Semiparametric Bayesian latent variable regression for skewed multivariate data

Apurva Bhingare 1,*, Debajyoti Sinha2, Debdeep Pati3, Dipankar Bandyopadhyay4 and 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

*email: bhingareapurva@gmail.com

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.

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

This paper has been submitted for consideration for publication in Biometrics
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 (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 antependence 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 \( Y_i = (Y_{i1}, \ldots, Y_{im}) \in \mathbb{R}^m \) be a row vector of skewed responses from subject (cluster) \( i = 1, \ldots, n \). For brevity, we use fixed number of \( m \) components per subject, which can be generalized to accommodate unequal \( m_i \). Our regression model is

\[
Y_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad \text{for } j = 1, \ldots, m, \tag{1}
\]

where \( \mu_{ij} = x_{ij}^T \beta_j \), \( x_{ij} = (1, x_{ij1}, \ldots, x_{ij,p-1}) \) is the row vector of \((p-1)\) covariates for \( Y_{ij} \) (with first term for the intercept), \( \beta_j \) is the corresponding \((p \times 1)\) vector of regression parameters, and \( \varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im}) \) 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 \( R_\rho \) is a \((m \times m)\) correlation matrix, \( h = (h_1, \ldots, h_m) \) is a vector of \( m \) unknown nonparametric (baseline) symmetric univariate densities, and \( \alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m \) is a vector of skewness parameters. This SMS\((0, h, R_\rho, \alpha)\) density has the following stochastic representation

\[
\varepsilon_i^T = A_\alpha |Z_{1i}| + A_\alpha^* Z_{2i}, \tag{2}
\]

where \( Z_{1i} = (Z_{1i1}, \ldots, Z_{1im})^T \) and \( Z_{2i} = (Z_{2i1}, \ldots, Z_{2im})^T \) are two independent \((m \times 1)\) vectors of latent variables with \( h_j(\cdot) \) being the common marginal density for \( Z_{1ij} \) and \( Z_{2ij} \), \( A_\alpha \) and \( A_\alpha^* \) 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 \( \varepsilon_{ij} \) in (2) can also be expressed as

\[
\varepsilon_{ij} = a_j Z_{2ij} + a_j^* |Z_{1ij}|, \quad \text{using the “skewing shock” } |Z_{1ij}| \text{ to the symmetric random noise } Z_{2ij} \text{ with weights } a_j^* \text{ and } a_j, \text{ such that } a_j^2 + a_j^{*2} = 1. \text{ Only when the skewness parameters } (\alpha_1, \ldots, \alpha_m) \text{ are all equal to zero, the } \varepsilon_i \text{ of (2) equals } Z_{2i} \text{ with symmetric marginal densities.
We assume \( m \) “skewing shocks” \( Z_{1ij}, j = 1, \ldots, m \) to be independent with marginal densities \( h_1(\cdot), \ldots, h_m(\cdot) \), and we model the association within \( \varepsilon_i \) via the \( m \)-variate Gaussian Copula \( \text{Nelsen} \) \( 2007 \).

\[
C_m(Z_{2i} \mid h, R_\rho) = \phi_m \left[ \Phi_1^{-1}\{H_1(z_{2i1})\}, \ldots, \Phi_1^{-1}\{H_m(z_{2im})\} \mid R_\rho \right] \times \prod_{j=1}^m \frac{h_j(z_{2ij})}{\phi_1 \left[ \Phi_1^{-1}\{H_j(z_{2ij})\} \right]} \tag{3}
\]

for \( Z_{2i} \), where \( \phi_m(\cdot \mid M) \) is the \( m \)-variate Gaussian density with mean 0 and variance-covariance matrix \( M \), \( R_\rho \) is a correlation matrix with unknown parameter vector \( \rho \), \( \Phi_1^{-1}(\cdot) \) 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 \( j = 1, \ldots, m \). We model \( h_j(\cdot) \) as a scale mixture of symmetric (around 0) density kernel \( K(\cdot \mid \sigma) \), given by

\[
h_j(u) = \int_0^\infty K(u \mid \sigma) dG_j(\sigma), \tag{4}
\]

where \( G_j(\cdot) \) is the unknown nonparametric mixing distribution of scale \( \sigma \) 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 \( K(\cdot \mid \sigma) \) in (4). However, other choices are possible. For example, using the result of \( \text{Feller} \) \( 1971 \) (p. 158), when the kernel \( K(\cdot \mid \sigma) \) 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 \( \text{Ferguson} \) \( 1974 \), \( \text{Müller et al.} \) \( 2015 \) \( \text{DP}(G_{0j}, C) \) on \( G_j(\cdot) \), with pre-specified prior mean \( G_{0j}(\cdot) \) and precision \( C \). In Section 3, we explain that this \( \text{DP}(G_{0j}, C) \) 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 \( \text{Sahu et al.} \) \( 2003 \). However, our semiparametric model with nonparametric marginals has two additional crucial differences: (a) a new parametrization of weights \( (A_\alpha, A^*_\alpha) \), 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 SMS($0, h, R, \alpha$) density of (2)-(3) is obtained when $Z_{1i}$ and $Z_{2i}$ in (2) have parametric mean-zero multivariate normal densities $N_m(0, D^2_\sigma)$ and $N_m(0, \Sigma)$ respectively with variance-covariance matrices $\Sigma = D_{\sigma} R_{\rho} D_{\sigma}$, where $D_{\sigma} = \text{diag}(\sigma_1, \ldots, \sigma_m)$, and $R_{\rho}$ is a positive definite correlation matrix. We call this a (multivariate) parametric Gaussian mixture (PGM) model, denoted by $\varepsilon_i \sim \text{PGM}(0, \alpha, \Sigma)$. Despite similar parametrization based on $(\alpha, \Sigma)$, 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 $\varepsilon_i \sim \text{MSN}(0, \alpha, \Sigma)$, with the corresponding joint density $f_m(\varepsilon_{1i}, \ldots, \varepsilon_{mi}) = 2\phi_m(\varepsilon_{i} | \Sigma) \Phi_1(\varepsilon_i^T \alpha)$. The marginal distribution of each component $\varepsilon_{ij}$ for PGM($0, \alpha, \Sigma$) density is the univariate skew-normal (SN) density $f_j(\varepsilon_{ij}) = 2\phi_1(\varepsilon_{ij} | \sigma_j) \Phi_1(\alpha_j \varepsilon_{ij})$ denoted by SN($0, \alpha_j, \sigma_j$) and dependent only on the component-specific parameters $(\alpha_j, \sigma_j)$, whereas, for the MSN($0, \alpha, \Sigma$) model of Azzalini and Dalla Valle (1996), the corresponding univariate marginal density is SN($0, \bar{\alpha}_j, \sigma_j$), however, with parameter $\bar{\alpha}_j$ being a function of the entire $(\alpha, \Sigma)$ (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 $\varepsilon_i^{(1)}$ of $\varepsilon_i = (\varepsilon_i^{(1)}, \varepsilon_i^{(2)})$ is from the same class of semiparametric density SMS($0, h^{(1)}, R_{11}, \alpha^{(1)}$) indexed by $(h^{(1)}, R_{11}, \alpha^{(1)})$, where $\alpha = (\alpha^{(1)}, \alpha^{(2)})$, $(h_1, \ldots, h_m) = (h^{(1)}, h^{(2)})$ and matrix $R = \begin{pmatrix} R_{11} & R_{12} \\ R_{12} & R_{22} \end{pmatrix}$ are partitions of $(\alpha, h, R)$. Unlike the multivariate model of (2) based on multiple “skewing shocks” $|Z_{1i1}|, \ldots, |Z_{1im}|$, the existing multivariate models, such as MSN($0, \alpha, \Sigma$) are based on univariate “skewing shock” $|Z_{1i1}|$ (common to all $m$ components). As a consequence, these models are closed under marginalization only
in a restricted sense because \( \varepsilon_i \sim MSN(0, \alpha, \Sigma) \Rightarrow \varepsilon_i^{(1)} \sim MSN(0, \alpha^{(1)}_1, \Sigma^{(1)}_1) \), where the parameters \( (\alpha^{(1)}_1, \Sigma^{(1)}_1) \) are different from \( (\alpha^{(1)}_1, \Sigma_{11}) \), 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 \( |Z_{i1}^*| \) common to both CAL and PPD, compared to the paired skewing shocks \( (|Z_{i1}|, |Z_{i2}|) \) in model (2). Consequently, \( \bar{\alpha}_1 \), depends on \( \alpha_1 \), as well as \( (\alpha_2, \sigma_2, \rho) \). This complicates the marginal interpretation and Bayesian prior specification of \( \alpha_1 \) for the MSN model.

