Published in final edited form as:

Biometrics. 2019 June ; 75(2): 528–538. doi:10.1111/biom.12989.

Semiparametric Bayesian latent variable regression for skewed multivariate data

Apurva Bhingare^{1,*}, Debajyoti Sinha², Debdeep Pati³, Dipankar Bandyopadhyay⁴, Stuart R. Lipsitz⁵

¹Eunice Kennedy Shriver National Institute of Child Health & Human Development, Bethesda, MD

²Department of Statistics, Florida State University, Tallahassee, FL

³Department of Statistics, Texas A & M University, College Station, TX

⁴Department of Biostatistics, Virginia Commonwealth University, Richmond, VA

⁵Department of Medicine, Brigham and Women's Hospital, Boston, MA

Summary:

For many real-life studies with skewed multivariate responses, the level of skewness and association structure assumptions are essential for evaluating the covariate effects on the response and its predictive distribution. We present a novel semiparametric multivariate model and associated Bayesian analysis for multivariate skewed responses. Similar to multivariate Gaussian densities, this multivariate model is closed under marginalization, allows a wide class of multivariate associations, and has meaningful physical interpretations of skewness levels and covariate effects on the marginal density. Other desirable properties of our model include the Markov Chain Monte Carlo computation through available statistical software, and the assurance of consistent Bayesian estimates of the parameters and the nonparametric error density under a set of plausible prior assumptions. We illustrate the practical advantages of our methods over existing alternatives via simulation studies and the analysis of a clinical study on periodontal disease.

Keywords

Dirichlet process; Kernel density; Markov Chain Monte Carlo; Periodontal disease; Skewed error

1 Introduction

In many biomedical and health-care cost studies, we often come across highly skewed multivariate responses. For example, the preliminary analysis (see Figure 2) of clinical data from a periodontal disease (PD) study (Fernandes et al., 2009) of Gullah-speaking African-Americans diabetics (henceforth, GAAD study) clearly reveals that the two major (correlated) clinical endpoints of subject-level PD status (Greenstein, 1997)– the mean

* bhingareapurva@gmail.com.

⁸Supplementary Materials

Web appendices and computer codes referenced in Sections 6 are available with this article at the *Biometrics* website on Wiley Online library and at the GitHub link: <https://github.com/bandyopd/Multivariate-Skewed-models>

(average) periodontal pocket depth (PPD) and the mean clinical attachment level (CAL) of all available teeth of each subject (cluster) are highly non-Gaussian (skewed), *with the skewness of CAL being different from the skewness of PPD. While the CAL is the accepted 'gold standard' endpoint for measuring the progression and history of PD, the PPD measures the current disease activity. Also, CAL can occur in presence, or absence of PPD, and vice versa. Hence, a holistic approach towards assessing the impact of a covariate (say, diabetes status) on PD should consider the effect on the joint distribution of CAL and PPD.* A typical multivariate (normal) regression analysis that ignores non-Gaussianity (say, skewness of CAL) may lead to biased parameter estimates (Azzalini and Capitanio, 2003), and the corresponding predictive density. Although analysis may proceed transforming the multivariate skewed responses to multivariate normal, difficulties remain in choosing a suitable transformation, and interpreting the covariate effects on the original non-transformed responses. As alternatives, the quantile regression (QR) methods (Koenker, 2005) and their multivariate extensions mostly focus on a single pre-specified quantile, without evaluating the levels of skewness, and the whole joint density of the PPD and CAL given covariates.

This paper addresses a number of major challenges and limitations of existing model-based regression analysis of multivariate skewed response data in biostatistical applications. First, many recently developed model-based approaches (Genton, 2004; Sahu et al., 2003; Bandyopadhyay et al., 2012) centers around the restricted parametric family pioneered by Azzalini and Dalla Valle (1996). Our less-restrictive semiparametric proposition with skewed nonparametric marginal error densities will remain attractive under concerns about the validity and consequence of the parametric assumptions on the error density. Second, unlike most existing multivariate skewed models, our elegant model class is closed under marginalization. Specifically, the distribution of any subset of the response vector is within the same multivariate distribution class of the original vector, however, the smaller vector's density involves only the relevant subset of the parameters. So, unlike existing models, where, say, the marginal variance and skewness of the CAL depend on the parameters of PPD, the marginal variance and skewness of CAL in our model do not depend on the parameters of PPD. This property alleviates the major roadblock for component-wise interpretation of skewness and covariate effects, and also paves way for specifying priors based solely on available marginal opinion for each response. Third, the accommodation of a flexible class of multivariate associations (such as antedependence, toeplitz, or autoregressive structures) via a covariance matrix continues to remain a major challenge. Only a recent proposal (Chang and Zimmerman, 2016) accommodates the antedependence association, however, at the cost of a strict condition on the skewness parameters. Our Bayesian semiparametric formulation allows a broader class of multivariate associations (including the aforementioned structures) within the response vector, without imposing any restriction on skewness parameters. Fourth, our proposal is computationally scalable, with easy implementation in freeware like **R** and **JAGS** using available Markov Chain Monte Carlo (MCMC) tools. Finally, we also explore and present some desirable asymptotic justification (Bayesian posterior consistency) of our method under practically reasonable prior assumptions.

The rest of the paper proceeds as follows. After introducing our semiparametric model in Section 2, we present the relevant components of Bayesian inference, such as prior choices, likelihood specifications, posteriors, and the MCMC implementation routines in Section 3. In Sections 4 and 5, we assess the posterior consistency and the finite sample properties of the Bayesian estimates (via simulation studies), respectively. In Section 6, we illustrate our methods via analyzing the motivating GAAD dataset. Finally, some concluding remarks are presented in Section 7.

2. Semiparametric Model for Multivariate Skewed Response

Let $Y_i = (Y_{i1}, \dots, Y_{im}) \in \mathbb{R}^m$ be a row vector of skewed responses from subject (cluster) $i = 1, \dots, n$. For brevity, we use fixed number of m components per subject, which can be generalized to accommodate unequal m_i . Our regression model is

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad \text{for } j = 1, \dots, m, \quad (1)$$

where $\mu_{ij} = x_{ij}\beta_j$, $x_{ij} = (1, x_{ij1}, \dots, x_{ij,p-1})$ is the row vector of $(p-1)$ covariates for Y_{ij} (with first term for the intercept), β_j is the corresponding $(p \times 1)$ vector of regression parameters, and $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im})$ is the row vector of errors following a m -variate multivariate density $f_m(\varepsilon)$, with possibly skewed marginal densities. For the PD study, $m = 2$ for every subject. We introduce a new class of semiparametric multivariate skew (SMS) density for $f_m(\varepsilon)$, denoted by $\varepsilon_i \sim \text{SMS}(0, h, R_\rho, \alpha)$, where R_ρ is a $(m \times m)$ correlation matrix, $h = (h_1, \dots, h_m)$ is a vector of m unknown nonparametric (baseline) symmetric univariate densities, and $\alpha = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m$ is a vector of skewness parameters. This $\text{SMS}(0, h, R_\rho, \alpha)$ density has the following stochastic representation

