Augmented mixed models for clustered proportion data

Dipankar Bandyopadhyay[1]*, Diana M. Galvis[2], Victor H. Lachos[2]

[1]Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455

[2]Departamento de Estatística, Universidade Estadual de Campinas, Campinas, SP, Brasil

Abstract

Often in biomedical research, we deal with continuous (clustered) proportion responses ranging between zero and one quantifying the disease status of the cluster units. Interestingly, the study population might also consist of relatively disease-free as well as highly diseased subjects, contributing to proportion values in the interval [0, 1]. Regression on a variety of parametric densities with support lying in $(0, 1)$, such as beta regression, can assess important covariate effects. However, they are deemed inappropriate due the presence of zeros and/or ones. To evade this, we introduce a class of general proportion density (GPD), and further augment the probabilities of zero and one to this GPD, controlling for the clustering. Our approach is Bayesian, and presents a computationally convenient framework amenable to available freeware. Bayesian case-deletion influence diagnostics based on q-divergence measures are automatic from the MCMC output. The methodology is illustrated using both simulation studies and application to a real dataset from a clinical periodontology study.

Keywords: Augment; Bayesian; Dispersion models; Kullback-Leibler divergence; Proportion data; Periodontal disease.

*Address for correspondence: Division of Biostatistics, University of Minnesota SPH, A460 Mayo MMC 303, 420 Delaware Street SE, Minneapolis, MN 55455. E-mail: dbandyop@umn.edu

1 Introduction

Continuous proportion data (expressed as percentages, proportions, and rates), such as the percent decrease in glomerular filtration rate at various follow-up times since baseline[1,2] are routinely analyzed in medicine and public health. Because the responses are confined in the open interval $(0, 1)$, one might be tempted to use the logistic-normal model[3] with Gaussian assumptions for logit-transformed proportion responses. However, covariate effects interpretation are not straightforward because the logit link is no longer preserved for the expected value of the response. Alternatively, to tackle this, the beta[4,5], beta rectangular (BRe)[6] and simplex[7] distributions (all with common support within the open unit interval), and their corresponding regressions were proposed under a generalized linear model (GLM) framework.

The flexible beta density[8] can represent a variety of shapes, accounting for uncorrectable non-normality and skewness[9] in the context of bounded proportion data. The beta regression (BR) reparameterizes the associated beta parameters, connecting the response to the data covariates through suitable link functions[5]. Yet, the beta density does not accommodate tail-area events, or flexibility in variance specifications[10]. To accommodate this, the BRe density[6], and associated regression models[10] were considered under a Bayesian framework. Note, the BRe regression includes the (constant dispersion) BR[5], and the variable dispersion BR[9] as special cases. The simplex regression[1] is based on the simplex distribution from the dispersion family[11], assumes constant dispersion, and uses extended generalized estimating equations for inference connecting the mean to the covariates via the logit link. Subsequently, frameworks with heterogenous dispersion[12], and for mixed-effects models[13] were explored. Yet, their potential were limited to proportion responses with support in $(0, 1)$.

A clinical study on periodontal disease (PrD) conducted at the Medical University of South Carolina (MUSC)[14] motivates our work. The clinical attachment level (CAL), a clinical marker of PrD was measured at each of the 6 sites of a subject's tooth, and we were interested to assess covariate-response relationships on 'tooth-type specific (such as incisors, canines, pre-molars and molars) proportion of diseased sites' to determine the status of PrD. Figure 1 (left panel) plots the raw (unadjusted) density histogram of the proportion responses, packed over all subjects and tooth-types. The responses are in the closed interval $[0, 1]$ where 0 and 1 represent 'completely disease free', and 'highly diseased' cases, respectively. For a simple parametric treatment to this data, one might be tempted to use one of the three distributions mentioned above after possible transformation[9] of the response from $[0, 1]$ to the interval $(0, 1)$. These ad hoc re-scalings might work out for small small proportions of 0's and 1's, but the sensitivity on parameter estimates can be considerable as the proportions increase. Transformations, in general, are not universal. In addition, presence of clustering (tooth-sites within mouth) brings in an extra level of heterogeneity, and these transformations which are usually applied component-wise may not guarantee a tractable (multivariate) joint distribution[15]. At this stage, we desire an appropriate theoretical model capable of handling all these challenges, yet avoiding data transformations.