To further demonstrate the lack of interpretation of \( \alpha \) 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 \( \alpha_0 \) 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 \( \rho \) 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 \( \alpha_0 \). However, for our multivariate model in (2), the Pearsonian skewness of each marginal density does not depend on \( \rho \) 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 \( \beta_{j0}^* = \beta_{j0} + \{\alpha_j/(1 + \alpha_j^2)^{1/2}\} \lambda_j \). The covariance matrix of \( Y_i \) is \( \text{var}(Y_i|X_i) = V \), where
the diagonal elements $V_{jj} = \sigma^2_{hj} - \{\alpha_j^2/(1 + \alpha_j^2)\}\lambda_j^2$, 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 $j \neq k \in \{1, \ldots, m\}$. Here, the common expectations $E(|Z_{1ij}|) = \lambda_j$ and variances $Var(Z_{1ij}) = Var(Z_{2ij}) = \sigma^2_{hj}$ are taken with respect to the nonparametric marginal density $h_j(\cdot)$ of $Z_{1ij}$ and $Z_{2ij}$, given in (4). The covariance $Cov(Z_{2ij}, Z_{2ik}) = E[Z_{2ij}Z_{2ik}] = \sigma_{hjk}$ 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 $\text{Corr}(\epsilon_{i1}, \epsilon_{i2})$ between PPD and CAL responses within subject/cluster $i$ is $\sigma_{h12}a_1^*a_2^*/[\{\sigma^2_{h1} - \alpha_1^2\lambda_1^2\}\{\sigma^2_{h2} - \alpha_2^2\lambda_2^2\}]^{1/2}$, 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 $D = \{y_i, X_i; i = 1, \ldots, n\}$ from $n$ independent multivariate responses, the likelihood function of the parameter space $\Theta = (\beta, \alpha, \rho, \sigma, G)$ is

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

(5)

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

$$p(\Theta | D) \propto L(\Theta | D)p_1(\beta)p_2(\alpha)p_3(\rho)p_4(G),$$

(6)

with the following priors.

(i) The priors $p_1(\beta)$ and $p_2(\alpha)$ for $\beta$ and $\alpha$ are independent mean zero multivariate normal with pre-specified covariance matrices $M_\beta$ and $M_\alpha$, respectively. In particular, we choose these as $\sigma^2 I$, where $\sigma^2 = 100$.

(ii) For the GAAD study with bivariate response, the prior $p_3(\rho)$ of the scalar parameter $\rho$ in
\( R_\rho \) is Uniform\((-1, 1)\). However, for a vector \( \rho \), this \( \pi_3(\rho) \) prior has to be a multivariate density with an appropriate support.

(iii) The joint prior process \( \pi_4(G) \) of \( G = (G_1, \ldots, G_m) \) is the product of \( m \) independent Dirichlet processes \( DP(G_0j, C_j) \), for \( j = 1, \ldots, m \) [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

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

where \( \phi_m(u|0, R) \) with \((m \times m)\) covariance matrix \( R \) is a mean-zero multivariate normal density evaluated at \( u \in \mathbb{R}^m \), \( W_i = (W_{i1}, \ldots, W_{im}) \) with \( W_{ij} = \Phi_1^{-1}[H_j\{(1 + \alpha_j^2)^{1/2}(y_{ij} - X_{ij}\beta) - \alpha_j|Z_{1ij}|\}] \), \( \Phi_1^{-1} \) 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 \( \sigma_j \), we use the parametric prior distributions \( \sigma_j \sim g_j(\cdot|\gamma_j) \) independently for \( j = 1, \ldots, m \) with pre-specified hyperparameters \( \gamma_j \). For this special case, \( W_{ij} \) simplifies to \( \{(1 + \alpha_j^2)^{1/2}(y_{ij} - X_{ij}\beta) - \alpha_j|Z_{1ij}|\}/\sigma_j \).

For a semiparametric analysis with a particular symmetric density kernel \( \mathcal{K}(z|\sigma) \) for the nonparametric kernel-mixture densities \( h = (h_1, \ldots, h_m) \) in (4), we implement the joint density in (7) by assuming (a) \((W_{i1}, \ldots, W_{im}) \) for \( i = 1, \ldots, n \) are mean-zero multivariate \( N_m(0, R_\rho) \), where \( W_{ij} = \Phi_1^{-1}[K_{ij}\{(1 + \alpha_j^2)^{1/2}(y_{ij} - X_{ij}\beta) - \alpha_j|Z_{1ij}|\}] \), and \( K_{ij} \) and \( K_{ij}^{-1} \) are the cdf and inverse-cdf of \( \mathcal{K}(\cdot|\sigma_{ij}) \); (b) \( Z_{1ij} \) for \( j = 1, \ldots, m \) are independent with density

\[
\]
Bayesian semiparametric multivariate skewed models

\( K(\cdot|\sigma_{ij}) \), and (c) \( \sigma_{ij} \) for \( j = 1, \ldots, m \) are independent with nonparametric cdf \( G_j \) following Dirichlet Process (DP) prior \( \text{DP}(G_{0j}, C) \). For the Gaussian kernel in [1], \( K_{ij}(u) = \Phi(u/\sigma_{ij}) \) and \( W_{ij} \) further simplifies to

\[
W_{ij} = \{ (1 + \alpha_j^2)^{1/2}(y_{ij} - X_{ij}\beta) - \alpha_j |Z_{ij}| \} / \sigma_{ij}.
\]

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 \( (\delta_1, \ldots, \delta_{K_0}) \) are generated independently from \( G_{0j} \), \( (B_1, \ldots, B_{K_0-1}) \) are generated independently from Beta\((1, C)\), \( \omega_1 = B_1 \), \( \omega_h = B_k \prod_{l<k}(1 - B_l) \) for \( k = 2, \ldots, K_0 - 1 \) and \( \omega_{K_0} = 1 - \sum_{l=1}^{K_0-1} \omega_l \). Then, the sample path of \( G_j \) is approximated by \( G_j = \sum_{l=1}^{K_0} \omega_l I_{\delta_l} \), 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, \( \beta \), 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,

\[
f_0(e) = \int_{\mathbb{R}^m} 2^m |J| \prod_{j=1}^m \phi_1 \left[ \phi^{-1}\{H_{0j}(e_j^*)\} \right] \phi_m \left[ \Phi^{-1}\{H_{01}(e_1^*)\}, \ldots, \Phi^{-1}\{H_{0m}(e_m^*)\}\right] R_\rho \, dz,
\]