$$\varepsilon_i^T = A_\alpha |Z_{1i}| + A_\alpha^* Z_{2i}, \quad (2)$$

where $Z_{1i} = (Z_{1i1}, \dots, Z_{1im})^T$ and $Z_{2i} = (Z_{2i1}, \dots, Z_{2im})^T$ are two independent $(m \times 1)$ vectors of latent variables with $h(\cdot)$ being the common marginal density for Z_{1ij} and Z_{2ij} , A_α and A_α^* are diagonal matrices with the j th diagonal elements $a_j = \alpha_j(1 + \alpha_j^2)^{-1/2}$ and $a_j^* = (1 + \alpha_j^2)^{-1/2}$ respectively. The univariate error ε_{ij} in (2) can also be expressed as $\varepsilon_{ij} = a_j Z_{2ij} + a_j^* |Z_{1ij}|$, using the “skewing shock” $|Z_{1ij}|$ to the symmetric random noise Z_{2ij} with weights a_j^* and a_j , such that $a_j^2 + a_j^{*2} = 1$. Only when the skewness parameters $(\alpha_1, \dots, \alpha_m)$ are all equal to zero, the ε_i of (2) equals Z_{2i} with symmetric marginal densities $h(\cdot)$. We assume m “skewing shocks” Z_{1ij} , $j = 1, \dots, m$ to be independent with marginal densities $h_1(\cdot), \dots, h_m(\cdot)$, and we model the association within ε_i via the m -variate Gaussian Copula (Nelsen, 2007)

$$\begin{aligned} \mathcal{C}_m(Z_{2i} | h, R_\rho) &= \phi_m[\Phi_1^{-1}\{H_1(z_{2i1})\}, \dots, \Phi_1^{-1}\{H_m(z_{2im})\} | R_\rho] \\ &\times \prod_{j=1}^m \frac{h_j(z_{2ij})}{\phi_1[\Phi_1^{-1}\{H_j(z_{2ij})\}]} \end{aligned} \quad (3)$$

for Z_{2i} , where $\phi_m(\cdot | M)$ is the m -variate Gaussian density with mean 0 and variance-covariance matrix M , R_ρ is a correlation matrix with unknown parameter vector ρ , $\Phi_1^{-1}(\cdot)$ is

the inverse cdf of $N(0, 1)$, and $H_j(\cdot)$ is the cdf of the unknown symmetric density $h_j(\cdot)$. The copula structure of (3) ensures that Z_{2ij} has the nonparametric marginal density $h_j(\cdot)$ (same as Z_{1ij}) for $j = 1, \dots, m$. We model $h_j(\cdot)$ as a scale mixture of symmetric (around 0) density kernel $\mathcal{K}(\cdot | \sigma)$, given by

$$h_j(u) = \int_0^\infty \mathcal{K}(u | \sigma) dG_j(\sigma), \quad (4)$$

where $G_j(\cdot)$ is the unknown nonparametric mixing distribution of scale σ with support in $\mathbb{R}_+ = (0, \infty)$. In this paper, we use the popular Gaussian kernel $N(0, \sigma^2)$ as the choice for kernel $\mathcal{K}(\cdot | \sigma)$ in (4). However, other choices are possible. For example, using the result of Feller (1971) (p. 158), when the kernel $\mathcal{K}(\cdot | \sigma)$ is $\text{Uniform}(-\sigma, \sigma)$, the nonparametric $h_j(u)$ in (4) represents the class of all unimodal and symmetric around 0 densities. The semiparametric Bayesian inference proceeds via specifying a Dirichlet process prior (Ferguson, 1974; Müller et al., 2015) $DP(G_{0j}, C)$ on $G_j(\cdot)$, with pre-specified prior mean $G_{0j}(\cdot)$ and precision C . In Section 3, we explain that this $DP(G_{0j}, C)$ prior process on $G_j(\cdot)$ induces a flexible prior on the nonparametric density $h_j(\cdot)$ in (4).

When Z_{1i} and Z_{2i} in (2) are assumed to have parametric multivariate elliptical densities, the resulting model is similar to the “multiple skewed shocks” model of Sahu et al. (2003). However, our semiparametric model with nonparametric marginals has two additional crucial differences: (a) a new parametrization of weights (A_α, A_α^*), and (b) a multivariate copula modeling of Z_{2i} as in (3). Also, our model is applicable even in studies where no actual latent skewing shock vectors are clinically interpretable in practice. Later, we demonstrate that the latent vector representation in (2) provides computational convenience.

Now, a parametric subclass of the semiparametric $\text{SMS}(0, h, R_\rho, \alpha)$ density of (2)-(3) is obtained when Z_{1i} and Z_{2i} in (2) have parametric mean-zero multivariate normal densities $N_m(0, D_\sigma^2)$ and $N_m(0, \Sigma)$ respectively with variance-covariance matrices $\Sigma = D_\sigma R_\rho D_\sigma$, where $D_\sigma = \text{diag}(\sigma_1, \dots, \sigma_m)$, and R_ρ is a positive definite correlation matrix. We call this a (multivariate) parametric Gaussian mixture (PGM) model, denoted by $\epsilon_i \sim \text{PGM}(0, \alpha, \Sigma)$. Despite similar parametrization based on (α, Σ) , our parametric PGM model has key differences with the location 0 parametric multivariate skew-normal (MSN) model of Azzalini and Dalla Valle (1996), denoted by $\epsilon_i \sim \text{MSN}(0, \alpha, \Sigma)$, with the corresponding joint density $f_m(\epsilon_{i1}, \dots, \epsilon_{im}) = 2\phi_m(\epsilon_i^T \Sigma \epsilon_i) \Phi_1(\epsilon_i^T \alpha)$. The marginal distribution of each component ϵ_{ij} for $\text{PGM}(0, \alpha, \Sigma)$ density is the univariate skew-normal (SN) density $f_j(\epsilon_{ij}) = 2\phi_1(\epsilon_{ij} | \sigma_j) \Phi_1(\alpha_j \epsilon_{ij})$ denoted by $\text{SN}(0, \alpha_j, \sigma_j)$ and dependent only on the component-specific parameters (α_j, σ_j) , whereas, for the $\text{MSN}(0, \alpha, \Sigma)$ model of Azzalini and Dalla Valle (1996), the corresponding univariate marginal density is $\text{SN}(0, \bar{\alpha}_j, \sigma_j)$, however, with parameter $\bar{\alpha}_j$ being a function of the entire (α, Σ) (Chang and Zimmerman, 2016). See Appendix A for details on the expression of $\bar{\alpha}_j$. In general, similar to the parametric formulation of Sahu et al. (2003), the model class in (2) is closed under marginalization, because the marginal density of any subset $\epsilon_i^{(1)}$ of $\epsilon_i = (\epsilon_i^{(1)}, \epsilon_i^{(2)})$ is from the same class of semiparametric density $\text{SMS}(0, h^{(1)}, R_{11}, \alpha^{(1)})$ indexed by $(h^{(1)}, R_{11}, \alpha^{(1)})$, where $\alpha = (\alpha^{(1)},$