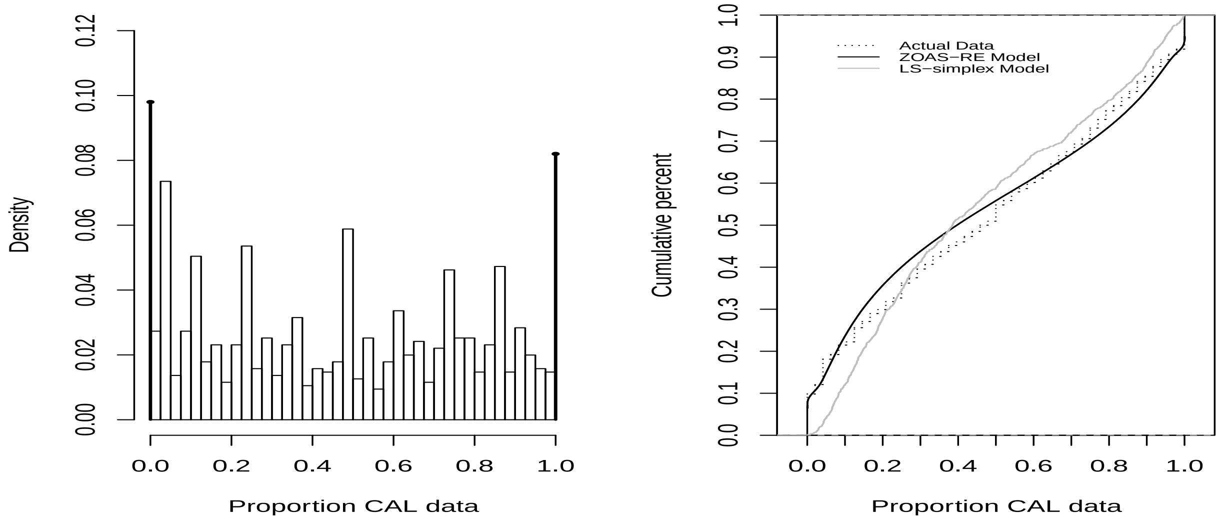

Figure 1: Periodontal proportion data. The (raw) density histogram combining subjects and tooth-types are presented in the left panel. The empirical cumulative distribution function of the real data, and that obtained after fitting the ZOAS-RE and the LS-simplex models appear in the right panel.

Note that the beta, BRe and simplex densities (and their regressions) present a noticeable analytic difference in their probability density function (pdf) specification. Motivated by these differences and the flexibility they provide, we seek to combine them into a new (parametric) class of density called the general proportion density (GPD), where these three popular models appear as particular cases. In this context, our paper generalizes the recent augmented beta proposition[16]. Next, we extend this GPD to a regression setup for independent responses in $(0, 1)$. Finally, for a unified (regression) framework for clustered responses in $[0, 1]$, we propose a generalized linear mixed model (GLMM) framework by augmenting the probabilities of occurrence of zeros, ones or both to the standard GPD regression model via an augmented GPD random effects (AugGPD-RE) model. Our inferential framework is Bayesian, and can be easily handled using freeware like `OpenBUGS`. Furthermore, case-deletion and local influence diagnostics[17] to assess outlier effects are immediate from the Markov chain Monte Carlo (MCMC) output.

The rest of the article is organized as follows. Section 2 formulates the GPD and the augmented GPD class of density as well as some useful statistical properties. Section 3 develops the Bayesian estimation framework for the AugGPD-RE regression model and related diagnostics. Application to the motivating PD data appear in Section 4. Section 5 presents simulation studies to compare finite-sample performance of parameter estimates among the GPD class members, and also under model misspecification. Finally, some concluding statements appear in Section 6.