with the \( j \)-th marginal density given by \( f_{0j}(e_j) = \int_{\mathbb{R}_+} 2^{1/2} h_{0j}(z_j) h_{0j}(e_j^*) \, dz_j \) \( j = 1, \ldots, m \), where \( h_{0j}(\cdot) \) is the true (unknown) symmetric around 0 density, \( H_{0j}(\cdot) \) is the
corresponding cumulative distribution function, $e_j^* = \sqrt{1 + \alpha_j^2 e_j - \alpha z_j}$, and the Jacobian of transformation $|J|$ is $|J| = \begin{vmatrix} I & 0 \\ A^{-1}_\alpha & -A^{-1}_\alpha A \end{vmatrix}$. Let $\Pi$ denote a prior on $\mathcal{C}\{(\mathcal{F})^m \times (0, 1)\}$, such that $\Pi = (\pi F)^m \otimes \pi_3(\rho)$, where $\pi F$ is a prior on $F$ defined by $G_j \sim \text{DP}(G_0^j, C)$ for $j = 1, \ldots, m$, $\alpha \sim \pi_2(\alpha)$ and $\pi_3(\rho)$ is a prior on $(0, 1)$. For ease of exposition, we will assume that $\alpha$ is known. To show that posterior consistency holds at the true density $f_0$, we first define the Kullback-Liebler divergence as $\text{KL}(f_0, f) = \int_{\mathbb{R}^m} f_0 \log(f_0/f)\, \text{d}x$ and Kullback-Leibler neighborhood of size $\epsilon$ as $\kappa_\epsilon(f_0) = \{f : \text{KL}(f_0, f) < \epsilon\}$. For any prior $\Pi^*$ on the density space $\mathcal{F}^*$, $f_0$ is said to be in the Kullback-Leibler support of $\Pi^*$ denoted by $\text{KL}(\Pi^*)$ if $f_0 \in S$, where $S = \{f_0 : \Pi^*(f : \text{KL}(f_0, f) < \epsilon) > 0\}$. Define $\mathcal{F}_{KL} = \{f_0 \in \mathcal{C}\{(\mathcal{F})^m \times (0, 1)\}, \int f_0 |\log f_0| < \infty\}$. We characterize the Kullback-Leibler support of $\Pi$ in the following lemma, with its proof available in Web Appendix C.

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

Suppose $D_n$ is the observed data with sample size $n$, and $\beta_0$ 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 $\beta$ 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 $\beta$; so, we define a strong $L_1$ neighborhood of radius $\delta > 0$ around the true value $\beta_0$ as the set $\mathcal{S}(\beta_0) = \{\beta : \|\beta - \beta_0\| < \delta\}$, and weak neighborhood around $f_0$ as $\mathcal{W}_\epsilon(f_0) = \{f : |\int \varphi f - \int \varphi f_0| < \epsilon\}$ for a bounded continuous function $\varphi$. 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, \( \hat{\Pi}\{f_m, \beta \in U^c|D_n}\} \to 0 \) a.s. under the true data generating distribution \( P_{f_0, \beta_0}\) that generates data \( D_n \).

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 \( (\varepsilon_{ij1}, \varepsilon_{ij2}) \) were generated from the bivariate PGM model, where \( \alpha = (\alpha_1, \alpha_2) = (2, 2) \), \( Z_{1ij} \sim N_2(0, I) \) and \( Z_{2ij} \sim N_2(0, R_\rho) \), \( R_\rho \) is a \((2 \times 2)\) 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 \( \beta \)) for both components (different from separate \( \beta \) 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 \( Z_{1ij} \sim N(0, \sigma_{ij}^2) \) and \( Z_{2ij} \sim N_m(0, \Sigma) \), 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 \( \beta_0, \beta_1 \) and \( \beta_2 \), and also for the skewness parameters \( \alpha_1 \) and \( \alpha_2 \). 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 Gamma\((2, 1)\) density and \( C = 1 \), implying the prior guess for \( \varepsilon_{ij} \) is mean 0 with ranges \( \pm 6 \), 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 \( \varepsilon_i \), we would like to demonstrate that even this guess produces posterior estimates with good finite sample properties. For the PGM, we specify Gamma\((2, 1)\) for \( \sigma_1^2 = \sigma_2^2 \). To test the performances, we use the mean squared error
\[
\text{MSE}(\hat{\theta}) = \frac{1}{N} \sum_{k=1}^{N} (\hat{\theta}_k - \theta)^2,
\]
where \( \hat{\theta}_k \) is the posterior estimate of \( \theta \) from the \( k \)th simulated dataset, \( k = 1, 2, \ldots, N \). We also report the relative bias (RB),
\[
\text{RB} = \frac{1}{N} \sum_{k=1}^{N} (\hat{\theta}_k - \theta)/N\theta,
\]
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
Bayesian semiparametric multivariate skewed models

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 $\geq 30$), 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 \texttt{nlme} package in \textit{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 $Z_{1i} \sim N_2(0, D_2^2)$ and $Z_{2i} \sim N_2(0, \Sigma)$, where $D_2^2 = \text{diag}(\sigma_1^2, \sigma_2^2)$ 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 $(\varepsilon_{i1}, \varepsilon_{i2})$, with $\nu$ degrees of freedom,

4. The Bivariate Normal (BVN) model, with a symmetric $N_2(0, \Sigma)$ density for $(\varepsilon_{i1}, \varepsilon_{i2})$, ignoring the skewness.

We use practically flat independent $N(0,100)$ priors for all the regression parameters
(the components of $\beta_1$ and $\beta_2$), and the skewness parameters $\alpha_1$ and $\alpha_2$. For the unknown mixing distributions $G_1$ and $G_2$, we choose independent DP($G_{0j}, C$) 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 $\sigma_1$ and $\sigma_2$ in the PGM model, the skew-$t$ and bivariate normal models. For the degrees of freedom $\nu$ of the skew-$t$ density, we choose Exp(0.1)$I_{(2,\infty)}$, the exponential density truncated at 2. For all four models, we use a Uniform($-1, 1$) as the prior for $\rho$. 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, $(\text{CPO})_i = \int f_m(y_i - X_i\beta | \Theta)p(\Theta | D_{(-i)})d\Theta$ [Gelfand et al. 1992], to compare the model performances, where $D_{(-i)}$ is the cross-validated data after deleting the $i^{th}$ observation from the full data $D$, and $p(\Theta | D_{(-i)})$ is the posterior density of $\Theta$ given $D_{(-i)}$. 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 $p(\Theta | D)$. 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=...
Bayesian semiparametric multivariate skewed models

-127.882 and \(\text{waic} = 441.89\)), followed by the SMS model (LPML = -134.03 and \(\text{waic} = 516.88\)). Recall that the PGM model is the parametric version of the SMS model. The LPML (\(\text{waic}\)) 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 \(\alpha_1\) (for PPD) and \(\alpha_2\) (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 \(\alpha_1\) 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 patient_1 (Female, low HbA1c) and patient_2 (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 patient_1 and patient_2.

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
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 \(|Z_{1i1}|, \ldots, |Z_{1im}|\) in (2) be equal to a single univariate cluster-specific skewing shock variable \(|Z^\ast_i|\). 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 \( \alpha \) and \( \rho \), 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 \( \varepsilon_{ij} \) is only a function of \((\alpha_j, h_j)\). 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 \( Z_{1i1}, \ldots, Z_{1im} \). 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 $R_p$. Also, for our model, the adopted association structure (form of $R_p$) 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.

8 Supplementary 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

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


Bayesian semiparametric multivariate skewed models


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). (This figure appears in color in the electronic version of this article).
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 $(\beta_0, \beta_1, \beta_2, \alpha_1, \alpha_2) = (2, 0.5, 0.5, 2, 2)$.

<table>
<thead>
<tr>
<th></th>
<th>SPGM</th>
<th>PGM</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_0$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>100×MSE</td>
<td>1.702</td>
<td>0.376</td>
<td>0.432</td>
</tr>
<tr>
<td>10×SE</td>
<td>0.948</td>
<td>0.616</td>
<td>0.636</td>
</tr>
<tr>
<td>RB</td>
<td>0.045</td>
<td>0.004</td>
<td>-0.025</td>
</tr>
<tr>
<td>CP</td>
<td>0.945</td>
<td>0.95</td>
<td>0.975</td>
</tr>
</tbody>
</table>
Table 2: Posterior estimates of the regression parameters and the skewness parameters ($\alpha_1$ and $\alpha_2$), 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.