$\alpha^{(2)}, (h_1, \dots, h_m) = (h^{(1)}, h^{(2)})$ and matrix $R = \begin{pmatrix} R_{11} & R_{12} \\ R_{12} & R_{22} \end{pmatrix}$ are partitions of (α, h, R) . Unlike the multivariate model of (2) based on multiple “skewing shocks” $|Z_{1i1}|, \dots, |Z_{1im}|$, the existing multivariate models, such as $\text{MSN}(0, \alpha, \Sigma)$ are based on univariate “skewing shock” $|Z_{1i}^*|$ (common to all m components). As a consequence, these models are closed under marginalization only in a restricted sense because $\varepsilon_i \sim \text{MSN}(0, \alpha, \Sigma) \Rightarrow \varepsilon_i^{(1)} \sim \text{MSN}(0, \alpha_{(1)}^*, \Sigma_{(1)}^*)$, where the parameters $(\alpha_{(1)}^*, \Sigma_{(1)}^*)$ are different from $(\alpha^{(1)}, \Sigma_{11})$, and in fact, they are even functions of $(\alpha^{(2)}, \Sigma_{12}, \Sigma_{22})$ (see Theorem 1 of Chang and Zimmerman (2016) for details). For the bivariate PD response in GAAD study, the existing MSN models are essentially based on a subject-specific univariate skewing shock $|Z_{1i}^*|$ common to both CAL and PPD, compared to the paired skewing shocks $(|Z_{1i1}|, |Z_{2i1}|)$ in model (2). Consequently, $\bar{\alpha}_1$, depends on α_1 , as well as $(\alpha_2, \sigma_2, \rho)$. This complicates the marginal interpretation and Bayesian prior specification of α_1 for the MSN model.

To further demonstrate the lack of interpretation of α as a skewness parameter within the MSN model compared to the straightforward interpretation within our model, Figure 1 presents the contour plots and marginal densities of the bivariate MSN (Chang and Zimmerman, 2016) model (right panel) and our bivariate PGM model (left panel), when both models have common skewness parameter $\alpha_1 = \alpha_2 = \alpha_0 = 10$, $\rho = -0.7$ and $\sigma_1 = \sigma_2 = 1$. While Figure 1(a) demonstrate substantial positive skewness of both components from the bivariate PGM model (as expected from the chosen α_0 as high as 10), both the (near elliptical) contour and marginal density plots in Figure 1(b) indicate approximately symmetric marginal densities from the MSN model. This illustrates the fact that, depending on the value of ρ and other parameters of the model, the value (even its sign) of the marginal Pearsonian skewness of the MSN model may be very different from the value of α_0 . However, for our multivariate model in (2), the Pearsonian skewness of each marginal density does not depend on ρ and scale parameters of other components. Further comparisons of contour plots of the bivariate PGM with the bivariate SN density for various choices of $\alpha_0 = 0, 2, 5, 10$ and $\sigma_1 = \sigma_2 = 1$ with $\rho = 0.7$ are presented in Web Appendix B.

The marginal mean of Y_i for our SMS model is $E(Y_{ij} | X_i) = \beta_{j0}^* + \sum_{l=1}^{p-1} x_{ijl} \beta_{jl}$ with $\beta_{j0}^* = \beta_{j0} + \left\{ \alpha_j / (1 + \alpha_j^2) \right\}^{1/2} \lambda_j$. The covariance matrix of Y_i is $\text{var}(Y_i | X_i) = V$, where the diagonal elements $V_{jj} = \sigma_{hj}^2 - \left\{ \alpha_j^2 / (1 + \alpha_j^2) \right\} \lambda_j^2$, and the off-diagonal elements $V_{jk} = \sigma_{hjk} a_j^* a_k^*$, with weight $a_j^* = 1 / (1 + \alpha_j^2)^{1/2}$ for all $j, k \in \{1, \dots, m\}$. Here, the common expectations $E(Z_{1ij}) = \lambda_j$ and variances $\text{Var}(Z_{1ij}) = \text{Var}(Z_{2ij}) = \sigma_{hj}^2$ are taken with respect to the nonparametric marginal density $h_j(\cdot)$ of Z_{1ij} and Z_{2ij} , given in (4). The covariance $\text{Cov}(Z_{2ij}, Z_{2ik}) = E[Z_{2ij} Z_{2ik}] = \sigma_{hjk}$ involves the expectation with respect to the bivariate density of the pair (Z_{2j}, Z_{2k}) based on the copula model of (3). For the GAAD study, the correlation $\text{Corr}(e_{i1}, e_{i2})$ between PPD and CAL responses within subject/cluster i is $\sigma_{h12} a_1^* a_2^* / \left[\left\{ \sigma_{h1}^2 - a_1^2 \lambda_1^2 \right\} \left\{ \sigma_{h2}^2 - a_2^2 \lambda_2^2 \right\} \right]^{1/2}$, where $\sigma_{h12} = E[Z_1 Z_2]$ is the expectation taken with respect to the bivariate version of (3) for $m = 2$.

3 Bayesian Inference: Likelihood, Prior and Posterior

In this section, we develop the semiparametric Bayesian inferential framework for the SMS model. Given observed data $\mathcal{D} = \{y_i, X_i; i = 1, \dots, n\}$ from n independent multivariate responses, the likelihood function of the parameter space $\Theta = (\beta, \alpha, \rho, \sigma, G)$ is

$$\mathcal{L}(\Theta | \mathcal{D}) = \prod_{i=1}^n f_m(y_i - X_i\beta), \quad (5)$$

where $f_m(\epsilon_i)$ is the integral $\int \mathcal{C}_m(A_\alpha^{*-1}\{\epsilon_i - A_\alpha Z_{1i}\} | h, R_\rho) \left[\prod_{j=1}^m h_j(Z_{1ij}) \right] dZ_{1i}$, taken over the support $\mathbb{R}^m$ of latent vectors Z_{1i} in (2), $h_j(\cdot)$ is the independent marginal density of Z_{1ij} , $\epsilon_i = y_i - X_i\beta$, and $\mathcal{C}_m(\cdot | h, R_\rho)$ is the copula specification in (3). Clearly, $f_m(\cdot)$ of $\epsilon_i = y_i - X_i\beta$ in (2) has a complicated analytical expression. Based on the likelihood in (5), the joint posterior density is

$$p(\Theta | \mathcal{D}) \propto \mathcal{L}(\Theta | \mathcal{D}) \pi_1(\beta) \pi_2(\alpha) \pi_3(\rho) \pi_4(G), \quad (6)$$

with the following priors.

- i. The priors $\pi_1(\beta)$ and $\pi_2(\alpha)$ for β and α are independent mean zero multivariate normal with pre-specified covariance matrices M_β and M_α , respectively. In particular, we choose these as $\sigma^2 I$, where $\sigma^2 = 100$.
- ii. For the GAAD study with bivariate response, the prior $\pi_3(\rho)$ of the scalar parameter ρ in R_ρ (3) is Uniform $(-1, 1)$. However, for a vector ρ , this $\pi_3(\rho)$ prior has to be a multivariate density with an appropriate support.
- iii. The joint prior process $\pi_4(G)$ of $G = (G_1, \dots, G_m)$ is the product of m independent Dirichlet processes $DP(G_{0j}, C)$, for $j = 1, \dots, m$ (Ferguson, 1974), where the prior mean G_{0j} of G_j is specified based on the “prior guess” $h_{0j}(u) = \int_0^\infty \mathcal{K}(u | \sigma) dG_{0j}(\sigma)$ of the density $h_j(u)$ of Z_{1ij} in (4). The user-specified concentration/precision parameter C is the measure of uncertainty of G_j around G_{0j} ; the larger the value of C , the closer the sample-paths of $h_j(u)$ are to the “prior guess” $h_{0j}(u)$, while smaller values of C allow the sample-path of $h_j(u)$ to be very different from $h_{0j}(u)$.