2 General proportion density

We start with the definition of proportion density (PD) models, and then proceed to establish the GPD density class.

Definition 1. *A random variable (rv) ξ with support in the unit interval $(0, 1)$ belongs to* *the class of PD with parameters λ and ϕ if it can be expressed as*

$$g_1(\xi; \lambda, \phi) = a_1(\lambda, \phi)a_2(\xi, \phi) \exp\{-\phi a_3(\xi, \lambda)\}, \quad \phi > 0, \ \lambda \in (0, 1), \tag{2.1}$$

where $E[\xi] = \lambda$, and $a_s(\cdot, \cdot)$, $s = 1, 2, 3$ are real-valued functions with $a_1, a_2 \geq 0$, and a_3 taking value on the real line. We use the notation $\xi \sim PD(\lambda, \phi)$ to represent ξ a member of the PD class defined in (2.1). Following[11]*, if $a_3(\xi, \lambda)$ in (2.1) is continuous and twice differentiable function with respect to ξ and λ and is non-zero, the variance function is* $V(\lambda) = -\left(\dfrac{\partial^2 a_3(\xi, \lambda)}{\partial\lambda\partial x}\right)^{-1}\Bigg|_{\xi=\lambda}$.

Next, consider the density of the rv X following the 2-component mixture $X = \eta U + (1-\eta)\xi$, where $\eta \in [0, 1]$ is a mixture parameter and U a Uniform$(0, 1)$ rv distributed independently of ξ with pdf in (2.1). Then, X follows the general proportion density (GPD), i.e., $X \sim$ GPD(η, λ, ϕ) with the pdf given by

$$g(X; \eta, \lambda, \phi) = \eta + (1 - \eta)g_1(X; \lambda, \phi), \tag{2.2}$$

where g_1 is as defined in (2.1). Note that for $\eta = 1$, the GPD reduces to the uniform distribution, and for $\eta = 0$ we retrieve the PD class of distributions. The mean and variance of X are $\mu = E[X] = \eta/2 + (1 - \eta)E[\xi]$, $\sigma^2 = \text{Var}(X) = \frac{\eta}{12} + (1 - \eta)^2\text{Var}(\xi)$, respectively.

2.1 Densities in the GPD class

The GPD class includes the beta, simplex, and the BRe densities with support in the interval (0,1), and can be used to model proportion data. These are described in the propositions below with their respective pdf's presented in Appendix A.

Proposition 1. *The beta density*[5] *reparametrized in terms of μ (the mean) and of ϕ (the precision parameter) belongs to the GPD class of distributions with its variance function given by $V(\mu) = \mu(1 - \mu)$.*

Proof. In (2.2), consider $\eta = 0$, $\lambda = \mu$ and $g_1(x;\mu,\phi) = \dfrac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}$, such that $a_1(\mu,\phi) = \dfrac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}$, $a_2(x,\phi) = \left(x(1-x)^{1-\phi}\right)^{-1}$ and $a_3(x,\mu) = \mu\log\frac{1-x}{x}$. Then, the variance function[5] (from Definition 1) is

$$V(\mu) = -\left(\frac{\partial^2 a_3(x,\mu)}{\partial\mu\partial x}\right)^{-1}\Bigg|_{x=\mu} = -\left(\frac{-1}{x(1-x)}\right)^{-1}\Bigg|_{x=\mu} = \mu(1-\mu)$$

. □

Proposition 2. *The simplex distribution[7] with parameters μ and ϕ belongs to the GPD class with the variance function given by $V(\mu) = \mu^3(1-\mu)^3$.*