<table>
<thead>
<tr>
<th></th>
<th>SMS</th>
<th>PGM</th>
<th>ST</th>
<th>BVN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>CI</td>
<td>Mean</td>
</tr>
<tr>
<td>PPD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>2.2624</td>
<td>0.3439</td>
<td>(1.6026, 2.9764)</td>
<td>2.1887</td>
</tr>
<tr>
<td>Age</td>
<td>-0.0062</td>
<td>0.0037</td>
<td>(-0.0137, 0.0011)</td>
<td>-0.0058</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.2005</td>
<td>0.0918</td>
<td>(-0.3793, -0.0313)</td>
<td>-0.2096</td>
</tr>
<tr>
<td>BMI</td>
<td>-0.0903</td>
<td>0.0797</td>
<td>(-0.2302, 0.0759)</td>
<td>-0.1378</td>
</tr>
<tr>
<td>Smoker</td>
<td>0.0544</td>
<td>0.0925</td>
<td>(-0.1244, 0.2343)</td>
<td>0.0746</td>
</tr>
<tr>
<td>HbA1c</td>
<td>0.0801</td>
<td>0.0797</td>
<td>(-0.0704, 0.2431)</td>
<td>0.1331</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.5862</td>
<td>0.1848</td>
<td>(0.1631, 0.9053)</td>
<td>0.6968</td>
</tr>
<tr>
<td>CAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1.2180</td>
<td>0.4328</td>
<td>(0.3866, 2.0272)</td>
<td>0.9896</td>
</tr>
<tr>
<td>Age</td>
<td>0.0062</td>
<td>0.0042</td>
<td>(-0.0018, 0.0146)</td>
<td>0.0081</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.3339</td>
<td>0.1059</td>
<td>(-0.5560, -0.1528)</td>
<td>-0.3556</td>
</tr>
<tr>
<td>BMI</td>
<td>-0.0392</td>
<td>0.0933</td>
<td>(-0.2249, 0.1658)</td>
<td>-0.1022</td>
</tr>
<tr>
<td>Smoker</td>
<td>0.1405</td>
<td>0.0921</td>
<td>(-0.0510, 0.4041)</td>
<td>0.2546</td>
</tr>
<tr>
<td>HbA1c</td>
<td>0.1468</td>
<td>0.0979</td>
<td>(0.0047, 0.3791)</td>
<td>0.2392</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>1.0500</td>
<td>0.1617</td>
<td>(0.7952, 1.4063)</td>
<td>1.1646</td>
</tr>
</tbody>
</table>
Web Appendix A: Restrictive parameterization in the Azzalini & Dalla Valle’s multivariate skew-normal model

Unlike our semiparametric multivariate skew (SMS) model (equation 2 in the paper), parameters \((\alpha^*_1, \Sigma^*_1)\) in the marginal multivariate skew-normal (MSN) specification (Azzalini and Dalla Valle, 1996) of \(\varepsilon_i^{(1)}\) are functions of \((\alpha^{(2)}, \Sigma_{22}, \Sigma_{12})\); see Theorem 1 of Chang and Zimmerman (2016) for details. In particular, the scalar component \(\varepsilon_{ij}\) has the univariate skew-normal distribution (Azzalini, 1985) denoted by \(\text{SN}(0; \bar{\alpha}_j, \sigma_j)\) with density

\[
f(\varepsilon_{ij}) = 2\phi_1(\varepsilon_{ij}|\sigma_j)\Phi_1(\bar{\alpha}_j \varepsilon_{ij}),
\]

where \(\phi_1(.|\sigma)\) is the \(N(0,\sigma^2)\) density, \((\Sigma)_{jj} = \sigma_j^2\), and the marginal skewness parameter \(\bar{\alpha}_j\) is

\[
\bar{\alpha}_j = \frac{\alpha_j + (1/\sigma_j^2)\Sigma^T_{(-j)}\alpha_{(-j)}}{1 + \alpha^T_{(-j)}\Sigma_{(-j,-j)}\alpha_{(-j)} - (1/\sigma_j^2)(\Sigma_{j(-j)}\alpha_{(-j)})^2},
\]

where \(\Sigma_{.j}\) is the \(j\)-th column of the covariance matrix \(\Sigma\), \(R_{(-j)}\) denotes the reduced vector after deleting \(j\)-th element of the vector \(R\), and \(Q_{(-j,-k)}\) denotes the reduced matrix after deleting \(j\)-th row and \(k\)-th column of matrix \(Q\). It is obvious from (A-1)-(A-2) that the marginal distribution of \(\varepsilon_{ij}\), including its mean, variance and skewness, clearly depends on all the components of vector \(\alpha\) and matrix \(\Sigma\) via \(\bar{\alpha}_j\), which is a function of \(\alpha_k\) and \(\sigma_k\), \(\forall\ k = 1, \ldots, m\).
To illustrate the difference between the marginal distribution from the bivariate PGM model (parametric subclass of SMS model) and the marginal distribution from the bivariate MSN model of Azallini & Dalla Valle, we consider $\alpha = (\alpha_1, \alpha_2)$ and $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}$, common to both bivariate models. Under the MSN model, the univariate SN marginal density (Azzalini, 1985) is $\text{SN}(0; \bar{\alpha}_1, \sigma_1)$, with the marginal skewness parameter $\bar{\alpha}_1 = \{\alpha_1 + \rho \alpha_2 \sigma_2 / \sigma_1\}/\{1 + \alpha_2^2 \sigma_2^2 (1 - \rho^2)\}$, with $\rho = \text{Corr}(\varepsilon_{i1}, \varepsilon_{i2}) = \sigma_{12}/\{\sigma_1 \sigma_2\}$, whereas, the marginal distribution of $\varepsilon_{i1}$ under the PGM model (parametric case of our model) is $\text{SN}(0; \alpha_1, \sigma_1)$. It is obvious that the magnitudes of $\alpha_1$ and $\bar{\alpha}_1$ are different, and even the signs of $\alpha_1$ and $\bar{\alpha}_1$ are different whenever $\alpha_1^2 \sigma_1 + \rho \alpha_1 \alpha_2 \sigma_2 < 0$. For Bayesian analysis, in particular, this complicates the prior specification for, say, $(\alpha_1, \sigma_1)$ based on available prior opinion about the marginal skewness $\bar{\alpha}_1$ and variability of PPD. This model also imposes a restriction that, even when $\alpha_1 = 0$, $\varepsilon_{i1}$ is skewed if $\alpha_2, \rho \neq 0$. In general, this model imposes several restrictions on within cluster association structure.

Web Appendix B: Contour plots of our bivariate PGM$(0, \alpha, \Sigma)$ density and the bivariate MSN density

Figure F1 compares the contour plot of our parametric bivariate PGM$(0, \alpha, \Sigma)$ density (right panel) with the contour plot of the parametric bivariate MSN$(0, \alpha, \Sigma)$ density of Chang and Zimmermann (2016) (left panel), for 4 choices of the common skewness parameter $\alpha_0 = \alpha_1 = \alpha_2 = 0, 2, 5, 10$, the common (for two components) scale parameter $\sigma_1 = \sigma_2 = 1$ and association parameter $\rho = 0.7$. Even when the common $\alpha_0$ is moderately large, the contours of these two bivariate densities are noticeably different. Both bivariate densities become increasingly non-elliptical as $\alpha_0$ increases. Although, the effect of an increasing $\alpha_0$ on the skewness of the marginal density of the bivariate MSN is not as profoundly obvious as this effect on the skewness of the marginal density of our PGM. For the bivariate MSN densities, there is even a moderately negative association in the left tail in spite of the association parameter $\rho$ being as large as 0.7.
Figure F1: Contour plots of the bivariate parametric Gaussian mixture (PGM) model density (right panel) and of the bivariate skew-normal (MSN) densities (left panel) for association parameter $\rho = 0.7$, common scale parameter $\sigma_1 = \sigma_2 = 1$, and different values of common skewness $\alpha_0 = \alpha_1 = \alpha_2$. 

(a) $\alpha_0 = 0$

(b) $\alpha_0 = 0$

(c) $\alpha_0 = 2$

(d) $\alpha_0 = 2$

(e) $\alpha_0 = 5$

(f) $\alpha_0 = 5$

(g) $\alpha_0 = 10$

(h) $\alpha_0 = 10$
Lemma 1 $\mathcal{F}_{KL} \subset KL(\Pi)$ if $G_{0j}$ is defined on support $\mathbb{R}_+$ for all $j = 1, \ldots, m$ and $G_{0j}$ is absolutely continuous with respect to Lebesgue measure and $\text{supp}(\pi_3(\rho)) = (0, 1)$.

