjing Forestry University, Nanjing, China
zhang_qi1998@163.com

Chenhui Yang
– College of Science, Nanjing Forestry University, Nanjing, China
chyang1215@163.com

Yemao Xia ✉
– College of Science, Nanjing Forestry University, Nanjing, China
ym_xia71@163.com

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

Abstract:

- A semiparametric Bayesian fitting is developed for two-part model (TPM) to guard against the distribution deviations. Parametric assumption on the continuous process is replaced by the Dirichlet process mixture (DPM) model. By taking advantage of the sparseness of DP, the semiparametric fitting automatically adjust random weights to accommodate heterogeneity such as skewness, multi-modality and/or extreme observations. We develop a variational Bayesian inference procedure. A mean-field based variational density is constructed to approximate posterior. Our empirical results show that the proposed method is more robust against the distributional deviations and/or outliers, and can efficiently improve model fitting.

Keywords:

- *Two-part model; semi-parametric Bayesian fitting; variational Bayes; mean-field family; household finance survey.*

✉ Corresponding author

AMS Subject Classification:

- 62C10, 62F15.

1. INTRODUCTION

In social surveys, researchers often encounter the semi-continuous variable. A typical feature of semi-continuous variable is that its realizations are nonnegative but with a large proportion of zeros. The dataset illustrates strong heterogeneity. In understanding such data structure, two-part model (TPM, [7, 8];[31]) is an appreciated statistical method. TPM assumes that the overall model is consisted of two submodels: a binary value model (Part one) and a continuous value model (Part two). The binary model is usually formulated via logistic or probit regression model to indicate the effects of explanatory factors on the proportion of zeros in data, and the continuous model is generally specified within the log-normal or log-skew-elliptical linear regression model to describe the effects of covariates on the means of responses. By integrating binary model and continuous process into one, TPM provides a unified and flexible way to describe various relevances for semi-continuous data. Now, TPM has been widely used in the fields of household finance, public health, psychological education, finance insurance or other fields, see for example, [8], [20], [27], [31], [33], [34]; [37], [39], [54] and references therein. The reader can also refer to, for example, [11], [19], [26], [53] and [55] for the extensions of TPM in the latent variable analysis context. However, most developments mentioned above are confined to the parametric fitting, that is, particular parametric assumptions are specified to the positive-valued part to facilitate statistical computation and/or theoretical results. As is the well-known, parametric assumption is more sensitive to the distributional deviations or outliers. It readily results in biased inferences or misleading conclusions when the posited model is not consistent with true population.

In recent years, much effort has been devoted to robustifying two-part model fitting. For example, Chai and Bailey [6] suggested using a skewed normal distribution to fit the log-transformed positive values to accommodate skewness arising from the log-transformation. Manning *et al.* [32] proposed a flexible one-part generalized gamma distribution that included log-normal, Weibull and exponential distributions as special cases to adjust the skewness of semi-continuous data. In the analysis of alcohol consumption data, Liu *et al.* [28] extended traditional TPM to the situation where the mixed-effects were incorporated into the analysis of TPM and argued that a generalized gamma distribution in Part two could provide the best fit in the comparisons with models with log-skew-normal distribution and Box-Cox transformation; for the spatial semi-continuous data, Neelon, Zhu and Neelon [37] developed a broad class of Bayesian two-part spatial models to address the heterogeneity underlying Part two. The skewed-normal and skewed t-distributions were fitted to the positive values on log-scale to accommodate skewness or heavier-than-normal tails. Moreover, a mixture model was applied to the overall model to explore the heterogeneity of population. Xing *et al.* [54] claimed that fitting a skewed -normal distribution to the logarithmic positive values may overcorrect the skewness of data. They suggested modeling a skewed normal or skewed t-distribution on the continuous positive values

directly. Though more appealing, these developments are still confined to the parametric fitting. It heavily depends on the form of parametric distribution, and hence is rather limited in dealing with the distributional deviation or model mis-specification.

In this paper, we proposed a Bayesian semiparametric fitting for the continuous part of semi-continuous data. Our approach is more along with lines of the mixture method in [37] but we pursue a Bayesian fitting for the log-transformed positive-valued part, not exploring the heterogeneity of entire population. We resort to the Dirichlet process mixture model(DPM, [2], [9], [10], [14], [21], [22], [29], [30], [35], [41] and among others). The normal assumption on the errors of Part two is replaced by the normal mixture model mixed with Dirichlet process(DP). By taking advantage of the sparsity of DP, the posited model can be automatically selected from the space of all possible distribution functions. Such a relaxation gives relatively weak assumptions on the sampling model and hence could provide a sufficiently flexible modeling for data.

Within the Bayesian paradigm, the analysis of TPM usually resorts to the approximation methods. Markov chain Monte Carlo sampling method (MCMC,[15]; [17]) is undoubtedly a powerful tool in dealing with the complex models and hierarchically structured data, see [11]; [53] and [52] in the context of MCMC sampling for TPM within the framework of latent variables. However, MCMC often suffers some limitations, e.g., slow convergence, time-consuming cycles and non-ignorable sampling errors. In this paper, we pursue a variational Bayesian (VB) inference procedure ([4, 5], [24], [42], [50], [51]). An appeal underlying VB is its elegant computation efficiency and deterministic solutions. It is also more convenient for dealing with big data. Up to now, variational Bayesian inference has been widely used in computer vision and robotics, and others, see [5] for a review of recent applications of VB. But to the best of our knowledge, no developments have been made on the analysis of semi-continuous data, especially for the analysis of TPM under the Bayesian semiparametric fitting.

The rest of the paper is organized as follows. Section 2 introduces the proposed model for semi-continuous data. Section 3 gives variational inference procedure under the Bayesian framework. Parameter estimation, variable selection and model assessment are also presented in this section. A simulation study to assess the performance of the proposed model is given in Section 4. In Section 5, we apply the proposed method to the China household finance data. Section 6 concludes the paper with a discussion. Some technical details are given in the supplementary material.

2. MODEL DESCRIPTION

2.1. Two-part semi-parametric model

Suppose that for $i = 1, \cdots, n$, y_i is a semi-continuous variable which takes value in $[0, \infty)$; $\mathbf{x}_i$ is a generic vector of fixed covariates, consisted of q categorical or continuous variables, used to identify the variability of y_i. To deal with the excess zeros in semi-continuous data, in the literature, y_i is usually identified with an indicator variable u_i and an intensity variable z_i as that $u_i = I\{y_i > 0\}$ and $z_i = \log(y_i|y_i > 0)$, where $I\{A\}$ is the indicator function of set A. In this case, the population of y_i is totally determined by the joint distribution of u_i and z_i.

In exploring the effects of $\mathbf{x}_i$ on y_i, two-part model(TPM,[8],[31]) specifies two submodels, say Part one and Part two, for u_i and z_i respectively. Part one assumes that conditional on $\mathbf{x}_i$, u_i satisfies the following generalized linear model [38]

$$h(P(u_i = 1|\mathbf{x}_i)) = \alpha + \mathbf{x}_i^T\boldsymbol{\beta}, \tag{2.1}$$

where $h(\cdot)$ is any link function used to link the predictor to the mean of u_i, α is the intercept and $\boldsymbol{\beta}$ is the vector of regression coefficients. Various constructs can be chosen for h (see for example, [46]). We here prefer using the sigmoid or logistic function for the inverse of h, i.e., $h^{-1}(x) = \sigma(x) = 1/(1 + e^{-x})$. Such a choice is also popular in the pattern recognition and machine learning ([3]).

Under the parametric modeling framework, Part two usually formulates z_i via linear normal regression model as follows

$$z_i|u_i = 1 \sim N(\gamma + \mathbf{x}_i^T\boldsymbol{\psi}, \sigma^2) \tag{2.2}$$

where $\boldsymbol{\psi}$ is the vector of regression coefficients, and γ and σ^2 are the scalars of intercepter and scale respectively. However, in many circumstances, single normal assumption on z_i may be inappropriate. It is especially true when dataset takes on the high skewness, heavy tails and/or multi-modalities. Although least square estimates (LSE) of regression coefficients are more robust against the distributional deviations, their standard deviations will be seriously distorted when the posited model is misspecified. To robustify model fitting, in this paper, we extend the parametric model (2.2) to the semi-parametric setting in which the normal distribution is replaced by the normal mixture model mixed with the Dirichlet process (DP, [13]). To this end, let $\epsilon_i = z_i - \mathbf{x}_i^T\boldsymbol{\psi}$ be the idiosyncratic part in z_i. Unlike that in (2.2), we first assume that conditional on $u_i = 1$, ϵ_i has the normal distribution with inhomogeneous mean and variance as follows:

$$\epsilon_i|u_i = 1 \overset{ind.}{\sim} N(\gamma_i, \sigma_i^2), \tag{2.3}$$

where γ_i and σ_i^2 are the subject-specific latent variables taking value in $\mathbb{R}$ and $\mathbb{R}^+$ respectively. Further, let $\boldsymbol{\theta}_i = (\gamma_i, \sigma_i^2)^T$. We assume that $\boldsymbol{\theta}_i$s are iid. with

the common distribution P, while P is treated to be random distributed with the DP prior $D(cF_0)$, where $c > 0$ is the concentration parameter controlling the variability of P around F_0, and F_0 is the baseline distribution representing the center of P. As pointed out by Lo [29], the normal mixture of DP is more flexible in the semiparametric Bayesian fitting and it can provide enough functions to be chosen from the space of all probability distributions on $\mathbb{R}$.