The posterior in (6) is analytically intractable, hence, we proceed via MCMC sampling from the joint distribution

$$p(\Theta, Z_1 | \mathcal{D}) \propto \left[\prod_{i=1}^n \phi_m(W_i | 0, R_\rho) \prod_{j=1}^m h_j(Z_{1ij}) \right] \pi_1(\beta) \pi_2(\alpha) \pi_3(\rho) \pi_4(G), \quad (7)$$

where $\phi_m(u|0, R)$ with $(m \times m)$ covariance matrix R is a mean-zero multivariate normal density evaluated at $u \in \mathbb{R}^m$, $W_i = (W_{i1}, \dots, W_{im})$ with

$W_{ij} = \Phi_1^{-1} \left[H_j \left\{ (1 + \alpha_j^2)^{1/2} (y_{ij} - X_{ij}\beta) - \alpha_j Z_{1ij} \right\} \right]$, Φ_1^{-1} is the inverse-cdf of the standard normal density, and $H_j(\cdot) = H_j(\cdot | G_j)$ is the cdf corresponding to the density $h_j(\cdot | G_j)$. For the parametric case, $h_j(\cdot | G_j)$ and $\pi_4(G)$ are replaced with the parametric density $h_j(\cdot | \sigma_j)$ and the

joint parametric prior of $\pi_4(\sigma)$. For example, when $h_j(\cdot|\sigma_j)$ is $N(0, \sigma_j^2)$ density with unknown σ_j , we use the parametric prior distributions $\sigma_j \sim g_j(\cdot|\gamma_j)$ independently for $j = 1, \dots, m$ with pre-specified hyperparameters γ_j . For this special case, W_{ij} simplifies to $\left\{ (1 + \alpha_j^2)^{1/2} (y_{ij} - X_{ij}\beta) - \alpha_j |Z_{1ij}| \right\} / \sigma_j$. For a semiparametric analysis with a particular symmetric density kernel $\mathcal{K}(z|\sigma)$ for the nonparametric kernel-mixture densities $h = (h_1, \dots, h_m)$ in (4), we implement the joint density in (7) by assuming (a) $(W_{i1}, \dots, W_{im})$ for $i = 1, \dots, n$ are mean-zero multivariate $N_m(0, R_\rho)$, where $W_{ij} = \Phi_1^{-1} \left[K_{ij} \left\{ (1 + \alpha_j^2)^{1/2} (y_{ij} - X_{ij}\beta) - \alpha_j |Z_{1ij}| \right\} \right]$, and K_{ij} and K_{ij}^{-1} are the cdf and inverse-cdf of $\mathcal{K}(\cdot | \sigma_{ij})$; (b) Z_{1ij} for $j = 1, \dots, m$ are independent with density $\mathcal{K}(\cdot | \sigma_{ij})$, and (c) σ_{ij} for $j = 1, \dots, m$ are independent with nonparametric cdf G_j following Dirichlet Process (DP) prior $DP(G_{0j}, C)$. For the Gaussian kernel in (4), $K_{ij}(u) = \Phi(u/\sigma_{ij})$ and W_{ij} further simplifies to $W_{ij} = \left\{ (1 + \alpha_j^2)^{1/2} (y_{ij} - X_{ij}\beta) - \alpha_j |Z_{1ij}| \right\} / \sigma_{ij}$.

For convenient MCMC implementation, we use the constructive definition of the DP (Sethuraman, 1994), truncated at a user specified finite number (K_0) of components for the prior on G_j . Suppose $(\delta_1, \dots, \delta_{K_0})$ are generated independently from G_{0j} , $(B_1, \dots, B_{K_0-1})$ are generated independently from $\text{Beta}(1, C)$, $\omega_1 = B_1$, $\omega_h = B_h \prod_{l < h} (1 - B_l)$ for $h = 2, \dots, K_0 - 1$ and $\omega_{K_0} = 1 - \sum_{l=1}^{K_0-1} \omega_l$. Then, the sample path of G_j is approximated by $G_j = \sum_{l=1}^{K_0} \omega_l \delta_l$, when K_0 is large. These likelihood steps and priors facilitate easy Bayesian semiparametric implementation for either the Gaussian or uniform kernels using available freeware, such as WinBUGS/JAGS.

4 Posterior Consistency

For theoretically valid inference from a complex semiparametric Bayesian model, it is important to assure that the parameter posteriors become increasingly concentrated around the true parameter values with increasing sample size. This necessitates the investigation of the asymptotic properties, along with the finite sample properties. In this section, we provide sufficient conditions for posterior consistency, the most important asymptotic property. We present our theory for the case where the dimension of the regression parameter, β , is p .

We first investigate if the support of the prior is large enough to cover all possible relevant densities. Suppose $\mathcal{F}$ denotes the class of all univariate asymmetric unimodal residual densities and $\mathcal{C}\{(\mathcal{F})^m \times (0, 1)\}$ denotes the collection of all m -dimensional residual densities f_0 with unknown correlation $\rho_{jk} \in (0, 1)$ having the form,

$$f_0(e) = \int_{\mathbb{R}_+^m} 2^m \left| J \right| \prod_{j=1}^m \frac{h_{0j}(z_j) h_{0j}(e_j^*)}{\phi_1 \left[\Phi^{-1} \{ H_{0j}(e_j^*) \} \right]} \phi_m \left[\Phi^{-1} \{ H_{01}(e_1^*) \}, \dots, \Phi^{-1} \{ H_{0m}(e_m^*) \} \mid R_\rho \right] dz,$$

with the j -th marginal density given by $f_{0j}(e_j) = \int_{\mathbb{R}_+} 2 \sqrt{1 + \alpha_j^2} h_{0j}(z_j) h_{0j}(e_j^*) dz_j \quad j = 1, \dots, m$, where $h_{0j}(\cdot)$ is the true (unknown) symmetric around 0 density, $H_{0j}(\cdot)$ is the corresponding