Proof of Lemma 1

Given a density $f_0 \in \mathcal{F}$, the idea is to construct a sequence of functions $f^{(M)} \in \mathcal{C}(\mathcal{F})^m \times (0, 1)$ $M \geq 1$ such that $KL(f_0, f^{(M)}) \to 0$ as $M \to \infty$. Let $e = (e_1, \ldots, e_m)$, $A_\alpha = \text{diag}(\alpha_1(1 + \alpha_1^2)^{-1/2}, \ldots, \alpha_m(1 + \alpha_m^2)^{-1/2})$ and $A_\alpha^* = \text{diag}((1 + \alpha_1^2)^{-1/2}, \ldots, (1 + \alpha_m^2)^{-1/2})$. We assume $\alpha_1 = \alpha_2 = \ldots = \alpha_m = \alpha_0$, and $\alpha_0$ is known. Define,

$$f^{(M)}(e) = \int_{\mathbb{R}^m} 2^m |J| \prod_{j=1}^m \frac{h_{jM}(z_j)h_{jM}(e_j^*)}{\phi_1(\Phi^{-1}(H_{jM}(e_j^*)))} \phi_m \left[ \Phi^{-1}(H_{1M}(e_1^*)), \ldots, \Phi^{-1}(H_{mM}(e_m^*)) \right] |R_{\rho M}|$$ $dz$, 

where $e_j^* = (1 + \alpha_0^2)^{1/2}e_j - \alpha_0 z_j$ for $j = 1, \ldots, m$, $H_{jM}$ is the cdf of $h_{jM}$, $\rho_M \to \rho$ as $M \to \infty$, and $|J| = \begin{vmatrix} I & 0 \\ A_\alpha^{*-1} & -A_\alpha^{*-1}A_\alpha \end{vmatrix} = \prod_{j=1}^m |\alpha_j|$ with the marginal density given by,

$$f_{jM}(e_j) = \int_0^\infty 2(1 + \alpha_0^2)^{1/2}h_{jM}(z_j)h_{jM}(e_j^*)dz_j.$$ 

We now construct $h_{jM}$. Suppose $F_{jM}(\cdot)$ denotes the cdf of $f_{jM}(\cdot)$, and $h_{0j}$ is continuous and symmetric (around 0) density. Clearly, $h_{0j}$ is increasing $\mathbb{R}^-$ and decreasing on $\mathbb{R}^+$. We define weights as in Wu and Ghosal (2008). Suppose $t_1 > 0$ and $t_2 > 0$ such that $h_{0j}(t_1) = a_1$ and $h_{0j}(t_2) = b_1$, where $0 < b_1 < 1$ and $b_1 < a_1 < h_{0j}(0)$. For given $M$, let $M_1$ and $M_2$ be such that $\frac{M_1}{M} \leq t_1 \leq \frac{M_1+1}{M}$ and $\frac{M_2}{M} \leq t_2 \leq \frac{M_2+1}{M}$. Set

$$w_{ji} = \begin{cases} \frac{i}{M}(h_{0j}(\frac{i}{M}) - h_{0j}(\frac{i+1}{M})), & 1 \leq i < M_1, \\
\frac{M}{M}(h_{0j}(\frac{M}{M}) - a_1), & i = M_1, \\
\frac{M+1}{M}(a_1 - h_{0j}(\frac{M+1}{M})), & i = M_1 + 1, \\
\frac{i}{M}(h_{0j}(\frac{i-1}{M}) - h_{0j}(\frac{i}{M})), & M_1 + 1 < i \leq M_2, \\
\frac{i}{M}(h_{0j}(\frac{i-1}{M}) - h_{0j}(\frac{i}{M})), & i \geq M_1 + 1 \end{cases}$$
We define \( h_{jM}^*(e) = \sum_{1}^{\infty} w_{ji}^* K(e; \frac{i}{M}) \), where \( K(e; \theta) = \frac{1}{2b}1_{(-\theta \leq e \leq \theta)} \). By the continuity of \( h_{0j} \), \( h_{jM}^*(e) \) converges to \( h_{0j}(e) \) pointwise. However, \( \sum_{1}^{\infty} w_{ji}^* \neq 1 \) and \( h_{jM}^*(e) \) is not a pdf. We define \( w_{ji} = w_{ji}^* \frac{1 - \sum_{1}^{M_1-1} w_{ji} - \sum_{M_2+1}^{\infty} w_{ji}}{\sum_{M_2}^{\infty} w_{ji}} \) for \( M_1 \leq i \leq M_2 \). Then, \( \sum_{1}^{\infty} w_{ji} = 1 \). Let \( h_{jM}(e) = \sum_{1}^{\infty} w_{ji} K(e; \frac{i}{M}) \), where \( K(e; \theta) = \frac{1}{2b}1_{(-\theta \leq e \leq \theta)} \). Observe that

\[
h_{jM}^*(e) - h_{jM}(e) = \left( \sum_{1}^{\infty} w_{ji}^* - \sum_{1}^{\infty} w_{ji} \right)
\]

\[
= \left( 1 - \sum_{1}^{M_1-1} w_{ji} - \sum_{M_2+1}^{\infty} w_{ji} \right) - \left( 1 - \sum_{1}^{M_1-1} w_{ji} - \sum_{M_2+1}^{\infty} w_{ji} \right)
\]

\[
\leq \left( 1 - \sum_{1}^{M_1-1} w_{ji} - \sum_{M_2+1}^{\infty} w_{ji} \right) - \left( 1 - \sum_{1}^{M_1} w_{ji} - \sum_{M_2}^{\infty} w_{ji} \right) \frac{M}{2M_1}
\]

\[
= \left( 1 - \frac{1}{M} \sum_{1}^{\infty} h_{0j}(i/M) - a_1 \right) \frac{M}{2M_1} \to 0
\]

as \( M \to \infty \), by definition of Riemann integral. Thus \( h_{jM}(e) \) converges to \( h_{0j} \). Let \( M \) be large such that the RHS of the equation is less than \( \frac{\alpha}{2} \). Define \( f_{jM}(e) = \frac{2}{b} \int_{0}^{\infty} h_{jM}(z_j) h_{jM} \left( \frac{e-az_j}{b} \right) dz_j \), where \( a = \frac{\alpha}{\sqrt{1+\alpha^2}} \) and \( b = \frac{1}{\sqrt{1+\alpha^2}} \). Hence,

\[
|\log f_{jM}(e_j)| = |\log \left( \frac{2}{b} \right) + \log \int_{0}^{\infty} h_{jM}(z_j) h_{jM}(e_j^* )| \]

\[
\leq |\log \left( \frac{2}{b} \right) | + |\log \int_{0}^{\infty} h_{jM}(z_j) h_{jM}(e_j^* )|
\]

Since \( h_{jM} \to h_{0j} \) pointwise, by construction of \( h_{jM} \), we have \( c_1 h_{0j}(z_j) \leq h_{jM}(z_j) \leq c_2 h_{0j}(z_j) \), and \( c_1 h_{0j}(e_j^*) \leq h_{jM}(e_j^*) \leq c_2 h_{0j}(e_j^*) \). Thus we have

\[
c_1 \int_{0}^{\infty} h_{0j}(z_j) h_{0j}(e_j^*) \leq \int_{0}^{\infty} h_{jM}(z_j) h_{jM}(e_j^*) \leq c_2 \int_{0}^{\infty} h_{0j}(z_j) h_{0j}(e_j^*).
\]