It follows from the Sethuraman's construction [47] that P can be stochastically expressed as the mixture model in the form $P(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}(\cdot)$, where $\boldsymbol{\theta}_k^* = (\gamma_k^*, \sigma_k^{*2})^T$s are the iid. atoms with the common distribution F_0, and π_k are the random weights constructed via stick-breaking procedure given by

$$\pi_1 = V_1, \cdots, \pi_k = V_k \prod_{\ell=1}^{k-1} (1 - V_\ell)(k = 2, 3, \cdots), \tag{2.4}$$

where V_k are the iid. $Beta(1, c)$ random variables. The discreteness of P induces an infinite normal mixture model $\sum_{k=1}^{\infty} N(\gamma_k^*, \sigma_k^{*2})$ for ϵ_i. In real applications, the infinite sum is usually truncated at the finite level, say K, which results in the following truncated DP (TDP, [22])

$$P \overset{D}{=} \sum_{k=1}^{K} \pi_k \delta_{\theta_k^*}(\cdot), \tag{2.5}$$

in which V_K in π_K is set at 1 to ensure the sum of π_k to be unity, and K is a positive integer representing the level of truncation in approximating DP. Essentially, TDP is a finite dimensional distribution prior. It puts the space of P on the union of the spaces of the finite mixtures with at most K atoms. In the context of linear models, Ishwaran and James [23] exploited some theoretical properties of TDP. They provided a guideline to assess the sufficiency of approximation. According to their developments, the setting $n = 400$, $K = 100$ and $c = 2$ yields an L_1 bound 1.272×10^{-19} for the distance between the marginal likelihoods under DP and TDP. Therefore, even for huge sample sizes, a mere truncation leads to an approximating hierarchical model that is virtually indistinguishable from one based on the DP prior (see [23] for more discussions). More importantly, they showed that working with TDP can lead to effective Gibbs updates in MCMC sampling. Note that TDP can be re-expressed by introducing the configuration variables $s_i \in \{1, 2, \cdots, K\}$ such that $\boldsymbol{\theta}_i = \boldsymbol{\theta}_k^*$ if $s_i = k$. In this case, $s_i|\boldsymbol{\pi} \overset{iid.}{\sim} \sum_{k=1}^{K} \pi_k \delta_k(s_i)$, and $\boldsymbol{\theta}_i$ is totally determined by $\boldsymbol{\theta}_k^*$ and s_i.

2.2. Pólya-gamma stochastic repression

The main challenge in the current analysis comes from the logistic function, a nonlinear model which forbids the explicit form of the posterior of regression parameters. To facilitate the posterior analysis, we follow the routine in [43] and

rewrite equation (2.1) as follows:

$$\frac{\exp(u_i\eta_i)}{1+\exp(\eta_i)} = 2^{-1}\exp\{\kappa_i\eta_i\}\int_0^\infty \exp\left\{-\frac{u_i^*}{2}\eta_i^2\right\} p_{\mathrm{PG}}(u_i^*)du_i^*, \tag{2.6}$$

where $\eta_i = \alpha + \mathbf{x}_i^T\boldsymbol{\beta}$ and $\kappa_i = u_i - 1/2$; $p_{\mathrm{PG}}(u_i^*)$ is the standard Pólya-Gamma PG$(1,0)$ probability density function [43]. An advantage of working with equation (2.6) is that if we introduce auxiliary variables u_i^* and augment them with u_i, then equation (2.1) can be considered as the marginal density of the joint distribution

$$p(u_i, u_i^*|\mathbf{x}_i, \alpha, \boldsymbol{\beta}) = 2^{-1}\exp\left\{\kappa_i\eta_i - \frac{u_i^*}{2}\eta_i^2\right\} p_{\mathrm{PG}}(u_i^*). \tag{2.7}$$

Note that the exponential part in the brackets is the kernel of normal density function with respect to η_i. Hence, it admits conjugate full-conditional distributions for all model parameters, leading to a straightforward Bayesian computation, though at the expense of sampling from $p_{\mathrm{PG}}(u_i^*)$.

Let $\mathbf{U} = \{u_i\}_{i=1}^n$ and $\mathbf{Z} = \{z_i\}_{i\in\mathcal{I}}$ be the sets of the observed responses, where $\mathcal{I}$ is the set of indices such that $u_i = 1$; we refer to $\mathbf{U}^* = \{u_i^*\}_{i=1}^n$ and $\mathbf{S} = \{s_i : i \in \mathcal{I}\}$ as the sets of local latent responses, and $\mathbf{V} = \{V_1, V_2, \cdots, V_K\}$ and $\boldsymbol{\theta}^* = \{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \cdots, \boldsymbol{\theta}_K^*\}$ as the global latent variables; we write $\boldsymbol{\eta}$ for the vector of unknown parameters $\{\alpha, \boldsymbol{\beta}, \boldsymbol{\psi}\}$. The joint distribution of $\mathbf{U}$, $\mathbf{U}^*$, $\mathbf{S}$, $\mathbf{V}$, $\boldsymbol{\theta}^*$ and $\mathbf{Z}$ (suppressing the fixed covariates) is given by

$$p(\mathbf{U}, \mathbf{U}^*, \mathbf{Z}, \mathbf{S}, \mathbf{V}, \boldsymbol{\theta}^*|\boldsymbol{\eta}, c) = p(\mathbf{U}, \mathbf{U}^*|\alpha, \boldsymbol{\beta})p(\mathbf{Z}|\mathbf{U}, \mathbf{S}, \boldsymbol{\theta}^*, \boldsymbol{\psi})p(\mathbf{S}|\mathbf{V})p(\boldsymbol{V}|c)p(\boldsymbol{\theta}^*).$$

The observed-data likelihood is obtained by integrating out the latent quantities, which touches on the high-dimensional integration. In the following, we pursue Bayesian analysis for $\boldsymbol{\eta}$.

2.3. Prior

Bayesian analysis requires assigning priors to the parameters and/or hyperparameters to complete Bayesian model specification. In the current context, the parameters are consisted of γ, $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$, and the hyper-parameters are formed by F_0 and c. By model convention, we assume they are mutually independent.

First, we assign a conjugate normal prior $N(\alpha_0, \sigma_{\alpha 0}^2)$ to α, where α_0 and $\sigma_{\alpha 0}^2$ are the hyper-parameters treated as fixed. The α_0 is usually set at zero and $\sigma_{\alpha 0}$ is taken a large value to ensure the distribution of α inflated enough. In the situation where the informative information about α_0 and $\sigma_{\alpha 0}$ can be available, one could expect such prior will be helpful in improving statistical inference. Next, we assume $F_0 = Ga(\mu_{\gamma 0}, \sigma_{\gamma 0}^2) \times IG(\alpha_{\epsilon 0}, \beta_{\epsilon 0})$, where '$Ga$' and '$IG$' refer to the Gamma and inverse Gamma distributions respectively. This is also the conjugate prior of γ and σ^2 under the parametric setting. Similarly, to avoid overly subjective information, the values of $\mu_{\gamma 0}$, $\sigma_{\gamma 0}^2$, $\alpha_{\epsilon 0}$, and $\beta_{\epsilon 0}$ can be chosen to

ensure F_0 to be dispersed enough. For example, setting $\alpha_{\epsilon 0} = \beta_{\epsilon 0} = 2.0$ produces the mean 2 and variance infinity for σ^2. However, some cautions should be imposed on the choice of concentration parameter c since it controls the amount of clustering and hence is critical for inference about f_ϵ, the density function of ϵ_i. To model c properly, we follow the routine in [23] and assign $c \sim Ga(\tau_1, \tau_2)$ with small values τ_1 and τ_2 (e.g., $\tau_1 = \tau_2 = 2.0$) which will encourage both small and large values for c.