Proof. In (2.2), consider $\eta = 0$, $\lambda = \mu$ and $g_1(x;\mu,\phi) = \dfrac{\sqrt{\phi}}{\sqrt{2\pi}\,(x(1-x))^{3/2}}\times$ $\exp\left\{-\phi\dfrac{(x-\mu)^2}{2x(1-x)\mu^2(1-\mu)^2}\right\}$, such that $a_1(\mu,\phi) = 1$, $a_2(x,\phi) = \dfrac{\phi}{\sqrt{2\pi}\,(x(1-x))^{3/2}}$ and $a_3(x,\mu) = \dfrac{(x-\mu)^2}{2x(1-x)\mu^2(1-\mu)^2}$. Then, the variance function[11] is given by $V(\mu) = -\left(\frac{\partial^2 a_3(x,\mu)}{\partial\mu\partial x}\right)^{-1}\Big|_{x=\mu} = -\left(\frac{-1}{x^3(1-x)^3}\right)^{-1}\Big|_{x=\mu} = \mu^3(1-\mu)^3.$ □

Proposition 3. *The BRe density[6] with parameters η, λ and ϕ belongs to the GPD class of distributions.*

Proof. The proof follows from (2.2), considering $\eta > 0$ and $g_1(x;\lambda,\phi)$ as in Proposition 1, replacing μ by λ. However, the BRe density is a mixture of a uniform and a beta density (see Appendix A in the supplementary material) and a closed form expression of the variance function is not available. □

A major shortcoming of these densities is that they are not appropriate for modeling datasets containing proportion responses at the extremes (i.e., 0, or 1, or both). We seek to address this via an augmented GPD framework defined as follows:

Definition 2. *The pdf of a rv Y with support in the interval $[0,1]$ belongs to the augmented GPD class if it has the form*

$$f(y;\eta,\lambda,\phi,p_0,p_1) = p_0 I_{\{y=0\}} + p_1 I_{\{y=1\}} + (1-p_0-p_1) g(y;\eta,\lambda,\phi) I_{\{y\in(0,1)\}}, \qquad (2.3)$$

where $I_{\{A\}}$ is the indicator function of the set A; $g(.)$ is as defined in Equation (2.2) and p_0, $p_1 \geq 0$, with $p_0 + p_1 < 1$.

From (2.3), the expectation and variance of Y are, respectively, $E[Y] = p_1 + (1-p_0-p_1)\mu = \delta$ and $\text{Var}(Y) = p_1(1-p_1) + (1-p_0-p_1)[\sigma^2 - 2p_1\mu + (p_0+p_1)\mu^2]$, where μ and σ^2 are as in Definition 1. Note, the augmented GPD class defined in (2) reduces to the GPD class when p_0 and p_1 are simultaneously equals to zero. When $p_0 > 0$ and $p_1 = 0$ we have the zero augmented GPD class, and for $p_0 = 0$ and $p_1 > 0$ we have the one augmented GPD class. Finally, when $p_0 > 0$ and $p_1 > 0$, we have the more general zero-one augmented GPD class. Motivated by the PrD data, we are particularly interested in the following three subfamilies of the augmented GPD class, corresponding to the densities specified in Subsection 2.1

- Zero-one augmented beta (ZOAB) density, if $\eta = 0$ and $g_1(.)$ the beta density
- Zero-one augmented simplex (ZOAS) density, if $\eta = 0$ and $g_1(.)$ the simplex density
- Zero-one augmented beta rectangular (ZOABRe) density, if $\eta > 0$ and $g_1(.)$ the beta density

3 Model development and Bayesian inference

3.1 GPD regression model

Let $Y_1, \ldots, Y_n$ be n independent rv's such that $Y_i \sim \text{GPD}(\eta_i, \lambda_i, \phi_i)$. Consider that $\mu_i = \eta_i/2 + (1-\eta_i)\lambda$ is directly modeled through covariates as $h_1(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$ where h_1 is a adequate link function with counterdomain the real line, $\boldsymbol{\beta}$ is the vector of regression parameters with the first element of $\mathbf{x}_i$ being 1. However, μ_i is a function of the mixture parameter η_i and λ, which leads to a restricted parametric space of η_i, defined as $0 < \eta_i < |2\mu_i - 1|$ that is dependent on μ_i. Hence, for a more appropriate regression framework that connects Y to covariates, we work with the reparameterization proposed in[10], and define $\alpha_i \in [0, 1]$ such that $\alpha_i = \dfrac{\eta_i}{1 - (1 - \eta_i)|2\lambda_i - 1|}$. Henceforth, the GPD class is parameterized in terms of μ_i, α_i and ϕ_i.