cumulative distribution function, $e_j^* = \sqrt{1 + \alpha_j^2} e_j - \alpha_j z_j$, and the Jacobian of transformation $|J|$ is $|J| = \begin{vmatrix} I & 0 \\ A_\alpha^* - 1 & -A_\alpha^* - 1 \end{vmatrix} A_\alpha$. Let Π denote a prior on $\mathcal{C}\{(\mathcal{F})^m \times (0, 1)\}$, such that $\Pi = (\pi_{\mathcal{F}})^m \otimes \pi_3(\rho)$, where $\pi_{\mathcal{F}}$ is a prior on $\mathcal{F}$ defined by $G_j \sim \text{DP}(G_{0j}, C)$ for $j = 1, \dots, m$, $\alpha \sim \pi_2(\alpha)$ and $\pi_3(\rho)$ is a prior on $(0, 1)$. For ease of exposition, we will assume that α is known. To show that posterior consistency holds at the true density f_0 , we first define the Kullback-Leibler divergence as $KL(f_0, f) = \int_{\mathbb{R}^m} f_0 \log(f_0/f)$ and Kullback-Leibler neighborhood of size ϵ as $\kappa_\epsilon(f_0) = \{f: KL(f_0, f) < \epsilon\}$. For any prior Π^* on the density space $\mathcal{F}^*$, f_0 is said to be in the Kullback-Leibler support of Π^* denoted by $KL(\Pi^*)$ if $f_0 \in \mathcal{S}$, where $\mathcal{S} = \{f_0: \Pi^*(f: KL(f_0, f) < \epsilon) \forall \epsilon > 0\}$. Define $\mathcal{F}_{KL} = \{f_0 \in \mathcal{C}\{(\mathcal{F})^m \times (0, 1)\}, f_0 | \log \int f_0 < \infty\}$. We characterize the Kullback-Leibler support of Π in the following lemma, with its proof available in Web Appendix C.

LEMMA 1:

$\mathcal{F}_{KL} \subset KL(\Pi)$ if G_{0j} is defined on support $\mathbb{R}_+$ for all $j = 1, \dots, m$ and G_{0j} is absolutely continuous with respect to Lebesgue measure and $\text{supp}(\pi_3(\rho)) = (0, 1)$.

Suppose $\mathcal{D}_n$ is the observed data with sample size n , and β_0 and f_0 are the true values of the regression parameter vector and the multivariate error density, respectively. We now present the sufficient conditions to ensure that as the sample size n increases, the posterior distributions of the parameter β and the error density f are concentrated around a small neighborhood around their true values. We are essentially interested in the inference on the regression parameter β ; so, we define a strong L_1 neighborhood of radius $\delta > 0$ around the true value β_0 as the set $\mathcal{S}(\beta_0) = \{\beta: \|\beta - \beta_0\| < \delta\}$, and weak neighborhood around f_0 as $\mathcal{W}_\epsilon(f_0) = \{f: |\int \phi f - \int \phi f_0| < \epsilon\}$ for a bounded continuous function ϕ . Consider $U = \mathcal{W}_\epsilon(f_0) \times \mathcal{S}(\beta_0)$ for any arbitrary $\delta > 0$. The following theorem presents the result on posterior consistency, with the proof given in Web Appendix C.

THEOREM 1:

Suppose $(f_0, \beta_0) \in \mathcal{F}_{KL} \times \mathbb{R}^p$. Consider a prior $\tilde{\Pi} = (\Pi \otimes \Pi_\beta)$ on $\mathcal{F}_{KL} \times \mathbb{R}^p$. Then, $\tilde{\Pi}\{(f_m, \beta) \in U \cap \mathcal{D}_n\} \rightarrow 0$ a.s. under the true data generating distribution P_{f_0, β_0} that generates data $\mathcal{D}_n$.

5 Simulation Study

We conduct three simulation studies to compare the finite sample properties of our Bayesian estimates to those obtained from competing models under various data generation schemes. Here, we present results from Simulation 1, where the data is generated from the model in (2). Details on Simulations 2 and 3 are presented in Web Appendix D, where the data are generated under the competing parametric skew- t assumptions.

Simulation 1:

We consider $N = 500$ replications of bivariate responses with sample size $n = 50$. The bivariate skewed errors (e_{1i}, e_{2i}) were generated from the bivariate PGM model, where $\alpha = (\alpha_1, \alpha_2) = (2, 2)$, $Z_{1i} \sim N_2(0, I)$ and $Z_{2i} \sim N_2(0, R_\rho)$, R_ρ is a (2×2) correlation matrix with

$R_{12} = 0.7$. The regression structure is $Y_{ij} = \beta_0 + x_{ij1}\beta_1 + x_{ij2}\beta_2 + \varepsilon_{ij}$, $j = 1, 2$ where $(\beta_0, \beta_1, \beta_2) = (2, 0.5, 0.5)$, with x_{ij1} and x_{ij2} sampled independently, such that $x_{ij2} \sim N(0, 1)$ and $x_{ij1} = 1$, or -1 , each with probability 0.5. This simulation model has the same regression parameters (common β) for both components (different from separate β for two responses in GAAD study). We fit the SMS model, where symmetric $h_1(\cdot)$ and $h_2(\cdot)$ are expressed as unknown scale mixtures of Gaussian kernels. We also fit the PGM model, assuming $Z_{1ij} \sim N(0, \sigma_j^2)$ and $Z_{2i} \sim N_m(0, \Sigma)$, with $(\Sigma)_{11} = \sigma_1^2$, $(\Sigma)_{22} = \sigma_2^2$ and $(\Sigma)_{12} = \rho\sigma_1\sigma_2$. We use independent $N(0, 1)$ priors for β_0, β_1 and β_2 , and also for the skewness parameters α_1 and α_2 . For the $DP(G_{01}, C)$ and $DP(G_{02}, C)$ priors corresponding to G_1 and G_2 respectively, we assume the prior guess G_{01} and G_{02} to be a $\text{Gamma}(2, 1)$ density and $C = 1$, implying the prior guess for ε_{ij} is mean 0 with ranges ± 6 , with high prior probability for a symmetric true density. This also implies the skewing shock has range of $(0, 6)$, with a high probability. Although these prior guesses for G_1 and G_2 corresponds to a prior guess of $f(\varepsilon_i)$ far away from the true density of ε_i , we would like to demonstrate that even this guess produces posterior estimates with good finite sample properties. For the PGM, we specify $\text{Gamma}(2, 1)$ for $\sigma_1^2 = \sigma_2^2$. To test the performances, we use the mean squared error

$\text{MSE}(\hat{\theta}) = \sum_{k=1}^N (\hat{\theta}_k - \theta)^2 / N$, where $\hat{\theta}_k$ is the posterior estimate of θ from the k th simulated dataset, $k = 1, 2, \dots, N$. We also report the relative bias (RB), $\sum_{k=1}^N (\hat{\theta}_k - \theta) / (N\theta)$ and the coverage probability (CP) of the 95% interval estimates from these competing methods. In conjunction, we also compared a generalized estimating equation (GEE) fit (Liang and Zeger, 1986), given that GEE is equivalent to a mixed model with a sandwich variance estimate, although producing a biased marginal estimate of the intercept

$\beta_{0j}^* = \beta_0 - \{\alpha_j / (1 + \alpha_j^2)\}^{1/2} \lambda_j$. The results are summarized in Table 1.

We observe that both the SMS and the PGM models yield smaller MSE's (at least 20% reduction), and overall better CPs, compared to the GEE estimates. Unlike the GEE, our methods provides the estimates of skewness, and more precise interval estimates of the regression parameters. Also, under model misspecification (data simulated from the PGM model), the posterior estimates from the SMS model are comparable to the PGM in terms of MSE and RB. Additional simulation studies (see Simulations 2 and 3 in the Web Appendix D) that avoid stringent assumptions on the true form of the error densities reveal the superiority of the SMS and PGM models over various existing alternatives, such as the skew- t , and GEE.