Finally, we assign the following Laplace (double-exponential) priors to $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ for variable selection:

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}_\beta) = \frac{1}{2}\prod_{k=1}^{q} \gamma_{\beta k} e^{-\gamma_{\beta k}|\beta_k|},\ p(\boldsymbol{\psi}|\boldsymbol{\gamma}_\psi) = \frac{1}{2}\prod_{k=1}^{q} \gamma_{\psi k} e^{-\gamma_{\psi k}|\psi_k|}, \tag{2.8}$$

where $\gamma_{\beta k}(>0)$ and $\gamma_{\psi k}(>0)$ are the tuning parameters used to control the amount of shrinkage of β_k and ψ_k respectively.

Laplace distribution is usually used to fit the distribution of errors or random effects to downweight the influence of outliers(e.g., [25]). In the context of regression shrinkage and selection, Tibshirani [49] noted that his pioneered lasso estimates are equivalent to the Bayes posterior mode under the independent Laplace priors for the regression coefficients. Park and Casella [40] later extended it to the situation with Bayesian feature extraction and developed Bayesian Lasso formally. In their seminar paper, Park and Casella [40] set the tuning parameters to be homogeneous across the components of regression coefficients. Here we adopt Zou's adaptive lasso [56] which allows tuning parameters to vary with the components.

Similar to [40], we recast prior (2.8) as hierarchical model as follows

$$p(\boldsymbol{\beta}|\boldsymbol{\tau}_\beta^2) = N_q(\mathbf{0}, diag(\boldsymbol{\tau}_\beta^2)),\ p(\boldsymbol{\tau}_\beta^2|\boldsymbol{\gamma}_\beta) = \prod_{k=1}^{q} Exp(\gamma_{\beta k}^2/2), \tag{2.9}$$

$$p(\boldsymbol{\psi}|\boldsymbol{\tau}_\psi^2) = N_q(\mathbf{0}, diag(\boldsymbol{\tau}_\psi^2)), p(\boldsymbol{\tau}_\psi^2|\boldsymbol{\gamma}_\psi) = \prod_{k=1}^{q} Exp(\gamma_{\psi k}^2/2), \tag{2.10}$$

where '$Exp(\lambda)$' is the exponential distribution with mean $1/\lambda$; $\boldsymbol{\tau}_\beta^2 = (\tau_{\beta 1}^2, \cdots, \tau_{\beta q}^2)^T$ and $\boldsymbol{\tau}_\psi^2 = (\tau_{\psi 1}^2, \cdots, \tau_{\psi q}^2)^T$ are the latent variables with $\tau_{\beta k} > 0$, $\tau_{\psi k} > 0 (k = 1, \cdots, q)$.

The choices of $\boldsymbol{\gamma}_\beta^2$ and $\boldsymbol{\gamma}_\gamma^2$ should be selected with care since they determine the amount of shrinkage of regression directly. It follows from (2.8) that larger values of γ_β and γ_ψ favor more penalty on the regression coefficients. Hence, we follow the routine in [40] and assign gamma priors to them, i.e.,

$$p(\boldsymbol{\gamma}_\beta^2) = \prod_{k=1}^{q} p(\gamma_{\beta k}^2) = \prod_{k=1}^{q} Ga(a_{k0}, b_{k0}),\quad p(\boldsymbol{\gamma}_\psi^2) = \prod_{k=1}^{q} p(\gamma_{\psi k}^2) = \prod_{k=1}^{q} Ga(c_{k0}, d_{k0}),$$

where the values of a_{k0}, b_{k0}, c_{k0} and d_{k0} are set to encourage small and large values for $\boldsymbol{\gamma}_\beta^2$ and $\boldsymbol{\gamma}_\psi^2$ respectively. In the context of Bayesian adaptive lasso for

ordinal regression analysis, Feng *et al.* [12] suggested using shape unity and scale 0.05 in gamma prior to enhance the robustness of inferences. Their routine is also followed in our empirical study.

2.4. Posterior Distribution

With the priors given above, the statistical inference about $\boldsymbol{\eta}$ is based on the posterior distribution $p(\boldsymbol{\eta}|\mathbf{U}, \mathbf{Z})$. Under the MCMC sampling framework, the posterior analysis is carried out via data augmentation technique [48], i.e., augmenting the observed data with latent variables to form the complete-data. The inference is made on the basis of the joint posterior of latent quantities and unknown parameters. Apparently, the joint posterior distribution is given by

$$p(\mathbf{U}^*, \mathbf{S}, \boldsymbol{\tau}_\beta^2, \boldsymbol{\tau}_\psi^2, \boldsymbol{\gamma}_\beta^2, \boldsymbol{\gamma}_\psi^2, \mathbf{V}, \boldsymbol{\theta}^*, \boldsymbol{\eta}, c|\mathbf{U}, \mathbf{Z}). \tag{2.11}$$

Markov Chains Monte Carlo ([17]; [18]) sampling, in particular, the blocked Gibbs sampler ([15]; [23]) is implemented by drawing observations from the full conditional distributions of (2.11). Upon the convergence, the posterior analysis is conducted based on the simulated observations. Surely, MCMC is very powerful and particularly suitable to the situations with complex data structure and/or hierarchical model. However, as a stochastic approximation method, MCMC sampling also suffers some limitations, e.g., slow convergence and time-consuming. In particular, the sampling errors often make model comparison infeasible since they may vary with the competing models. In this paper, we consider variational Bayesian inference. Compared with the Monte Carlo sampling method, variational Bayes provides a deterministic solution.

3. VARIATIONAL BAYESIAN INFERENCE

3.1. Variational density

For ease of exposition, we write $\mathbf{W}$ as the collection of latent quantities in the posterior and $\mathbf{D}$ as the collection of observed data. Rather than working with $p(\mathbf{W}|\mathbf{D})$, variational Bayes aims to find a computationally feasible distribution $q(\mathbf{W})$ within the variational density family to approximate posterior. The approximation is achieved via maximizing the following evidence lower bound (ELBO)

$$\text{ELBO}(q, \mathbf{D}) = \int q(\mathbf{W}) \log(p(\mathbf{D}, \mathbf{W})/q(\mathbf{W}))d\mathbf{W} \leq \log p(\mathbf{D}) \tag{3.1}$$

where $p(\mathbf{D})$ is the model evidence and $p(\mathbf{D}, \mathbf{W})$ is the joint density of $\mathbf{D}$ and $\mathbf{W}$. Generally, the complexity of optimization (3.1) depends on the complexity of the

variational density family. If the variational density family is restricted within the mean-field family, i.e., $q(\mathbf{W}) = \prod_{m=1}^{M} q_m(\mathbf{w}_m)$, then it can be shown that the optimal solution satisfies

$$q_j(\mathbf{w}_j) \propto \exp\left\{E_{q\setminus j} \log p(\mathbf{D}, \mathbf{W})\right\}, (j = 1, \cdots, m) \tag{3.2}$$

where $E_{q\setminus j}$ denotes the expectation with respect to all variational density factors but q_j. Notice that equation (3.2) provides an iterative solution to $q(\cdot)$. It depends on the ordinates $q_m : m \neq j$ conditionally. Hence, the coordinate ascent variation inference algorithm can be implemented by solving $q_j : j = 1, \cdots, M$ via (3.2) one by one till the convergence of ELBO is achieved.

Let's return to the problem (2.11). Based on the model formulation, it is natural to consider the following mean-field variational family $\mathcal{Q}$:

$$q(\mathbf{U}^*, \mathbf{S}, \mathbf{V}, \boldsymbol{\theta}^*, \boldsymbol{\eta}, c) = q(\mathbf{U}^*)q(\mathbf{S})q(\mathbf{V})q(\boldsymbol{\theta}^*)q(\boldsymbol{\eta})q(c).$$

where $q(\boldsymbol{\eta}) = q(\alpha)q(\boldsymbol{\beta})q(\boldsymbol{\psi})$. Furthermore, we take the variational density factors as follows:

$$\begin{aligned}
&q(\mathbf{U}^*) = \prod_{i=1}^{n} q(u_i^*) = \prod_{i=1}^{n} \mathrm{PG}(1, \eta_i^u),\ q(\mathbf{S}) = \prod_{i=1}^{n} q(s_i|\phi_{i1}, \cdots, \phi_{iK}),\\
&q(\mathbf{V}) = \prod_{k=1}^{K-1} q(V_k) = \prod_{k=1}^{K-1} Beta(\zeta_{k1}, \zeta_{k2}),\ q(c) = Ga(\alpha_c, \beta_c),\\
&q(\boldsymbol{\theta}^*) = \prod_{k=1}^{K} q(\gamma_k^*, \sigma_k^{*2}) = \prod_{k=1}^{K} IG(\sigma_k^{*2}|\alpha_{\epsilon k}, \beta_{\epsilon k}) \times N(\gamma_k^*|\mu_{\gamma k}, \sigma_k^{*2}\sigma_{\gamma k}^2),\\
&q(\boldsymbol{\beta}) = N_q(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta),\ q(\boldsymbol{\psi}) = N_q(\boldsymbol{\mu}_\psi, \boldsymbol{\Sigma}_\psi),\\
&q(\boldsymbol{\tau}_\beta^{-2}) = \prod_{j=1}^{q} IGauss(\mu_{\beta j}, \lambda_{\beta j}),\quad q(\boldsymbol{\tau}_\psi^{-2}) = \prod_{j=1}^{q} IGauss(\mu_{\psi j}, \lambda_{\psi j}),\\
&q(\boldsymbol{\gamma}_\beta^2) = \prod_{j=1}^{q} Ga(a_{\beta j}, b_{\beta j}),\quad q(\boldsymbol{\gamma}_\psi^2) = \prod_{j=1}^{q} Ga(a_{\psi j}, b_{\psi j}),
\end{aligned}$$

where '$IGauss(\mu, \lambda)$' denotes the inverse Gaussian distribution with mean μ and scale λ; the scalars η_i^u, ϕ_k, ζ_{k2}, ζ_{k2}, α_c, β_c, $\alpha_{\epsilon k}$, $\beta_{\epsilon k}$, $\mu_{\gamma k}$, $\sigma_{\gamma k}^2$, $\mu_{\beta j}$, $\lambda_{\beta j}$, $\mu_{\psi j}$, $\lambda_{\psi j}$, $a_{\beta j}$, $b_{\beta j}$, $a_{\psi j}$, $b_{\psi j}$, the vectors $\boldsymbol{\mu}_\beta$, $\boldsymbol{\mu}_\psi$, and the matrices $\boldsymbol{\Sigma}_\beta$, $\boldsymbol{\Sigma}_\psi$ are the variational parameters which are required to be estimated.

We implement coordinate ascent (CA) variational algorithm as follows:
Step 1: Give the initial values of variational parameters;
Step 2: Update variational parameters via coordinate ascent algorithm;
Step 3: Compute ELBO;
Step 4: Repeat Steps 2 and 3 till convergence.

Upon the convergence, the variational Bayesian estimate of $\boldsymbol{\eta}$ and its variance and covariance estimates are given by $E_q\boldsymbol{\eta}$ and $Cov_q(\boldsymbol{\eta})$ respectively, where

the expectations are taken with respect to the variational density evaluated at the converged variational parameters. In particular, the posterior predictive distribution of ϵ_i is given by

$$\widehat{f}(\epsilon) = \sum_{k=1}^{K} E_q(\pi_k) E_q p(\epsilon|\gamma_k^*, \sigma_k^{*2}) = \sum_{k=1}^{K} (E_q \pi_k) t_{\nu_k}(\epsilon|\mu_{\gamma k}, m_k^2) \tag{3.3}$$

where $\nu_k = 2\alpha_{\epsilon k}$, $m_k^2 = 2\beta_{\epsilon k}(1 + \sigma_{\gamma k}^2)$, and $t_\nu(a, \sigma^2)$ is the t-distribution with degrees of freedom ν, location a and dispersion σ^2.

As noted in [5], ELBO is (generally) a nonconvex objective function and coordinate ascent variational algorithm only guarantees convergence to a local optimum, which can be sensitive to the initialization. In practice, one can implement a small amount of MCMC sampling to explore rough starting values. The convergence of algorithm can be monitored by observing the change of ELBOs against the number of iterations and terminating iterations when $|\text{ELBO}^{(k+1)} - \text{ELBO}^{(k)}| < \epsilon$, where $\epsilon > 0$ is a previously specified positive number. A simpler way is to observe the stability of the estimates by plotting their traces against the number of iterations. The update scheme of variational parameters is given in Section 1 in the supplementary material.

Computing ELBO involves the calculations of integrals of complete-data log-likelihood with respect to the variational density q and the K-L divergence between the prior densities and target distribution, that is,

$$\text{ELBO} = E_q \log p(\mathbf{U}, \mathbf{U}^*, \mathbf{Z}, \mathbf{S}, \boldsymbol{\theta}^*, \mathbf{V}^*|\boldsymbol{\eta}, c) + E_q \log p(\boldsymbol{\eta}, c) - E_q q(\mathbf{U}^*, \mathbf{S}, \mathbf{V}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}, c). \tag{3.4}$$

Due to the independence of variational density factors, most calculations are straightforward. However, the expectations $E_q \log q(u_i^*)$ are more intractable since they touch on the infinite sum. We address this issue in Section 2 in the supplementary material.

3.2. Variable selection and model determination

The Laplace prior introduced before is mainly used to shrink regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$. However, unlike that under the frequenist statistical framework, Bayesian lassos produce the estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\psi}}$ not exactly equal to 0. This requires determining whether or not the regression coefficient is zero. The problem is generally solved via posterior confidence intervals. For example, [12] suggested a hard threshold guideline to select variables. Note that in the current analysis, both of $q^*(\boldsymbol{\beta})$ and $q^*(\boldsymbol{\psi})$ are the multivariate normal distributions. Therefore, we assume that: if

$$|\widehat{\beta}_j|/\text{sd}(\widehat{\beta}_j) < z_{\alpha/2}, \quad |\widehat{\psi}_j|/\text{sd}(\widehat{\psi}_j) < z_{\alpha/2}, \tag{3.5}$$

then β_{xj} and ψ_{xj} can be considered as 0, where z_α is the upper α percentile point of standard normal distribution and α is the nominal level specified in advance, e.g., $\alpha = 0.05$.

Beyond the estimation, it is practically important to evaluate the adequacy of model fits. In the current context, it is of interest to determine whether the semi-parametric fitting is helpful in improving model fits. In the Bayesian paradigm, the model evaluation is often addressed via model selection procedure. Among various-easy-constructs, we prefer using the Akaike information criterion (AIC; [1]) and the Bayesian information criterion (BIC, [45]) for simplicity. The AIC and BIC are formally defined as AIC $= -2\log\widehat{L} + d\log(n)$ and BIC $= -2\log\widehat{L} + 2d$, where $\widehat{L}$ is the observed-data likelihood evaluated at the maximum likelihood (ML) estimate of the unknown parameters, d is the model complexity and n is the sample size of the observations. The model with the smallest value(s) of AIC and/or BIC is being selected. In practice, the Bayesian estimates are close to the ML estimates, hence they can be used to replace the ML estimates in computing AIC and BIC. In this paper, we employ the variational Bayesian estimates.

4. SIMULATION STUDY

In this section, a simulation study is conducted to assess the performance of the proposed method. The main objective is to investigate the accuracy of variational Bayesian estimates, the model adequacy via variable selection and the model selection among the competing models. We consider one semi-continuous variable of which the indicator variable u_i satisfies equation (2.1) while given $u_i = 1$, the intensity variable z_i is generated from the following equations:

- scenario I: $z_i - \mathbf{x}_i^T\boldsymbol{\psi} \sim N(0.8, 1.0)$;
- scenario II: $z_i - \mathbf{x}_i^T\boldsymbol{\psi} \sim 0.3N(-3.5, 1.0) + 0.7N(3.0, 4.0)$;
- scenario III: $z_i - \mathbf{x}_i^T\boldsymbol{\psi} \sim \sum_{k=0}^{7} \frac{1}{8} N(3\{(\frac{2}{3})^k - 1\}, (\frac{2}{3})^{2k})$;

in which scenario I denotes the single normal model, scenario II represents the multi-modal model with two modes, and scenario III, followed by [36], is the strongly skewed model (see Figure 1).

We generate five fixed covariates from the multivariate normal distribution $N_5(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{jk} = 0.5^{|j-k|}(j, k = 1, \cdots, 5)$, where Σ_{jk} is the (j, k) element of $\boldsymbol{\Sigma}$. The true values of population parameters are set as follows: $\alpha = 0.7$, $\boldsymbol{\beta} = (0.7, 0.0, 0.7, 0.0, 0.7)$, $\boldsymbol{\psi} = (0.0, 0.0, 0.7, 0.7, 0.7)$. To investigate the effect of sample size on the accuracy of estimates, we take $n = 300$ and 1000, which represents the small and large sample sizes.