The parameters ϕ_i and α_i can be assumed constants, or regressed onto covariates through convenient link functions. For μ_i and α_i, link functions such as, logit, probit or complementary log-log can be used. Finally, for ϕ_i, the log, square-root, or identity link functions can be considered. Parameter estimation can follow either the (classical) maximum likelihood (ML), or the Bayesian route through MCMC methods.

3.2 Augmented GPD random effects model

The augmented GPD model described in (2.3) is only appropriate for independent responses in $(0, 1)$. To accommodate clustering (as in our case) or longitudinal subject-specific profiles, we proceed with the augmented GPD random effects (henceforth, AugGPD-RE) model. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be n independent continuous random vectors, where $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ is the vector of length n_i for the sample unit i, with the components $y_{ij} \in \zeta$, where ζ is an element of the set $\{[0, 1), (0, 1], [0, 1]\}$. Thus, under the AugGPD-RE model, the parameters μ_{ij}, p_{0ij} and p_{1ij} can be connected with covariates through suitable link functions as

$$\begin{aligned} h_1(\mu_{ij}) &= \mathbf{X}_{\mu_{ij}}^\top \boldsymbol{\beta} + \mathbf{X}_{b_i}^\top \mathbf{b}_i, && (3.1)\\ h_2(p_{0ij}) &= \mathbf{X}_{0ij}^\top \boldsymbol{\psi}, && (3.2)\\ h_3(p_{1ij}) &= \mathbf{X}_{1ij}^\top \boldsymbol{\rho}, && (3.3) \end{aligned}$$

where $\mathbf{X}_{\mu_{ij}}$, $\mathbf{X}_{0ij}$ and $\mathbf{X}_{1ij}$ correspond to the j-th column from the design matrices $\mathbf{X}_{\mu_i}$, $\mathbf{X}_{p_{0i}}$ and $\mathbf{X}_{p_{1i}}$ of dimension $p \times n_i$, $r \times n_i$ and $s \times n_i$, related with the i-th unit sample, corresponding to the vectors of fixed effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_r)^\top$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_s)^\top$, respectively, and $\mathbf{X}_{bi}$ is the design matrix of dimension $q \times n_i$ corresponding to REs vector $\mathbf{b}_i = (b_{i1}, \ldots, b_{iq})^\top$. Choice of link functions for h_1, h_2 and h_3 remain the same as for μ_i and α_i in Subsection 3.1. For purpose of interpretation, we focus on the logit link. In this work, we consider ϕ and α as constants despite those parameters can also be regressed onto covariates through suitable link functions. Also, to avoid over-parameterization, the probabilities p_{0ij} and p_{1ij} are free of REs, however, both could be considered considered constants across subjects. Finally, we denote our AugGPD-RE model as $Y_{ij} \sim \text{AugGPD-RE}(p_{0ij}, p_{1ij}, \mu_{ij}, \alpha, \phi)$ $i = 1, \ldots, n,\ j = 1, \ldots, n_i$.

Let $\boldsymbol{\mathcal{D}} = (\mathbf{X}_\mu, \mathbf{X}_{p_0}, \mathbf{X}_{p_1}, \mathbf{X}_b, \mathbf{y})^\top$ be the full observed data and $\boldsymbol{\Omega} = (\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\rho}, \phi, \alpha)^\top$ be the parameter vector in the AugGPD-RE model. The joint data likelihood, conditional on the random-effects $\mathbf{b}_i$, $L(\boldsymbol{\Omega}; \boldsymbol{\mathcal{D}}, \mathbf{b})$ is given by:

$$L(\boldsymbol{\Omega}; \mathbf{b}, \boldsymbol{\mathcal{D}}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} p_{0ij}^{\,I_{y_{ij}=0}}\, p_{1ij}^{\,I_{y_{ij}=1}} \left[(1 - p_{0ij} - p_{0ij}) g(y_{ij}; \alpha, \mu_{ij}, \phi)\right]^{I_{y_{ij} \in (0,1)}}, \tag{3.4}$$