6 Analysis of GAAD Study

The GAAD study aims to evaluate the PD status of the Gullah-speaking African-Americans as specified by the subject-level (mean) CAL and PPD endpoints, and quantify the effects of subject-level covariates such as age (in years), gender (1=Female, 0=Male), Body Mass Index or BMI (obese=1 if BMI ≥ 30 , = 0, otherwise), glycemic/HbA1c status (1= High/uncontrolled, 0 = controlled) and smoking status (1 = smoker, 0 = never smoker), on these endpoints. For our analysis, we select $n = 288$ patients (subjects) with complete covariate information. About 31% of the subjects are smokers, with a mean age of 55 years, ranging from 26–87 years. Female subjects are predominant in our data (about 76%), which is not

uncommon in Gullah population (Johnson-Spruill et al., 2009). Also, 68% of subjects are obese (BMI ≥ 30), and 59% are with uncontrolled HbA1c level.

Here, the bivariate correlated responses, the mean PPD and mean CAL, calculated as averages of the corresponding measurements across all sites and tooth for that subject, are non-Gaussian (heavily skewed). Hence, the validity of estimation under a standard linear mixed model (LMM) framework remains questionable. For further illustration, Figure 2 [panels (a)-(f)] presents the histograms of the responses, the histogram and Q-Q plot of the empirical Bayes estimates of the random effects, and the histograms of the residuals, after fitting LMMs separately to the responses using the nlme package in R. These plots clearly reveal evidence of different degrees of skewness for the two error terms, and also for the random subject effects. However, to avoid a overly complicated model, we assume that the skewness of, say, CAL is the same for all subjects/clusters. To accommodate this, we illustrate the application of our models developed in Sections 2 and 3 on this dataset.

Specifically, we compare the fit of the following 4 competing models:

1. The proposed SMS model, with unknown symmetric densities $h_1(\cdot)$ and $h_2(\cdot)$ as nonparametric scale mixtures of Gaussian kernels,
2. The PGM model, with latent vectors $Z_{1i} \sim N_2(0, D_\sigma^2)$ and $Z_{2i} \sim N_2(0, \Sigma)$, where $D_\sigma^2 = \text{diag}(\sigma_1^2, \sigma_2^2)$ and $(\Sigma)_{11} = \sigma_1^2$, $(\Sigma)_{22} = \sigma_2^2$, $(\Sigma)_{12} = \rho\sigma_1\sigma_2$.
3. The skew- t model (ST), with a Skew- t (Sahu et al., 2003) density for (e_{1i}, e_{2i}) , with ν degrees of freedom,
4. The Bivariate Normal (BVN) model, with a symmetric $N_2(0, \Sigma)$ density for (e_{1i}, e_{2i}) , ignoring the skewness.

We use practically flat independent $N(0, 100)$ priors for all the regression parameters (the components of β_1 and β_2), and the skewness parameters α_1 and α_2 . For the unknown mixing distributions G_1 and G_2 , we choose independent $DP(G_0; C)$ priors, with $C = 1$, and assuming same Gamma(2, 1) density for the prior guess of G_{01} and G_{02} . These prior guesses matches with the prior choices for σ_1 and σ_2 in the PGM model, the skew- t and bivariate normal models. For the degrees of freedom ν of the skew- t density, we choose $\text{Exp}(0.1) \mathcal{I}_{[2, \infty)}$, the exponential density truncated at 2. For all four models, we use a Uniform(-1, 1) as the prior for ρ . It is important to note that our choice of priors, although practically flat, is primarily for illustration of our statistical methods, and may not represent the prior opinion of clinical investigators. We generated two parallel MCMC chains of size 150,000 and computed the posterior estimates after discarding the first 100,000 iterations (burn-in). To guard against potential autocorrelation among successive iterations, we used a thinning of 25. To assess model convergence, we use the trace plots, autocorrelation plots and the Gelman-Rubin $\hat{R}$. Compared to the SMS model, the PGM required larger number of iterations to converge. *The relevant R and JAGS codes for implementing these models on simulated data are available at a GitHub repository (see Web Appendix E).*

We use the Conditional Predictive Ordinate (CPO) statistic, $(\text{CPO})_i = \int f_m(y_i - X_i\beta | \Theta) p(\Theta | \mathcal{D}_{(-i)}) d\Theta$ (Gelfand et al., 1992), to compare the model performances, where $\mathcal{D}_{(-i)}$ is the cross-

validated data after deleting the i^{th} observation from the full data $\mathcal{D}$, and $p(\Theta | \mathcal{D}_{(-i)})$ is the posterior density of Θ given $\mathcal{D}_{(-i)}$. The corresponding summary statistic using the CPO _{i} is the log of the pseudo-marginal likelihood (LPML), given by $\text{LPML} = \sum_{i=1}^n \log(\text{CPO}_i)$. We also use the Watanabe-Akaike information criterion (waic) (Vehtari et al., 2017), considered a state-of-the-art Bayesian model selection tool. The computations of the waic and LPML are conveniently based on MCMC samples from the full posterior distribution $p(\Theta | \mathcal{D})$. Models with larger LPML and smaller waic indicate more support for observed data. These values suggest the PGM model to be the most appropriate (LPML = -127.882 and waic = 441.89), followed by the SMS model (LPML = -134.03 and waic = 516.88). Recall that the PGM model is the parametric version of the SMS model. The LPML (waic) values for the skew- t and bivariate normal models are -244.546 (618.94) and -300.315 (654.60) respectively, both substantially smaller (larger) than our new semiparametric and parametric Bayesian methods. These numbers also suggest that compared to existing parametric competitors, our proposal is more appropriate for the GAAD dataset.

The posterior estimates, standard deviations, and 95% credible intervals (CIs) of model parameters obtained from fitting the PGM and SMS models are presented in Table 2, along with the corresponding estimates (without the 95% CIs) from the skew- t and the BVN models. The CIs for the skewness parameters a_1 (for PPD) and a_2 (for CAL) do not contain 0, and are positive under both PGM and SMS models, implying substantial evidence of right-skewness for both responses, also revealed in Figure 2. However, the CIs for a_1 for the skew- t contains 0, and fail to detect skewness for PPD. Posterior estimates of the effect of HbA1c on both PPD (not substantial data evidence under SMS) and CAL are positive, with 95% CIs excluding zero, implying higher glycemic status may lead to substantially higher levels of PD. Among smokers, there is a trend (without strong evidence) of higher PPD and CAL, compared to non-smokers, with significant evidence of higher CAL only from the PGM model. Also, strong evidence of higher PPD and CAL are observed in males compared to females, from both SMS and PGM models. However, under the skew- t and BVN models, gender does not have enough posterior evidence. In addition, there is also not enough evidence for the effects of age and BMI on PPD and CAL. On the overall, for most parameters, our new proposals provide more precise posterior estimates, as revealed by tighter 95% CIs. In terms of sensitivity analysis, we observe that moderate changes in the choice of priors do not affect the parameter estimates (both magnitude and sign) and the model comparison measures (LPML and waic) greatly.