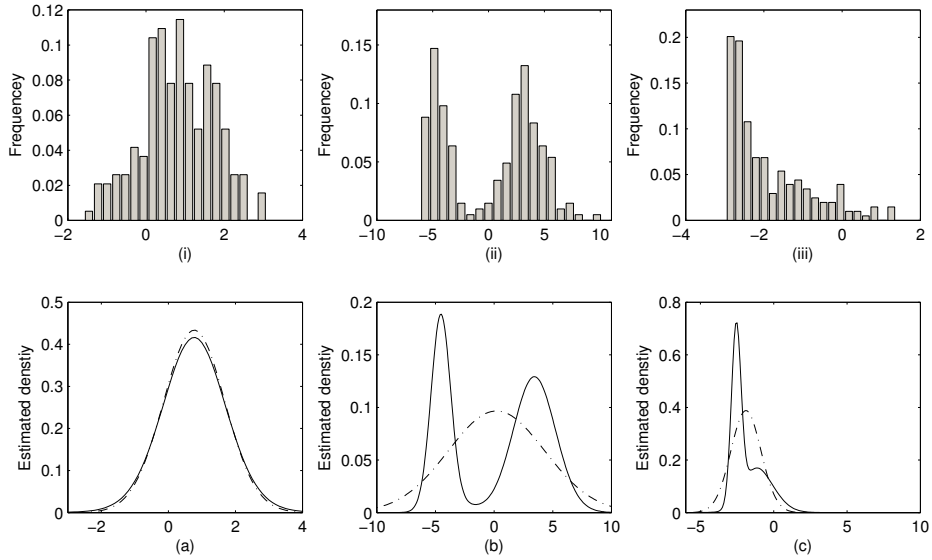

Figure 1: Plot of the histograms of data and the variational Bayesian density estimates under $n = 300$. Panels (i) to (iii) correspond to the data under scenarios I to III, and panels (a) to (c) correspond to the posterior predictive densities: the solid lines represent the semiparametric fitting ($K = 100$) and the dotted lines denote the parametric fitting.

For the Bayesian analysis, we consider the following inputs for the hyper-parameters: for the intercept parameter α, we take $\alpha_0 = 0$ and $\sigma^2_{\alpha 0} = 100.0$. The large variance $\sigma^2_{\alpha 0}$ ensures a large interval with high probability to accommodate initial value of α; for the baseline distribution F_0, we set mean and variance of γ^*_k as the same as that of α, while the scale $\beta_{\epsilon 0}$ and the shape $\alpha_{\epsilon 0}$ in the inverse gamma are fixed at 2.0; the inputs of hyper-parameters in the Laplace priors are taken as the same as those in [12]. Finally, for τ_1 and τ_2, we set them at 2.0. Note that these values can guarantee these distributions to be less informative.

The proposed algorithm given in Section 3 is implemented to produce the variational Bayes estimates. The truncated level involved in (2.5) is taken as $K = 100$. Before formal implementation, a few test runs are conducted as a pilot to monitor the convergence of VB algorithms. For comparison, we also investigate the convergence of MCMC sampling via blocked Gibbs sampler [22]. The full conditionals involved in the blocked Gibbs sampler are easily derived and omitted to save spaces. We examine the trace plots of the estimates against the number of iterations under three different starting values. Figure 2 gives the traces of EPSR (e.g.,[16]) of unknown parameters in the first 64 iterations for the MCMC sampling, and $\|\Delta\boldsymbol{\eta}^{(t)}\| = \|\boldsymbol{\eta}^{(t+1)} - \boldsymbol{\eta}^{(t)}\|$ in the first 30 iterations for the CA algorithm. It can be seen that the CA algorithm converges faster than the blocked Gibbs sampling. It only needs about three iterations to achieve convergence, while

Gibbs sampler requires about at least 50 iterations to guarantee EPSR less than 1.2. Moreover, MCMC sampling algorithm illustrates strong variation due to its stochastic approximation, while VB produces a deterministic solution.

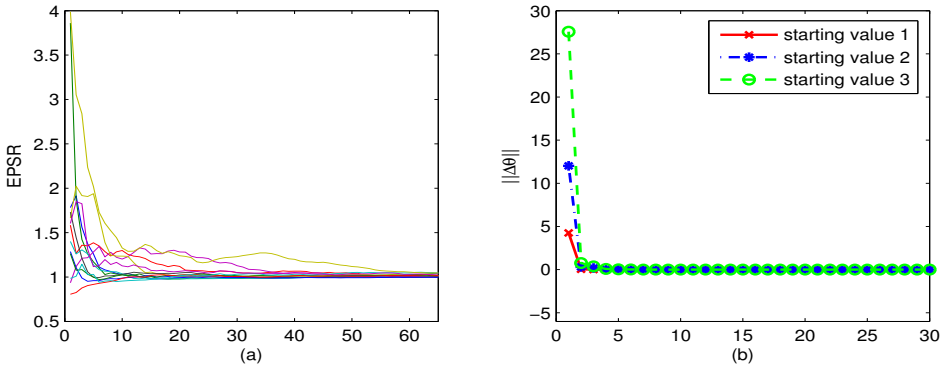

Figure 2: Plots of the values of EPSR and $\|\Delta\boldsymbol{\eta}^{(t)}\|$ of unknown parameters against the number of iterations under three different starting values ($n = 300$): (a) MCMC, (b) VB.

We calculate the bias(BIAS), the root of mean squares(RMS) and the standard deviations(SD) of VB estimates under the parametric and semiparametric fittings across 100 replications. The resulting summary is given in the section 3 in the supplementary material (see Tables 1 and 2 therein). Examinations of results show that: (i) for scenario I, the performance of VB estimates under semiparametric fitting is the same as those under parametric fitting. The BIAS, RMS and SD of two estimates are exactly identical regardless of $n = 300$ or 1000; This suggests that the semiparametric fitting can automatically adjust their random weights (2.5) to shrink data into one cluster to meet the assumption of the single parametric model; (ii) for scenario II, we find there exist larger differences between two estimates. The totals of RMS and SD under normal fitting equal to 2.304 and 1.703 at $n = 300$, while under semi-parametric fitting amount to 1.403 and 1.032. This reveals the fact that the normal fitting accommodates multiple modes via inflating the variances of the estimates; (iii) for scenario III, though not significant, the VB estimates under semiparametric fitting still outperforms those under parametric fitting; (iv) As expected, with the increase of sample sizes, two estimates becomes more and more accurate but the differences between them are still not ignorable when the posited models are not specified correctly; (v) Since the estimates of regression coefficients involved in Part one are not affected by the semiparametric fitting in Part two, the behavior of VB estimates in Part one under two fittings are wholly identical.

Another simulation focuses on the performance of variable selection using (3.5) under the parametric and semi-parametric Bayesian fittings. In this study, the simulation design is taken as the same as scenario II except that the sample sizes are set at 300, 500 and 1000, and the true values of regression coefficients are taken as $\boldsymbol{\beta} = (1, 0, 1, 0, 1)^T$ and $\boldsymbol{\psi} = (0, 0, 1, 1, 1)^T$. Table 1 presents the

Table 1: Summary of variable selection in the simulation study: scenario II data.

Para.	True value	Norm.				Semi.		
		$n = 300$	$n = 500$	$n = 1000$		$n = 300$	$n = 500$	$n = 1000$
β_1	$\neq 0$	100	100	100		–	–	–
β_2	$= 0$	96	95	97		–	–	–
β_3	$= 0$	100	100	100		–	–	–
β_4	$\neq 0$	93	97	98		–	–	–
β_5	$= 0$	100	100	100		–	–	–
ψ_1	$= 0$	94	95	98		96	96	98
ψ_2	$= 0$	95	97	96		96	97	97
ψ_3	$\neq 0$	74	98	100		100	100	100
ψ_4	$\neq 0$	80	100	100		100	100	100
ψ_5	$\neq 0$	83	98	100		99	100	100

summary of variable selection over 100 replications, in which the summary related to $\boldsymbol{\beta}$ under the semiparametric fitting are omitted. Based on Table 3, it can be found that for ψ_1 and ψ_2, the semiparametric fitting produces the result similar to that under the parametric fitting. However, for $\psi_j (j = 3, 4, 5)$, the false rates under the parametric setting are higher than those under the semiparametric fitting. This is not surprising since our variable selection method is closely related to the standard deviations, and parametric method easily inflates the standard deviations when model is misspecified.