where $p_{0ij} = \text{logit}^{-1}(\mathbf{X}_{0_{ij}}^{\top}\boldsymbol{\psi})$, $p_{1ij} = \text{logit}^{-1}(\mathbf{X}_{1_{ij}}^{\top}\boldsymbol{\rho})$, I is an indicator function, and g is given by

$$g(y_{ij}; \alpha, \mu_{ij}, \phi) = \eta_{ij} + (1 - \eta_{ij}) a_1(\lambda_{ij}, \phi) a_2(y_{ij}, \phi) \exp\left\{-\phi a_3(y_{ij}, \lambda_{ij})\right\}, \tag{3.5}$$

with $\eta_{ij} = \alpha(1 - 2|\mu_{ij} - \frac{1}{2}|)$, $\lambda_{ij} = \dfrac{\mu_{ij} - \frac{\eta_{ij}}{2}}{1 - \eta_{ij}}$ and $\mu_{ij} = \text{logit}^{-1}\left(\mathbf{X}_\mu ij^\top \boldsymbol{\beta} + \mathbf{X}_b ij^\top \mathbf{b}_i\right)$.

Although ML estimation of $\boldsymbol{\Omega}$ is certainly feasible using standard softwares such as (`SAS`, `R`, etc), we seek a Bayesian treatment here. The Bayesian approach accommodates full parameter uncertainty through appropriate choice of priors choices, proper sensitivity investigations, and provides direct probability statement about a parameter through credible intervals (C.I.)[18]. Next, we investigate the choice of priors on our model parameters to conduct Bayesian inference.

3.3 Priors, hyperpriors and posterior distributions

In order to complete the Bayesian specification, we need to consider prior distributions for all the unknown model parameters. In particular, we specify practical weakly informative prior opinion on the fixed effects regression parameters $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, $\boldsymbol{\rho}$, ϕ (dispersion parameter), α, and the random effects $\mathbf{b}_i$. In general, for the regression components, we can assume $\boldsymbol{\beta} \sim \text{Normal}_p(\mathbf{0}, \boldsymbol{\Sigma}_{\beta}^{-1})$, $\boldsymbol{\psi} \sim \text{Normal}_r(\mathbf{0}, \boldsymbol{\Sigma}_{\psi}^{-1})$, $\boldsymbol{\rho} \sim \text{Normal}_s(\mathbf{0}, \boldsymbol{\Sigma}_{\rho}^{-1})$. A Uniform$(0, 1)$ density[10] was adopted as prior for α. Prior on each element of $\mathbf{b}_i$ are $N(0, \sigma_b^2)$, where $\sigma_b \sim$ Uniform$(0, 100)$, the usual Gelman[19] specification. The prior on ϕ for the specific models in Subsection 2.1 were chosen as follows:

(i) *Beta and BRe models*: $\phi \sim$ Gamma(a, c), with small positive values of a and c ($c \ll a$).

(ii) *Simplex model*: $\phi^{-1/2} \sim$ Uniform$(0, a_1)$, with large positive value for a_1.

Assuming the elements of the parameter vector to be independent, the posterior conclusions are obtained combining the likelihood in (3.4), and the joint prior densities, given by

$$p(\boldsymbol{\Omega}, \mathbf{b}, \sigma_b | \boldsymbol{\mathcal{D}}) \propto L(\boldsymbol{\Omega}; \boldsymbol{\mathcal{D}}) \times \pi(\boldsymbol{\Omega}, \mathbf{b}, \sigma_b),$$

where $\pi(\boldsymbol{\Omega}, \mathbf{b}, \sigma_b) = \pi_0(\boldsymbol{\beta})\pi_1(\boldsymbol{\psi})\pi_2(\boldsymbol{\rho})\pi_3(\alpha)\pi_4(\phi)\pi_5(\mathbf{b}|\sigma_b)\pi_6(\sigma_b)$ and $\pi_j(.), j = 0, \ldots, 6$ denote the prior/hyperprior distributions on the model parameters as described above. The full conditional distributions necessary for the MCMC algorithm (combination of Gibbs sampling and Metropolis-within-Gibbs) in the AugGPD-RE model are as follows:

- The full conditional density for $\boldsymbol{\psi}|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\boldsymbol{\psi})}$, $\pi\left(\boldsymbol{\psi}|\mathbf{y}, \mathbf{b}, \sigma_b, \boldsymbol{\Omega}_{(-\boldsymbol{\beta})}\right)$ is proportional to

  $\exp\left\{-\frac{1}{2}(\boldsymbol{\psi} - \boldsymbol{\psi}_0)^\top \boldsymbol{\Sigma}_{\psi}^{-1}(\boldsymbol{\psi} - \boldsymbol{\psi}_0)\right\} p_{0ij}^{I_{y_{ij}=0}} (1 - p_{0ij} - p_{1ij})^{I_{y_{ij} \in (0,1)}}$.

- The full conditional density for $\boldsymbol{\rho}|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\boldsymbol{\rho})}$, $\pi\left(\boldsymbol{\rho}|\mathbf{y}, \mathbf{b}, \sigma_b, \boldsymbol{\Omega}_{(-\boldsymbol{\rho})}\right)$ is proportional to

$\exp\left\{-\frac{1}{2}(\boldsymbol{\rho}-\boldsymbol{\rho}_0)^\top \boldsymbol{\Sigma}_\rho^{-1}(\boldsymbol{\rho}-\boldsymbol{\rho}_0)\right\} p_{1ij}^{I_{y_{ij}=1}}(1-p_{0ij}-p_{1ij})^{I_{y_{ij}\in(0,1)}}$.

- The full conditional density for $\boldsymbol{\beta}|\mathbf{y},\mathbf{b},\sigma_b,\Omega_{(-\boldsymbol{\beta})}$, $\pi\left(\boldsymbol{\beta}|\mathbf{y},\mathbf{b},\sigma_b,\Omega_{(-\boldsymbol{\beta})}\right)$ is proportional to $\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\} \prod_{i=1}^{n}\prod_{j=1}^{n_i} g(y_{ij};\alpha,\mu_{ij},\phi)^{I_{y_{ij}\in(0,1)}}$, with $g(y_{ij};\alpha,\mu_{ij},\phi)$ given by (3.5).

- The full conditional density for $\phi|\mathbf{y},\mathbf{b},\sigma_b,\Omega_{(-\phi)}$, $\pi(\phi|\mathbf{y},\mathbf{b},\sigma_b,\Omega_{(-\phi)})$ is proportional to $\pi(\phi)\prod_{i=1}^{n}\prod_{j=1}^{n_i} g(y_{ij};\alpha,\mu_{ij},\phi)^{I_{y_{ij}\in(0,1)}}$.
- The full conditional density for $\alpha|\mathbf{y},\mathbf{b},\sigma_b,\Omega_{(-\alpha)}$, $\pi(\alpha|\mathbf{y},\mathbf{b},\sigma_b,\Omega_{(-\alpha)})$ is proportional to $\prod_{i=1}^{n}\prod_{j=1}^{n_i} g(y_{ij};\alpha,\mu_{ij},\phi)^{I_{y_{ij}\in(0,1)}} I_{\alpha\in[0,1]}$.
- The full conditional density for $\mathbf{b}_i|\mathbf{y},\sigma_b,\Omega$, $\pi(\mathbf{b}_i|\mathbf{y},\sigma_b,\Omega)$ is proportional to $\exp\left\{-\sum_{k=1}^{q}\frac{1}{2\sigma_b^2}b_{ik}^2\right\}\prod_{j=1}^{n_i} g(y_{ij};\alpha,\mu_{ij},\phi)^{I_{y_{ij}\in(0,1)}}$ with b_{ik} the k-th element of $\mathbf{b}_i=(b_{i1},\dots b_{iq})^\top$.
- The full conditional density for $\sigma_b|\mathbf{y},\mathbf{b},\Omega$, $\pi(\sigma_b|\mathbf{y},\mathbf{b},\Omega)$ is proportional to $\exp\left\{-\frac{1}{2\sigma_b^2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} b_{ij}^2\}\right\} I_{\sigma_b\in(0,c_1)}$.