To illustrate the practical usefulness of Bayesian inference using the SMS model, we focus on the difference in median responses (PPD and CAL) for two future patients with same age, BMI and smoking status, but different gender and HbA1c values. The 95% CI estimates of difference in median PPD (D_{PPD}) and difference in median CAL (D_{CAL}) between patient₁ (Female, low HbA1c) and patient₂ (Male, high HbA1c) is (-0.5562, -0.0897) and (-0.8076, -0.2570), respectively. The actual PPD differences are wider than the difference in the estimated medians, however the interval is still way above 0. Under the PGM model, the 95% CI estimates of D_{PPD} and D_{CAL} are (-0.6496, -0.0204) and (-1.0301, -0.2071) respectively, both wider than those from the SMS model, but well above 0. However, under the skew- t model, the 95% interval estimates for D_{PPD} and D_{CAL} are (-0.5712, 0.1156) and

(−0.8983, 0.0890), respectively, both covering 0 and indicating lack of posterior evidence in support of the difference in the median PPD and median CAL values between patient₁ and patient₂. Figures 3(a) and Figure 3(b) present the marginal density histogram of residuals for PPD and CAL responses, respectively, obtained after fitting the SMS model, while Figure 3(c) presents the contour plot of the joint bivariate density of these residuals. Evidently, the marginal, as well as the joint bivariate density of the residuals for PPD and CAL is skewed, with one outlying residual corresponding to a 53 year old non-smoker female patient with low BMI but with extremely high PPD and CAL values.

7 Discussion

Our methods development follows the recent joint EU/USA Periodontal Epidemiology Working Group (Holtfreter et al., 2015) recommendations on modeling the full-mouth average PPD and CAL, and is different from prior published work (Bandyopadhyay et al., 2010; Reich and Bandyopadhyay, 2010) in terms of the modeling objectives and clinical endpoints considered. Separate regression modeling of full-mouth average PPD and CAL is not uncommon (Jentsch et al., 2016). However, to cast new light on the biology of PD, this paper breaks new ground via joint modeling of the two responses, instead of analyzing them separately.

In biomedical research, a semiparametric model is often preferred, because they avoid restrictive and hard to verify parametric assumptions, specially on model features that are not of direct interest, such as the $h_j(\cdot)$ function in our model. A major strength is the results on posterior consistency that establishes the theoretical validity of our model. Also, our inferential framework is substantially different from the estimating equations (EE) approach of Ma et al. (2005). EE based models and existing parametric “single shock” models simultaneously adjust for the skewness and the within-cluster association via essentially assuming $|Z_{1i1}|, \dots, |Z_{1im}|$ in (2) be equal to a single univariate cluster-specific skewing shock variable $|Z_i^|$. This assumption of a single skewing shock per cluster is sometimes biologically untenable, and impedes the interpretation of the parameters and the link between skewness and association. In particular, for Bayesian inference, the consequence of this assumption is a major impediment for prior specifications on the parameters α and ρ , based on available prior opinions about the marginal skewness and the within-cluster association. For our multiple-shock model, the marginal Pearsonian coefficient of skewness for each e_{ij} is only a function of (α_j, h_j) . As a consequence of having independent skewing shocks for each component, our model in (2) has comparatively moderate within-cluster association in tails, and is more applicable when the association is not entirely driven by the tail-association. To accommodate strong tail-association, a possible extension of the model in (2) is to assume a multivariate distribution for Z_{1i} , allowing dependence among $Z_{1i1}, \dots, Z_{1im}$. However, such a model is less parsimonious than our current model and lacks advantages of ease of prior specifications and computational scalability via standard software. Furthermore, our methods can also estimate any desired quantile function, and are more appropriate compared to available quantile regression techniques focusing on effects of covariates on certain quantiles.*

Our class of models can incorporate a large and flexible within-subject association including Toeplitz, AR(1) and even higher-dimensional structures for the association matrix R_ρ . Also, for our model, the adopted association structure (form of R_ρ) does not affect the marginal densities. In addition, our latent variable representation facilitates the MCMC-based computation, and allows the joint density class to be closed under marginalization. *The later property helps to extend our methods straightforwardly to studies where the dimensions m_i and association matrix R_i of the response vectors vary over subjects.*

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank the editor, the associate editor and two referees, for their constructive comments. They also acknowledge NIH grants R01-DE024984 and R03-CA205018 and the Gustavus & Louise Pfeiffer Research Foundation grant for their support.

References

- Azzalini A. and Capitanio A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65, 367–389.
- Azzalini A. and Dalla Valle A. (1996). The multivariate skew-normal distribution. *Biometrika* 83, 715–726.
- Bandyopadhyay D, Lachos VH, Abanto-Valle CA, and Ghosh P. (2010). Linear mixed models for skew-normal/independent bivariate responses with an application to periodontal disease. *Statistics in Medicine* 29, 2643–2655. [PubMed: 20740568]
- Bandyopadhyay D, Lachos VH, Castro LM, and Dey DK (2012). Skew-normal/independent linear mixed models for censored responses with applications to Hiv viral loads. *Biometrical Journal* 54, 405–425. [PubMed: 22685005]
- Chang S-C and Zimmerman DL (2016). Skew-normal antedependence models for skewed longitudinal data. *Biometrika* 103, 363–376. [PubMed: 27279663]
- Feller W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. 2 Wiley, New York, NY, second edition.
- Ferguson T. (1974). Prior distribution on spaces of probability measures. *The Annals of Statistics* 2, 615–629.
- Fernandes J, Wiegand R, Salinas C, Grossi S, Sanders J, Lopes-Virella M, and Slate E. (2009). Periodontal disease status in Gullah African Americans with type 2 diabetes living in South Carolina. *Journal of Periodontology* 80, 1062–1068. [PubMed: 19563285]
- Gelfand AE, Dey DK, and Chang H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, DTIC Document.
- Genton MG (2004). *Skew-elliptical distributions and Their Applications: A Journey Beyond Normality*. Chapman and Hall/CRC.
- Greenstein G. (1997). Contemporary interpretation of probing depth assessments: diagnostic and therapeutic implications. A literature review. *Journal of Periodontology* 68, 1194–1205. [PubMed: 9444595]
- Holtfreter B, Albandar JM, Dietrich T, Dye BA, Eaton KA, Eke PI, Papapanou PN, and Kocher T. (2015). Standards for reporting chronic periodontitis prevalence and severity in epidemiologic studies: Proposed standards from the Joint EU/USA Periodontal Epidemiology Working Group. *Journal of Clinical Periodontology* 42, 407–412. [PubMed: 25808877]

- Jentsch HF, Buchmann A, Friedrich A, and Eick S. (2016). Nonsurgical therapy of chronic periodontitis with adjunctive systemic azithromycin or amoxicillin/metronidazole. *Clinical Oral Investigations* 20, 1765–1773. [PubMed: 26685849]
- Johnson-Spruill I, Hammond P, Davis B, McGee Z, and Loudon D. (2009). Health of Gullah families in South Carolina with Type-2 diabetes: Diabetes self-management analysis from Project SuGar. *The Diabetes Educator* 35, 117–123. [PubMed: 19244567]
- Koenker R. (2005). *Quantile Regression* Number 38 in *Econometric Society Monographs*. Cambridge University Press.
- Liang K-Y and Zeger SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.
- Ma Y, Genton MG, and Tsiatis AA (2005). Locally efficient semiparametric estimators for generalized skew-elliptical distributions. *Journal of the American Statistical Association* 100, 980–989.
- Müller P, Quintana FA, Jara A, and Hanson T (2015). *Bayesian Nonparametric Data Analysis*. Springer.
- Nelsen RB (2007). *An Introduction to Copulas*. Springer Science & Business Media.
- Reich B. and Bandyopadhyay D. (2010). A latent factor model for spatial data with informative missingness. *Annals of Applied Statistics* 4, 439–459.
- Sahu SK, Dey DK, and Branco MD (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics* 31, 129–150.
- Sethuraman J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica* 4, 639–650.
- Vehtari A, Gelman A, and Gabry J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27, 1413–1432.