Finally, as suggested by two reviewers, we consider two special cases for ϵ_i. One case is that ϵ_i has the t distribution with 4 degrees of freedom, and the other is that ϵ_i contains outliers. The former corresponds to the heaver-than-normal data while the latter represents the contaminated data. For the latter, we generate ϵ_is independently from the standard normal distribution and add 10.0 to them with proportions 5% and 10%, respectively. The sample size, analogous to the real example, is taken as $n = 1000$. The other model settings are the same as the previous situation. We calculated BIAS, RMS and SD of the estimates of regression coefficients ψ_j under the parametric and semiparametric fittings respectively. The summary is reported in section 3 in the supplementary material (see Table 3 therein). We also investigate the performance of feature selection via confidence intervals. Figure 4 gives the box plots of the estimates of regression coefficient ψ_j under two different fittings. Examinations of the simulated results and Figure 4 give the following facts: for the heavy-tailed and symmetric data with single mode, the semiparametric fitting is very close to the parametric fitting and slightly outperforms the parametric fitting. The underlying reason is that for the unimodal and symmetric data, the atoms in the semiparametric fitting are adaptively clustered into one group and the fitting model behaves more like the parametric model. Instead, for the contaminated data, the results produced by the semiparametric method are better than those under the parametric method. Moreover, as the levels of contamination increase, the estimates produced by semiparametric method are more stable and their confidence intervals

are uniformly narrower. It reveals that the proposed method is rather effective in dealing with outliers.

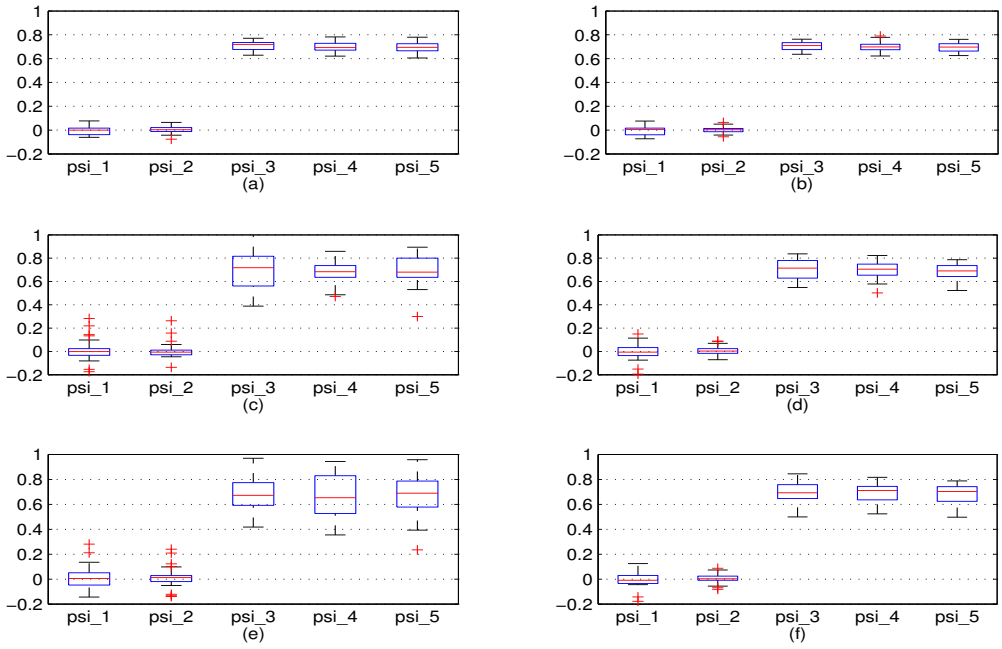

Figure 3: Boxplots of ψ_j under the parametric and semiparametric fittings: panels (a),(c) and (e) corresponds to the parametric fitting and panels (b), (d) and (f) correspond to the semiparametric fitting; The panels in the first row correspond to the heavier-than-normal data while the panels in the second and last rows correspond to the contaminated data with contamination levels at 5% and 10% respectively. The sample size $n = 1000$.

In summary, although the estimates of regression coefficients are less dependent on the distributional assumptions of the responses, the standard deviations obtained under the parametric fitting are more sensitive to the model specifications. In contrast, as a correction to the parametric method, the semiparametric method can accommodate various assumptions of the underlying distribution, thus more robust against the distributional deviations. For computation, all programs are coded in C language and implemented on Inter(R) Core(TM), i5-6500 processor with CPU 3.20GHz on Microsoft Windows 7 operating system. For $n = 1000$, it takes about two minutes to complete 100 replications for semiparametric fitting with $K = 100$ and 30 seconds to complete 100 replications for parametric fitting. Request on codes can be send to the corresponding author.

5. HOUSEHOLD FINANCIAL SURVEY DATA

In this section, we analyze a small portion of household financial debt data to illustrate the actual merits of the methodology. The data set is selected from the China Household Finance Survey(CHFS)(https://chfs.swufe.edu.cn) conducted by the institute of China Finance survey, a nonprofit institute organized by the Southeast University of Finance and Economics. The survey covers a series of questions which touch on the information about various aspects of the household's financial situation. In this study, we only focus on the measurement 'gross debts per household (DEB)', the amount of the secured debt and unsecured debt of a household under investigation. This is a primary measure of interest in the finance survey. It indicates whether a family would be willing to hold the financial debt and if so, how much debt a household holds. After removing the missing data, the sample size is 1047. A rough data analysis shows that the measurement DEB contains excessive zeros and the proportion of zeros is about 72.58%. Naturally, we treat this measurement as the semi-continuous y_i, and identify it with u_i and z_i. Figure 4 presents the histogram of DEB as well as the logarithms of the positive values. It can be seen clearly that data set illustrates strong heterogeneity. The skewness and kurtosis of DEB are 1.1042 and 2.3361, respectively, which indicates that single parametric model on DEB may be unappreciated.

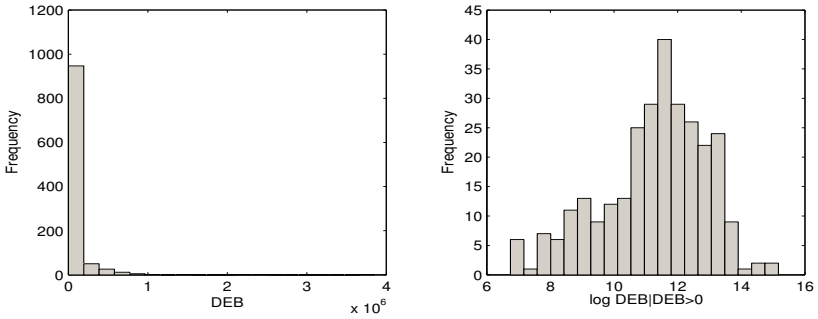

Figure 4: Histograms of DEB and the logarithms of their positive values: China household finance survey data. Left panel corresponding the DEB and right panel corresponding to log(DEB|DEB > 0).

Motivated by the existing economic literature (e.g.,[44]), the following measurements were included as the explanatory factors to explore the causal effects: gender (x_1; scaled via 0-1 point: 1 to male and 0 to female), age(aged from 21 to 94), marital status (x_5; scaled by points 2 to 6), health condition (x_6), educational degrees (x_7; scaled via 1 to 7 according to the educational level), employment(x_8; scaled on 0-1: 0 to unemployed and 1 to employed), the number of adults (aged > 16) in a family (x_9); and the household income (x_{10}). Moreover, in view of the large spread of ages, we group the subjects into five categories according to their ages, each assigned to a binary indicator: aged from 21 to 35(x_2; including 35, the same below), 35 to 45 (x_3), 45 to 60 (x_4), and over 60. Moreover, we refer to

Table 2: Descriptive statistics of explanatory variables: CHFS data .

Variable.	Description.	Mean.	Max.	Min.	SD
Gender (x_1)	=1, male; =0, otherwise	0.756	1	0	0.430
Age:					
aged 21 to 35 (x_2)	=1, yes; 0, otherwise	0.061	1	0	0.239
aged 36 to 45 (x_3)	=1, yes; 0, otherwise	0.132	1	0	0.339
aged 46 to 60 (x_4)	=1, yes; 0, otherwise	0.240	1	0	0.427
Marital status (x_5)	=1, married; 0, otherwise	0.863	1	0	0.344
Health condition (x_6)	=1, good; 0, otherwise	0.833	1	0	0.373
Education degree (x_7)	=1, high school or above; =0, otherwise	0.352	1	0	0.478
Employment (x_8)	=1, yes; 0, otherwise	0.092	1	0	0.290
No. of adults (x_9)		3.002	3	0	1.301
Income (CYN)(x_{10})*		2000	20000	1000	0.232

Note: the measurement is the middle value of the option in the questionnaire.

the last group as the reference and remove it for model identification. As a result, there are ten covariates in total. Table 2 provides the description summary of the explanatory variables under consideration. To unify the scale, all covariates are standardized.