Hence, we also have \( |\log f_{jM}(e_j)| \to 0 \). If \( f_0 \in F_{KL} \), we have \( \int f_{0j} |\log f_{0j}| < \infty \) for all \( j \). Since \( |\log h_{0j}(e_j)| \) is \( h_0 \)-integrable, using dominated convergence theorem, we have \( \int h_{0j} \log \frac{h_{0j}}{h_{jM}} \to 0 \) as \( M \to \infty \). Also, \( |\log f_{jM}(e_j)| \) is bounded above by an integrable function and hence \( \int f_{0j} \log \frac{f_{0j}}{f_{jM}} \to 0 \) as \( M \to \infty \). Below we show that \( f^{(M)} \to f_0 \) pointwise, and construct an \( f_0 \) integrable upper bound of \( g_M = \log \frac{f_0}{f_0^{(M)}} \). Since, \( f_{jM} \to f_{0j} \) pointwise for \( j = 1, \ldots, m \), by Scheffe’s theorem.
such that

By dominated convergence theorem, we conclude that $f_M \to f_0$ pointwise. Observe that

$$|g_M| \leq |\log f_0| + |\log f|^M|$$

and

$$|\log f|^M| \leq |\log(2^m|J|)| +$$

$$\left| \log \int_{\mathbb{R}^n} \prod_{j=1}^m h_{jM}(z_j) h_{jM}(e_j^*) \frac{1}{|\Sigma_M|^{1/2}} \exp \left[ -\frac{1}{2} H_M'(e^*)(\Sigma_M^{-1} - I) H_M'(e^*) \right] dz \right|. \quad (C-1)$$

Let, $I_M = \frac{1}{|\Sigma_M|^{1/2}} \exp \left[ -\frac{1}{2} H_M'(e^*)(\Sigma_M^{-1} - I) H_M'(e^*) \right]$. Then,

$$\log(I_M) = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} H_M'(e^*)(\Sigma_M^{-1} - I) H_M'(e^*)$$

$$= C_1 - \frac{1}{2} (H_M'(e^*) - H_0'(e^*))'(\Sigma_M^{-1} - I)(H_M'(e^*) - H_0'(e^*))$$

$$- \frac{1}{2} H_0'(e^*)(\Sigma_M^{-1} - \Sigma_0^{-1}) H_0'(e^*) - \frac{1}{2} H_0'(e^*)(\Sigma_0^{-1} - I) H_0'(e^*) \quad (C-2)$$

We have $H_M = (H_{1M}, \ldots, H_{mM})$ and $H_0 = (H_{01}, \ldots, H_{0m})$, where $H_{jM}(e_j^*) = \Phi^{-1}\{H_{jM}(e_j^*)\}$ and $H_0'(e_j^*) = \{H_0'(e_j^*)\}$ for $j = 1, \ldots, m$. Using the Taylor series expansion of $H_{jM}$ around $H_0$, we have for $\zeta \in [0, 1]$

$$H_{jM}^* = H_0^* + \frac{H_{jM}(e_j^*) - H_0(e_j^*)}{\phi_1(H_0(e_j^*))} + \frac{\phi'(\zeta)(H_{jM}(e_j^*) - H_0(e_j^*))}{\phi'(\zeta)}$$

Hence, $|H_{jM}^* - H_0^*| \to 0$ uniformly in $e_j^*$, and the first term in (C-4) is 0 as $M \to \infty$. As $\rho_M \to \rho_0$ as $M \to \infty$, the $m$ eigenvalues of $\Sigma_M^{-1}$ converge to the $m$ eigenvalues of $\Sigma_0^{-1}$. Therefore, $\frac{1}{2} H_0'(e^*)(\Sigma_M^{-1} - \Sigma_0^{-1}) H_0'(e^*) \leq \max_j |\lambda_{0j} - \lambda_{Mj}| H_0'(H_0'). \quad \text{Thus, there exists constants } C_1, C_2 > 0 \text{ such that}$

$$\exp\{-(C_1 + C_2|\log(I_0)|)\} \leq I_M \leq \exp\{C_1 + C_2|\log(I_0)|\}.$$ 

We know that $h_{jM}(\cdot) \to h_{0j}(\cdot)$ for all $j$ as $M \to \infty$. Hence

$$|\log f^{(M)}(e)| \leq |\log(2^m|J|)| + |\log \Psi|, \quad \text{with } \int f_0|\log f^{(M)}(e)| < \infty.$$ 

By dominated convergence theorem, we conclude that $\int f_0 \log \frac{f_0}{f^{(M)}} \to 0$ as $M \to \infty$ and $\mathcal{F}_{KL} \subset KL(\Pi)$. 

6
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
\[
\tilde{\Pi}\{ (f_m, \beta) \in U^c | \mathcal{D}_n \} \to 0 \text{ a.s. under the true data generating distribution, } P_{f_0, \beta_0} \text{ that generates data } \mathcal{D}_n.
\]

Proof of Theorem 1

Suppose for any two densities \(g_1\) and \(g_2\), \(K(g_1, g_2) = \int_{\mathbb{R}} g_1(w) \log \{g_1(w)/g_2(w)\} dw\) and \(V(g_1, g_2) = \int_{\mathbb{R}} g_1(w) \log_+ \{g_1(w)/g_2(w)\}^2 dw\), where \(\log_+(u) = \max(\log(u), 0)\). Set \(K_i(f, \beta) = K(f_0, f_{\beta_i})\) and \(V_i(f, \beta) = V(f_0, f_{\beta_i})\). The proof of Theorem 1 follows from (Pati and Dunson, 2014) and (Tang et al., 2015) with minor changes under the condition that there exist test functions \(\{\Phi_i\}_{n=1}^\infty\), sets \(\Theta_n = \mathcal{W}_i(f_0) \times \Theta_{\beta_n}, n \geq 1\), and constants \(C_1, C_2, c_1, c_2, c > 0\) such that

1. \(\sum_{n=1}^\infty E_{\Pi_{n=1}^n f_0} \Phi_n < \infty\)
2. \(\sup_{(f, \beta) \in U^c \cap \Theta_n} E_{\Pi_{n=1}^n f_0} (1 - \Phi_n) \leq C_1 e^{-c_1 n}\)
3. \(\Pi_\beta(\Theta^c_{\beta_n}) \leq C_2 e^{-c_2 T_n^2}\)
4. For all \(\delta > 0\) and for almost every data sequence \(\{y_i, x_i\}_{i=1}^\infty\),
\[
\tilde{\Pi}\{ (f, \beta) : K_i(f, \beta) < \delta \forall i, \sum_{i=1}^\infty \frac{V_i(f, \beta)}{\epsilon^2} < \infty \} > 0
\]

To verify the above conditions, we construct sieves \(\Theta_n = \mathcal{W}_i(f_0) \times \Theta_{\beta_n}, \text{ where } \Theta_{\beta_n} = \{\beta : \|\beta\| < T_n\}\) for sequences \(T_n\) to be chosen later.

Our claim is: \(\Pi_\beta(\Theta^c_{\beta_n}) \leq \frac{4p}{\sqrt{2\pi} T_n} e^{-T_n^2/2}\), \(\log N(\epsilon, \Theta_{\beta_n}, \|\cdot\|) \sim o(n)\)
\(\Theta_{\beta_n} = \{\beta : \|\beta\| < T_n\}\), then \(\log N(\epsilon, \Theta_{\beta_n}, \|\cdot\|) \leq \log\left(\frac{2T_n}{\epsilon}\right)\). Since \(\beta\) follows Gaussian distribution, \(\Pi_\beta(\Theta^c_{\beta_n}) \leq \frac{4p}{\sqrt{2\pi} T_n} e^{-T_n^2/2}\). Choosing \(T_n = O(\sqrt{n})\), \(\Pi_\beta(\Theta^c_{\beta_n})\) can be made exponentially small and \(\log N(\epsilon, \Theta_{\beta_n}, \|\cdot\|) \sim o(n)\). Hence condition (3) is satisfied.