Figure 1:

Comparing contour plots, when $\alpha_1 = \alpha_2 = 10$, $\rho = 0.7$ and $\sigma_1 = \sigma_2 = 1$. Bivariate parametric Gaussian mixture (PGM) model with marginal densities (left panel); bivariate version of the multivariate skew-normal (MSN) model with marginal densities (right panel).

Figure 2: GAAD data: Histograms of the Periodontal Pocket Depth (PPD) responses (panel a) and Clinical Attachment Level (CAL) responses (panel b); Histogram (panel c) and Q-Q plot (panel d) of the empirical Bayesian estimates of the random effects; Histograms of residuals for PPD (panel e) and CAL (panel f), obtained after fitting linear mixed models.

Figure 3: GAAD data: Histograms of residuals for PPD (panel a) and CAL (panel b) responses, obtained after fitting the SMS model. Contour plot of the bivariate density of the residuals are presented in panel c

Table 1:

Simulation results based on 500 replicates of the data comparing the Mean Squared Error (MSE), Standard Error (SE), Relative Bias (RB), and Coverage Probability (CP) for association parameter $\rho_{12} = 0.7$ sample size is $n = 50$. True parameter values are $(\beta_0, \beta_1, \beta_2, a_1, a_2) = (2, 0.5, 0.5, 2, 2)$.

	SPGM						PGM						GEE					
	β_0	β_1	β_2	a_1	a_2	β_0	β_1	β_2	a_1	a_2	β_0	β_1	β_2	a_1	a_2	β_0	β_1	β_2
100×MSE	1.702	0.376	0.432	16.22	19.995	1.461	0.359	0.397	18.07	14.88	56.897	0.479	0.545					
10×SE	0.948	0.616	0.636	2.361	3.036	0.766	0.600	0.627	2.966	2.776	0.687	0.632	0.640					
RB	0.045	0.004	-0.025	-0.016	-0.016	0.047	0.005	-0.016	-0.153	0.134	0.375	0.019	-0.009					
CP	0.945	0.95	0.975	0.975	1	0.95	0.96	0.96	1	0.98	0.2	0.91	0.9					

Table 2:

Posterior estimates of the regression parameters and the skewness parameters (a_1 and a_2), corresponding to the responses 'periodontal pocket depth' (PPD) and 'clinical attachment level' (CAL), obtained after fitting the Semiparametric Multivariate Skew (SMS) response model, parametric Gaussian mixture (PGM), parametric skew- t (ST), and bivariate normal (BVN) models. SD and CI denote the posterior standard deviation and 95% credible intervals, respectively.

	SMS				PGM				ST				BVN			
	PPD	Mean	SD	CI	Mean	SD	CI	Mean	SD	CI	Mean	SD	Mean	SD	Mean	SD
Intercept		2.2624	0.3439	(1.6026, 2.9764)	2.1887	0.3881	(1.4807, 2.9230)	2.0865	0.6557	(1.4807, 2.9230)	2.0865	0.6557	2.1169	0.6298	2.1169	0.6298
Age		-0.0062	0.0037	(-0.0137, 0.0011)	-0.0058	0.0043	(-0.0137, 0.0032)	-0.0069	0.0076	(-0.0137, 0.0032)	-0.0069	0.0076	-0.0098	0.0073	-0.0098	0.0073
Gender		-0.2005	0.0918	(-0.3793, -0.0313)	-0.2096	0.1219	(-0.4301, -0.0448)	-0.1458	0.1861	(-0.4301, -0.0448)	-0.1458	0.1861	-0.0811	0.1753	-0.0811	0.1753
BMI		-0.0903	0.0797	(-0.2302, 0.0759)	-0.1378	0.1021	(-0.3769, 0.0505)	-0.2259	0.1402	(-0.3769, 0.0505)	-0.2259	0.1402	-0.1056	0.1645	-0.1056	0.1645
Smoker		0.0544	0.0925	(-0.1244, 0.2343)	0.0746	0.0987	(-0.1270, 0.2534)	0.1663	0.1767	(-0.1270, 0.2534)	0.1663	0.1767	0.1485	0.1696	0.1485	0.1696
HbA1c		0.0801	0.0797	(-0.0704, 0.2431)	0.1331	0.0634	(0.0156, 0.2618)	0.2191	0.1546	(0.0156, 0.2618)	0.2191	0.1546	0.1701	0.1501	0.1701	0.1501
a_1		0.5862	0.1848	(0.1631, 0.9053)	0.6968	0.1176	(0.4709, 0.9378)	0.0687	0.0843	(0.4709, 0.9378)	0.0687	0.0843	-	-	-	-
CAL																
Intercept		1.2180	0.4328	(0.3866, 2.0272)	0.9896	0.4851	(-0.0245, 1.8539)	1.4133	0.6813	(-0.0245, 1.8539)	1.4133	0.6813	1.1401	0.6485	1.1401	0.6485
Age		0.0062	0.0042	(-0.0018, 0.0148)	0.0081	0.0057	(-0.0025, 0.0198)	0.0050	0.0078	(-0.0025, 0.0198)	0.0050	0.0078	0.0073	0.0075	0.0073	0.0075
Gender		-0.3339	0.1059	(-0.5560, -0.1528)	-0.3556	0.1592	(-0.6681, -0.0743)	-0.2505	0.1968	(-0.6681, -0.0743)	-0.2505	0.1968	-0.3094	0.1872	-0.3094	0.1872
BMI		-0.0392	0.0993	(-0.2249, 0.1658)	-0.1022	0.1229	(-0.3778, 0.1188)	-0.1936	0.1619	(-0.3778, 0.1188)	-0.1936	0.1619	-0.1218	0.1725	-0.1218	0.1725
Smoker		0.1495	0.1441	(-0.0510, 0.3444)	0.2546	0.1145	(0.0259, 0.4605)	0.2835	0.1865	(0.0259, 0.4605)	0.2835	0.1865	0.3581	0.1813	0.3581	0.1813
HbA1c		0.1668	0.0979	(0.0047, 0.3791)	0.2392	0.0864	(0.0708, 0.3981)	0.3005	0.1623	(0.0708, 0.3981)	0.3005	0.1623	0.3143	0.1573	0.3143	0.1573
a_2		1.0500	0.1617	(0.7552, 1.4063)	1.1646	0.1333	(0.9216, 1.4275)	0.5294	0.0919	(0.9216, 1.4275)	0.5294	0.0919	-	-	-	-