We fit the parametric and semi-parametric models mentioned before to the data. Similar to those in the simulation study, the inputs of hyperparameters in the prior are taken as follows: $\alpha_0 = \gamma_0 = 0$, $\sigma^2_{\alpha 0} = \sigma^2_{\gamma 0} = 0.01$, $a_0 = c_0 = 1.0$ and $b_0 = d_0 = 0.05$, $\alpha_{\epsilon 0} = \beta_{\epsilon 0} = 2.0$ and $\tau_1 = \tau_2 = 2.0$. Moreover, we investigate the effects of truncation levels in the semiparametric fitting on model fits. We implement CA algorithm to calculate the estimates of AIC and BIC across competing models. The convergence of algorithm is monitored by observing the trace of $\|\Delta\boldsymbol{\eta}\|$. We terminate the cycles when $\|\Delta\boldsymbol{\eta}\|$ is less than 1.0×10^{-6}. Table 3 presents the summary of AIC and BIC in the selection of six competing models. It can be seen that the values of AIC and BIC under the normal mixture model with $K = 10$ are the minimum among competing models. Hence such a model is chosen as the posited model. Further examinations show that large values of K are not necessarily to favor the improvement of model fits. As an illustration, Figure 5 gives the posterior density estimates of $f(\epsilon)$ under the parametric and semiparametric fittings with $K = 10$ based on 400 grids on the interval $[0, 20]$. It can be seen clearly that semiparametric fitting captures the left skewness of data successfully while parametric fitting fails.

Table 4 presents the estimates of unknown parameters and their standard deviations obtained under the parametric model and semiparametric mdoel with $K = 10$, in which the variables being selected via hard threshold method (e.g., [12]) (denoted by HT) and confidence interval method (3.5) (denoted by CI) are reported in the columns four, five, eight and nine. Due to the same reason in the simulation study, the estimates of unknown parameters involved in Part one are not reported to save spaces. Based on Table 4, it can be found that: (i) by the

Table 3: Summary of the model selection in analyzing China household finance data: 'norm' denotes the parametric fitting while 'Semi.' refers to the semiparametric fitting.

	Norm.	Semi.				
		$K = 2$	$K = 5$	$K = 10$	$K = 20$	$K = 100$
AIC	3212.565	3003.3928	5186.536	2837.479	2897.661	3378.497
BIC	3326.499	3132.1887	5359.915	3085.163	3293.955	4963.676

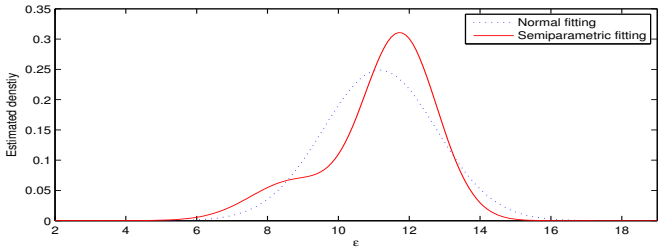

Figure 5: Plot of the posterior density estimates $\widehat{f}_\epsilon(\epsilon)$ in Part two: blue doted linear corresponds to the parametric fitting while red solid line corresponds to the semiparametric fitting with $K = 10$.

measures of HT and CI, the gender, the marital status and the health condition are all excluded from Part one, while the age, the employment, the number of adults and the household income are included as the exogenous factors to interpret the variability of u_i. This indicates that the proportion of households in holding financial debts depends less on the gender, the martial status and the health condition of the households. Instead, they rely on the ages, the employment of the head of the household and the household income. However, for the educational degree, there is a conflict between TH and CI: the former favors it while the latter discards it. Recall that in China, most household financial debts are consisted of the secured debts such as mortgage, care loans, liabilities for production and operation and so on, hence, these selected variables in some extent reflect the basic situations of the household finance in China. (ii) for Part two, there exists obvious difference in selecting relevant variables under the parametric and semi-parametric fittings. For the parametric fitting, only household income is included to explain the amount of household financial debts, while under the semiparametric fitting, besides it, health condition is also included. That is, the actual level of household financial debts is affected by the member's health condition of a household. Such a difference indicates that the skewness of data has an important impact on the choice of relevant variables.

Table 4: Summary of the estimates of unknown parameters and the variable selection in the analysis of household financial survey data: 0 corresponding to the variables being included and 1 corresponding to the variables being excluded.

	Norm.				Semi.			
Para.	Est.	SD	HT	CI	Est.	SD	HT	CI
α	-1.104	0.066	–	–	–	–	–	–
β_1	-0.056	0.056	1	1	–	–	–	–
β_2	0.354	0.078	0	0	–	–	–	–
β_3	0.648	0.082	0	0	–	–	–	–
β_4	0.475	0.083	0	0	–	–	–	–
β_5	0.029	0.056	1	1	–	–	–	–
β_6	0.030	0.052	1	1	–	–	–	–
β_7	0.108	0.065	0	1	–	–	–	–
β_8	0.154	0.062	0	0	–	–	–	–
β_9	0.167	0.067	0	0	–	–	–	–
β_{10}	0.173	0.068	0	0	–	–	–	–
ψ_1	-0.016	0.064	1	1	-0.010	0.047	1	1
ψ_2	-0.028	0.068	1	1	-0.007	0.05	1	1
ψ_3	0.018	0.065	1	1	-0.003	0.048	1	1
ψ_4	0.025	0.071	1	1	-0.008	0.052	1	1
ψ_5	0.096	0.091	1	1	0.025	0.056	1	1
ψ_6	0.084	0.076	1	1	0.142	0.059	0	0
ψ_7	0.003	0.061	1	1	0.064	0.053	1	1
ψ_8	0.068	0.067	1	1	0.029	0.044	1	1
ψ_9	-0.008	0.067	1	1	-0.042	0.055	1	1
ψ_{10}	0.404	0.090	0	0	0.453	0.061	0	0

6. DISCUSSION

In analyzing semi-continuous data, TPM is surely a popular method to identify the pattern of how the exogenous factors work on two parts of data. Semiparametric fitting for TPM is a natural extension of the ordinary TPM to the situation where the robust issue is to take into account to downweight the influence of distributional deviations on the continuous part. To relax the crucial assumption on the normal distribution of the idiosyncratic part, we propose a normal mixture model in which the mixing distribution is treated to be arbitrary and random, and selected from the space of distribution functions according to DP prior. By taking advantage of discreteness of DP, the proposed method induces a normal mixture model with at most finite atoms. Such modeling strategy is very appreciated in the Bayesian density estimation. It allows us to select the posited model via data-driven technique. However, with the increase of the

model complexity, the computation becomes more challenging. We develop a variational Bayesian procedure. Compared with the MCMC sampling method, the variational Bayes inference enjoys high computational efficiency and deterministic solution. Within the mean-field family framework, the variational density as well as the update scheme of variational parameters are obtained via coordinate ascent algorithm. Posterior inferences including parameter estimation, posterior density estimates, variable selection and model evaluation are carried out based on the variational distribution.

The closed form approximation to the posterior distribution of the parameters in our proposal benefits from the Pólya-Gamma stochastic representation of the logistic function in Part one, in which the nonlinear function of parameters is decomposed into the pattern as that in the normal setting. It leads to a conjugate posterior with Gaussian prior over the regression coefficients. In the case of Bayesian logistic regression analysis, Jaakkola and Jordan [24] proposed a nice approximation to the logistic function. By taking advantage of ξ- transformation, they obtained a normal style low bound for the observed-data likelihood. Within the mean-field family, the variational density is achieved based on the approximated likelihood. The main difference between our proposal and [24] underlies that Jaakkola and Jordan's formulation is grounded on the approximated likelihood while we utilize the data-augmentation technique.

The further research includes the robust model fitting for the TPM with latent variable analysis, and/or with missing data. These extensions surely raise theoretical and computational challenges and therefore require further study.

ACKNOWLEDGMENTS

The authors thank the editor and two anonymous reviewers for their insightful suggestions and comments. This work was supported by the National Nature Science Foundation of China (NNSF 11471161) and the General Research Projects of Philosophy and Social Science in College and University of China (2022JYB0140). The authors also thank Professor Nian-Sheng Tang, School of Mathematics and Statistics, Yunnan University for his valuable suggestions of variational method.

REFERENCES

[1] Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*. In "Second International Symposium on Information Theory" (B. N. Petrov and B. F. Csaki, Eds.), Academiai Kiado, Budapest, 267–281.

[2] Antoniak, C.E. (1974). Mixtures of Dirichlet processess with Applications to Nonparametric Problems, *The Annals of Statistics*, **2**, 4, 1152–1174.

[3] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*, Oxford Univerisity Press, New York..