Condition 1 & 2:

We write \(U = \mathcal{W}_{1n} \cup \mathcal{W}_{2n}\), where \(\mathcal{W}_{1n} = \mathcal{W}_i(f_0) \times \{\beta : \|\beta - \beta_0\| < \delta\}\) and \(\mathcal{W}_{2n} = (f, \beta) : \|\beta - \beta_0\| > \delta\). First, we prove the existence of exponentially consistent sequence of tests for \(H_0 : (f, \beta) = (f_0, \beta_0)\) against \(H_0 : (f, \beta) \in \mathcal{W}_{1n} \cap \Theta_n\) Without loss of generality,

\[
\mathcal{W}_i(f_0) = \{f : \int \Phi(y)f(y)dy - \int \Phi(y)f_0(y)dy < \epsilon\}, \quad (C-3)
\]
where $0 \leq \Phi \leq 1$ and $\Phi$ is Lipschitz continuous. Hence there exists $M > 0$ such that $|\Phi(y_1) - \Phi(y_2)| < M \sum_{j=1}^{m} |y_{1j} - y_{2j}|$. Set $f_0 = f_0(y_i - x_i^T \beta_0)$ and $\tilde{\Phi}_i(y) = \Phi(y_i - x_i^T \beta_0)$. Hence, $E_{f_0} \tilde{\Phi}_i = E_{f_0} \Phi$. 

$$E_{f_0} \tilde{\Phi}_i(y) = \int \tilde{\Phi}_i f_0(y) dy$$

$$= \int \Phi(y) f(y - \{x_i^T \beta - x_i^T \beta_0\}) dy$$

$$\geq \int \Phi(y - \{x_i^T \beta - x_i^T \beta_0\}) f(y - \{x_i^T \beta - x_i^T \beta_0\}) dy$$

$$- \int |\Phi(y) - \Phi(y - \{x_i^T \beta - x_i^T \beta_0\})| f(y - \{x_i^T \beta - x_i^T \beta_0\}) dy$$

$$= \int \Phi(y - \{x_i^T \beta - x_i^T \beta_0\}) f(y - \{x_i^T \beta - x_i^T \beta_0\}) dy$$

$$- \int |\Phi(y) - \Phi(y - x_i^T \{\beta - \beta_0\})| f(y - x_i^T \{\beta - \beta_0\}) dy$$

$$\geq \int \Phi(y) f(y) dy - M \|x_i\| \|\beta - \beta_0\|$$

$$\geq \int \Phi(y) f(y) dy - M_1 \|\beta - \beta_0\|.$$ 

Hence, $E_{f_0} \tilde{\Phi}_i(y) \geq E_{f_0} \Phi(y) + \epsilon - M_1 \delta$ for any $f \in U^c$ and for some constant $M_1 > 0$ depending on $M$. We complete the proof by choosing the $\delta$ sufficiently small and applying the Proposition 3.1 of Amewou-Atisso et al. (2003).

The proof of the existence of consistent sequence of tests for $H_0 : (f, \beta) = (f_0, \beta_0)$ against $H_0 : (f, \beta) \in W_{2n} \cap \Theta_n$ follows from the Proposition 3.1 of Amewou-Atisso et al. (2003) (when applied to $\|\beta - \beta_0\| > \delta$).

**Condition 4:**

Note that $V(f_0, f^{(M)}) \leq \int |\log f^{(M)} - \log f_0|^2 f_0$ is finite as $|\log f^{(M)}|^2 |\log f^{(M)}|$ are bounded by integrable functions. Hence condition 4 trivially follows from the Lemma1
Web Appendix D: Additional simulation studies

Simulation 2

For this simulation study, we compare the performance of the regression estimates based on our SMS and PGM models with the regression estimates based on the competing parametric skew-t (ST) model. Unlike the simulation study-1, we compare these competing methods when the true error densities do not follow the SMS model assumption of equation (2). For this purpose, we simulate \( N = 100 \) replications of data sets of bivariate responses for sample size \( n = 200 \) and another 100 replications for \( n = 400 \). For each replication of dataset, we first generate

\[
\mu_{ij} = \beta_{0j} + \beta_{1j}x_{ij1} + \beta_{2j}x_{ij2} \quad \text{for} \quad i = 1, \ldots, n \quad \text{and} \quad j = 1, 2 \quad \text{with true parameters} \quad (\beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}) = (0.5, 1, 0.5, 1)
\]

and random covariates \( x_{ij1} \sim \text{Ber}(0.5) \) and \( x_{ij2} \sim \text{Half-Normal}(0, 1) \). Then we simulate the bivariate response \( Y_i = \left( Y_{i1}, Y_{i2} \right) \) with marginal densities \( Y_{i1} \sim N(\mu_{i1}, 1) \) and \( Y_{i2} \sim \text{Exp}\left(\frac{1}{\mu_{i2}}\right) \) using inverse probability transformation while incorporating the within-pair association via a Gaussian copula with correlation \( \rho = 0.7 \). The resulting marginal regression model is

\[
Y_{ij} = \beta_{0j} + x_{ij1}\beta_{1j} + x_{ij2}\beta_{2j} + \epsilon_{ij} \quad \text{for} \quad i = 1, \ldots, n \quad \text{and} \quad j = 1, 2.
\]

The comparison of three competing methods of estimation are based on the average (based on \( N = 100 \) replications) mean-squared error (MSE), estimated standard error (SE), percentage of relative bias (RB%), and approximate coverage probability (CP) of the 95% interval estimates of the regression parameters \((\beta_{11}, \beta_{21}; \beta_{12}, \beta_{22})\). The summary of the results is given in Table T1.

For each method, the MSE, SE as well as the RB of the parameter estimates for \( n = 400 \) are lower than corresponding measures for the estimates with sample-size \( n = 50, 200 \). Overall, the SMS method yields better CP for each 95% interval-estimate, compared to the parameteric methods– PGM and ST. This is not surprising because the estimates obtained from methods using wrong parametric assumptions are expected to be less robust than the estimates obtained from semiparametric methods. However, it is important to note that even the estimates of the regression parameters of \( Y_1 \) from parametric joint models are performing worse than the corresponding estimates from semiparametric joint models, even when the parametric assumptions are only erroneous for \( Y_2 \).

Under this simulation setting, the true distributional form of the component \( Y_2 \) is not from the SMS class in equations (2)-(4) of the paper. So, we report the summary results only for \( \alpha_1 \) in the last rows of Table T1. Under this simulation model, the true value of the skewness parameter
Table T1: For sample sizes $n = 200$ and $400$, the simulation results to compare the approximate (based on 100 replications) Mean Squared Error (MSE), average Standard Error (SE), percentage of Relative Bias (RB%), and Coverage Probability (CP) of 95% interval estimates from three competing methods: SMS, PGM and ST. The datasets are simulated using Gaussian copula with association $\rho_{12} = 0.7$, true parameter values $(\beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}) = (0.5,1,0.5,1)$, and univariate Gaussian as the marginal density of $Y_1$ and the exponential density as the marginal density of $Y_2$.