[4] Blei, D. M. and Jordan M. I. (1974). Variational Inference for Dirichelt Process Mixtures, *Journal of Bayesian Analysis*, **1**, 4, 121–144.

[5] Blei, D. M.; Kucukelbir A. and McAuliffe, J. D. (2004). Variational Inference: A Review for Statisticians, *Journal of the American Statistical Association*, **112**, 518, 859–877.

[6] Chai, H. and Bailey, K. (2008). Use of Log-skew-normal Distribution in Analysis of Continuous Data with a Discrete Component at Zero. *Stat. Med.*, **27**, 3643–3655.

[7] Cragg, J. G. (1971). Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica*,1971, **39**, 5, 829–844.

[8] Duan, N.; Manning, W. G.; Morris, C. N. and NewHouse, J. P. (1983). A Comparison of Alternative Models for the Demand for Medical Care. *Journal of Business and Economic Statistics*, **1**, 115–126.

[9] Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, **68**, 268–277.

[10] Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

[11] Feng, X. N.; Lu, B.; Song, X. Y. and Ma, S. (2019). Financial literacy and Household Finances: A Bayesian Two-Part Latent Variable Modeling Approach. *Journal of Empirical finance*, **51**, 119–137.

[12] Feng, X. N.; Wu, H. T. and Song, X. Y. (2017). Bayesian Adaptive Lasso for Ordinal Regression With Latent Variables. *Sociological Methods & Research*, **46**, 4, 926–953.

[13] Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, **1**, 2, 209–230.

[14] Franzolini, B.; Lijoi, A. and Prünster, I. (2023). Model selection for maternal hypertensive disorders with symmetric hierarchical Dirichlet processes. *The Annals of Applied Statistics*, **85**, 17, 313–332.

[15] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 410, 398–409.

[16] Gelman, A. (1996). *Inference and Monitoring Convergence*. In "Markov Chain Monte Carlo in Practice" (W.R. Gilks, S. Richardson and D.J. Spiegelhalter, Eds.), Chapman and Hall, London, 131–144.

[17] Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

[18] Geyer, C. J. Practical Markov Chain Monte Carlo. *Statistical Science* ,1992, **7**, 473–511.

[19] Gou, J. W.; Xia, Y. M. and Jiang, D. P. (2022). Bayesian Analysis of Two-Part Nonlinear Latent Variable Model: Semiparametric Method. *Statistical Modeling*, Published online. doi:0.1177/1471082X211059233.

[20] Han, D. X.; Liu, L.; Su, X. G., Johnson, B. and Sun, L. Q. (2019). Variable Selection for Random Effects Two-part Models. *Statistical Methods in Medical Research*, **28**, 9, 2697–2709.

[21] Hanson, T.; Johnson, W. (2002). Modelling Regression Error with a Mixture of Polya Trees. *Journal of the American Statistical Association*, **97**, 1020–1033.

[22] Ishwaran, H, and Zarepour, M. (2000). Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models. *Biometrika*, **87**, 371-390.

[23] Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-breaking Priors. *Journal of the American Statistical Association*, 2001, **96**, 161–173, .

[24] Jaakkola, T. and Jordan, M. I. (2000). Bayesian Parameter Estimation via Variational Methods. *Statistics and Computing*, **10** 25–37.

[25] Koenker, R. W and Bassett, G. W. (1982). Tests of Linear Hypotheses and L_1 Estimation. *Econometrika*, 1982, 50:1577–1583.

[26] Kim, Y. and Muthén, B. O. (2009). Two-Part Factor Mixture Modeling: Application to an Aggressive Behavior Measurement Instrument. *Structural Equation Modeling: A Multidisciplinary Journal*, **16**, 4,602–624.

[27] Liu, L. (2009). Joint Modeling Longitudinal Semi-continuous Data and Survival with Application to Longitudinal Medical Cost Data. *Statist. Med.*, **28**, 6, 972–986.

[28] Liu, L.; Strawderman, R., Cowen M, et al. (2010). A flexible Two Part Random Effects Model for Correlated Medical Costs. *J. Health Econ.*, **29**, 110–123.

[29] Lo, A. Y. (1984). On a Class of Bayesian Nonparametric Estimates: I, Density Estimates. *The Annals of Statistics*, **12**, 351–357.

[30] MacEachern, S. N. and Müller, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, **7**, 2, 223–238.

[31] Manning, W. G.; Morris, C. N.; Newhouse, J. P.; Orr, L. L., et al. (1981). *A Two-Part Model of the Demand for Medical Care: Preliminary Results From the Health Insurance Experiment.* In "Health, Economics, and Health Economics" (J. van der Gaag and M.Perlman, Eds.), Amsterdam, North-Holland, 103–104.

[32] Manning, W. G.; Basu, A. and Mullahy, J. (2005). Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data. *J. Health Econ.*, **24**, 465–488.

[33] Maruotti, A. (2011.) A Two-part Mixed-effects Pattern-Mixture Model to Handle Zero-inflation and Incompleteness in a Longitudinal Setting. *Biometrical Journal*, **53**, 5, 716–734.

[34] Merlo, L., Maruotti, A. and Petrella, L. (2022). Two-part Quantile Regression Models for Semi-continuous Longitudinal Data: A finite mixture approach. *Statistical Modelling*. Published online. doi:10.1177/1471082X21993603.

[35] Mukhopadhyay, S. andD Gelfand, A. E. (1997.) Dirichlet Process Mixed Generalized Linear Models. *Journal of the American Statistical Association*, **92** 438, 633–639.

[36] Marron, J. S. and Wand, M. P. (1992). Exact Mean Integrated Squared Error *The Annals of Statistics*, **20**, 2, 712–736.

[37] Neelon, B.; Zhu, L. and Neelon, S. E. B. (2015). Bayesian Two-part Spatial Models for Semicontinuous Data with Application to Emergency Department Expenditures. *Biostatistics*, **16**, 3,465–480.

[38] Nelder, J. A. and Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, 1972, **135**, 370–384.

[39] Olsen, M. K. and Schafer, J. L. (2001). A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association*, **454** 96, 730–745.

[40] Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, **103**, 482, 681–686.

[41] Pati, D. and Dunson, D. B. (2014). Bayesian nonparametric regression with varying residual density. *Annals of the Institute of Statistical Mathematics*, **66**, 1–31.

[42] Peterson, C. and Anderson, J. (1987). A Mean Field Theory Learning Algorithm for Neural Networks. *Complex Systems*, **1**, 995–1019.

[43] Polson, N. G.; Scott, J. G. and Windle, J. (2013). Bayesian Inference for Logistic Models Using Polya-Gamma Latent Variables. *Journal of American statistician Association*,

108, 504, 1339–1349.

[44] Rooij, M. V.; Lusardi, A. and Alessie, R. (2011). Financial Literacy and Stock Market Participation. *Journal of Financial Economics*, **101**, 2, 449–472.

[45] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461–464.

[46] Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modelling: Multilevel, Longitudinal and Structural Equation Models*, Chapman & Hall/CRC, London.

[47] Sethuraman, J. A. (1994). Constructive Definition of Dirichlet Priors. *Statistical Sinica*, **4**, 639–650.

[48] Tanner, M. A. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation(with discussion). *Journal of the American statistical Association*, **82**, 528–550.

[49] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Ser. B*, **58**, 267–288.

[50] Wainwright, M. J. and Jordan, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, **1**, 1–305.

[51] Wang. Y. and Blei, D. M. (2019). Frequentist Consistency of Variational Bayes. *Journal of the American Statistical Association*, **114**, 527, 1147–1161.

[52] Xia, Y. M.; Lin, Y. B. and Xiong, S. C. (2018). Bayesian Inference of Two-part Factor Analysis Model, *Mathematica Applicata*, **31**, **4**, 761–778.

[53] Xia, Y. M.; Lu, B. and Tang, N. S. (2019). Inference on Two-Part Latent Variable Analysis Model with Multivariate Longitudinal Data, *Structural Equation Modeling: A Multidisciplinary Journal*, **26**, 5, 685–709.

[54] Xing, D. Y.; Huang, Y. X., Chen, H. N., et al. (2017). Bayesian Inference for Two-part Mixed Effects Model Using Skew Distributions, with Application to Longitudinal Semicontinuous Alcohol Data, *Statistical Methods in Medical Research*, **26**, 4, 1838–1853.

[55] Xu, S.; Blozis, S. A. and Vandewater, E. A. (2014). On Fitting a Multivariate Two-Part Latent Growth Model. *Structural Equation Modeling: A Multidisciplinary Journal*, **21**, 1, 131–148.

[56] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties., *Journal of the American Statistical Association*, **101**, 1, 1418–1429.