|          | SMS    |          |          |          | PGM    |          |          |          | ST     |          |          |          |          |          |          |          |          |          |          |          |          |          |
|----------|--------|----------|----------|----------|--------|----------|----------|----------|--------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|          |
|          | Mean   | MSE      | SE       | RB%     | CP     | Mean     | MSE      | SE       | RB%     | CP     | Mean     | MSE      | SE       | RB%     | CP     | Mean     | MSE      | SE       | RB%     | CP     | Mean     | MSE      | SE       | RB%     | CP     |
| $n=50$   |        |          |          |         |        |          |          |          |         |        |          |          |          |         |        |          |          |          |         |        |          |          |          |         |        |         |
| $\beta_{11}$ | 0.5228 | 0.0418   | 0.2221   | 2.1     | 0.96   | 0.5106   | 0.0372   | 0.2246   | 2.1     | 0.98   | 0.5114   | 0.0464   | 0.2273   | 2.3     | 0.96   |
| $\beta_{21}$ | 1.0519 | 0.0362   | 0.1963   | 4.8     | 0.95   | 1.0172   | 0.0283   | 0.1932   | 1.7     | 0.98   | 1.0023   | 0.0324   | 0.1935   | 2.1     | 0.95   |
| $\beta_{12}$ | 0.4070 | 0.1083   | 0.3087   | 9.6     | 0.91   | 0.4349   | 0.0924   | 0.3372   | 12.6    | 0.94   | 0.4342   | 0.0894   | 0.3233   | 13.2    | 0.94   |
| $\beta_{22}$ | 0.9247 | 0.0963   | 0.2808   | 7.5     | 0.94   | 0.8616   | 0.0943   | 0.2924   | 13.8    | 0.93   | 0.8574   | 0.0933   | 0.2864   | 14.3    | 0.94   |
| $\alpha_1$  | 0.2507 | 0.1899   | 0.5162   | –       | 0.97   | 0.4398   | 0.2863   | 0.5604   | –       | 0.70   | 0.5029   | 0.3932   | 0.5322   | –       | 0.83   |
| $n=200$   |        |          |          |         |        |          |          |          |         |        |          |          |          |         |        |          |          |          |         |        |          |          |          |         |        |         |
| $\beta_{11}$ | 0.5185 | 0.0105   | 0.1050   | 3.7     | 0.94   | 0.5191   | 0.0103   | 0.1075   | 3.8     | 0.97   | 0.5182   | 0.0111   | 0.1074   | 3.6     | 0.97   |
| $\beta_{21}$ | 1.0097 | 0.0069   | 0.0892   | < 1     | 0.96   | 1.0012   | 0.0064   | 0.0900   | < 1     | 0.96   | 0.9996   | 0.0065   | 0.0892   | < 1     | 0.98   |
| $\beta_{12}$ | 0.4255 | 0.0392   | 0.1505   | 15      | 0.91   | 0.4360   | 0.0274   | 0.1580   | 13      | 0.97   | 0.4383   | 0.0284   | 0.1576   | 13      | 0.91   |
| $\beta_{22}$ | 0.8820 | 0.0393   | 0.1317   | 14      | 0.96   | 0.8850   | 0.0350   | 0.1344   | 12      | 0.81   | 0.8919   | 0.0354   | 0.1360   | 11      | 0.79   |
| $\alpha_1$  | 0.1921 | 0.1250   | 0.3745   | –       | 0.93   | 0.4111   | 0.2119   | 0.3699   | –       | 0.84   | 0.4905   | 0.3115   | 0.3309   | –       | 0.84   |
| $n=400$   |        |          |          |         |        |          |          |          |         |        |          |          |          |         |        |          |          |          |         |        |          |          |          |         |        |         |
| $\beta_{11}$ | 0.4990 | 0.0061   | 0.0752   | < 1     | 0.97   | 0.5001   | 0.0059   | 0.0755   | < 1     | 0.96   | 0.5012   | 0.0060   | 0.0758   | < 1     | 0.95   |
| $\beta_{21}$ | 0.9945 | 0.0041   | 0.0634   | < 1     | 0.92   | 0.9919   | 0.0052   | 0.0635   | < 1     | 0.91   | 0.9935   | 0.0051   | 0.0633   | < 1     | 0.78   |
| $\beta_{12}$ | 0.4328 | 0.0208   | 0.1063   | 12      | 0.91   | 0.4385   | 0.0186   | 0.1132   | 12      | 0.90   | 0.4398   | 0.0176   | 0.1118   | 12      | 0.90   |
| $\beta_{22}$ | 0.8518 | 0.0323   | 0.0924   | 14      | 0.65   | 0.8604   | 0.0318   | 0.0957   | 14      | 0.65   | 0.8639   | 0.0298   | 0.0968   | 14      | 0.66   |
| $\alpha_1$  | 0.1945 | 0.1156   | 0.3015   | –       | 0.92   | 0.4523   | 0.2356   | 0.2956   | –       | 0.70   | 0.5173   | 0.3248   | 0.2603   | –       | 0.65   |
\(\alpha_1\) corresponding to the component \(Y_1\) is 0 (with Gaussian marginal density). Under the SMS model, the average (based on 100 replications) estimated value of \(\alpha_1\) (0.19 for \(n = 200, 400\) and 0.25 for \(n = 50\)) is substantially smaller than the average estimates from the methods based on parametric models PGM and ST. Even the approximate CP (0.93 and 0.92 respectively for \(n = 200\) and \(n = 400\) and 0.97 for \(n = 50\)) of the interval estimates of \(\alpha_1\) are better than corresponding CP of interval estimates from parametric methods. Moreover, these coverage probabilities from parametric methods seem to get substantially worse for increasing the sample size \(n\)!

Under this simulation model, the marginal density of the response component \(Y_2\) is exponential with Pearsonian skewness = 2. Under the SMS model, the average estimated value for the Pearsonian skewness parameter for the residuals is 1.88, 1.93 and 1.98 respectively for \(n = 50, n = 200\) and \(n = 400\). This suggests that our SMS model captures the true Pearsonian skewness of the response even when the true marginal density is different from that of the SMS model. This is a valuable practical advantage for analysis of many real life biomedical studies, when the goal is to estimate the covariate effects as well as to account for the level of skewness.

**Simulation 3**

The goal of this small scale simulation study is to compare the performance of the regression estimates obtained from our SMS and PGM models to the regression estimates obtained from the generalized estimating equations (GEE), the popular analysis tool for repeated measures. We want to make this comparison even when the true distribution of the errors is not a member of the class of SMS models in equations (2)-(4). For this purpose, we simulate \(N = 100\) replications of datasets of sample-size \(n = 50, 200\) from a bivariate repeated measures model with true marginal regression model \(Y_{ij} = \beta_0 + x_{ij1}\beta_1 + x_{ij2}\beta_2 + \varepsilon_{ij}\) for \(i = 1, \ldots, n\) and \(j = 1, 2\) (same regression parameters \((\beta_1, \beta_2)\) for both components \(Y_{i1}, Y_{i2}\)). The bivariate joint density of error \((\varepsilon_{i1}, \varepsilon_{i2})\) is same as in the Simulation 2 with Gaussian marginal density for \(Y_1\) and exponential marginal density for \(Y_2\). The summary of the results is given in Table T2.

Given that the estimates of the skewness parameters \(\alpha = (\alpha_1, \alpha_2)\) are not estimable under GEE, we do not report them in the summary. Also, the intercepts of SMS and PGM models are not comparable to intercept \(\beta_0\) under GEE. Hence, following Simulation 2, we report the MSE, SE, RB\% and CP summaries (based on 100 replicates) for only the regression parameters \(\beta_1\) and \(\beta_2\).
We observe that the GEE produces slightly biased regression estimates compared to the estimates obtained using SMS and PGM models. Compared to the GEE estimates, the SMS and PGM models yield smaller MSE (at least 30% reduction), average SE (at least 20% reduction) as well as smaller RB% even when the sample size is as small as \( n = 50 \). Thus, even when the true error model is not same as the assumptions of the SMS model, Bayesian estimates of the regression parameters derived from our semiparametric SMS model are comparable to the estimates from parametric model, and more precise than those obtained via the GEE method.
Table T2: Based on 100 replicates of the data-sets of sample size $n = 50, 200$, the simulation results comparing the approximate Mean Squared Error (MSE), Standard Error (SE), Relative Bias (RB), and Coverage Probability (CP) of regression estimates from three competing methods (SMS, parametric PGM and GEE). The true simulation model of bivariate repeated measures is Gaussian copula with association parameter $\rho_{12} = 0.7$, true regression parameters $\left(\beta_1, \beta_2\right) = (0.5, 1)$, and with Gaussian and exponential marginal densities.

<table>
<thead>
<tr>
<th></th>
<th>SMS</th>
<th>PGM</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>MSE</td>
<td>SE</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.4817</td>
<td>0.0307</td>
<td>0.1855</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.9646</td>
<td>0.0229</td>
<td>0.1594</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SMS</th>
<th>PGM</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>MSE</td>
<td>SE</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.5015</td>
<td>0.0096</td>
<td>0.0889</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.9912</td>
<td>0.0069</td>
<td>0.0747</td>
</tr>
</tbody>
</table>
Web Appendix E: Computer codes

R and JAGS codes for fitting the SMS and PGM models to simulated data are available in the GitHub link:

https://github.com/bandyopd/Multivariate-Skewed-models
References