For specific densities of the GPD class, the full conditionals for the beta, BRe and simplex models are presented in Appendix B. For computational simplicity, we avoid the multivariate prior specifications for $\boldsymbol{\beta}$, $\boldsymbol{\psi}$ and $\boldsymbol{\rho}$ (multivariate zero mean vector with inverted-Wishart covariance) and instead assign simple i.i.d Normal$(0, \text{Variance} = 100)$ priors on the elements of these vectors, which centers the 'odds-ratio' type inference at 1 with a sufficiently wide 95% interval. When p_0 and p_1 represent constant proportions for the whole data, we allocate the Dirichlet prior with hyperparameter $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\alpha_3)^\top$ for the probability vector $(p_0,p_1,1-p_0-p_1)^\top$, with $\alpha_s\sim$ Gamma$(1,0.01)$, $s=1,2,3$. After discarding the first 50000 burn-in samples, we used 50000 more samples (with a spacing of 10) from 2 independent chains with widely dispersed starting values for posterior summaries. Convergence was monitored via MCMC trace plots, autocorrelation plots and the Brooks-Gelman-Rubin $\hat{\mathrm{R}}$ statistics. Associated `R` code is available on request from the corresponding author.

3.4 Bayesian model selection and influence diagnostics

For model selection, we use the conditional predictive ordinate (CPO) and the log pseudo-marginal likelihood (LPML) statistic[20], derived from the posterior predictive distribution (ppd). Larger values of LPML indicate better fit. Computing CPO via the harmonic mean identity can lead to instability[21]. Hence, we consider a more pragmatic route and compute the CPO (and LPML) statistics using 500 non-overlapping blocks of the Markov chain, each of size 2000, post-convergence and report the expected LPML computed over the 500 blocks. In addition, we also apply the expected AIC (EAIC), expected BIC (EBIC)[20] and the DIC_3[22] criteria. The DIC_3 was used as an alternative to the usual DIC[23] because of the ease of computation directly from the MCMC output, and also due to the mixture modeling framework. All these criteria abide by the 'lower is better' law.

In addition, as a direct byproduct from the MCMC output, we develop some influence diagnostic measures to assess outlier effects on the fixed effects parameters based on case-deletion statistics[24], and the q-divergence measures[25,26] between posterior distributions. We consider three choices of these divergences, namely, the Kullback-Leibler (KL) divergence, the J-distance (symmetric version of the KL divergence), and the L_1-distance. We use the calibration method[17] to obtain the cut-off values as 0.90, 0.83 and 1.32 for the L_1, KL and J-distances, respectively.

4 Data analysis and findings

The motivating PrD dataset assessed the PrD status of Gullah-speaking African-Americans with Type-2 diabetes via a detailed questionnaire focusing on demographics, social, medical and dental history. The dataset contain measurements on 28 teeth (considered full dentition, excluding the 4 third-molars) from 290 subjects, recording proportion of diseased tooth-sites (with CAL value $\geq$ 3mm) per tooth type as the response for each subject. Hence, this clustered data framework has 4 observations (corresponding to the 4 tooth-types) for each subject. If a tooth is missing, it was considered 'missing due to PrD' where all sites for that tooth contributed to the diseased category. Subject-level covariables in the dataset include gender (0=male,1= female), age of subject at examination (in years, ranging from 26 to 87 years), glycosylated hemoglobin (HbA1c) status indicator (0=controlled,$<$ 7%; 1=uncontrolled,$\geq$ 7%) and smoking status (0=non-smoker,1=smoker). We also considered a tooth-level variable representing each of the four tooth types, with 'canine' as the baseline.

From Figure 1 (left panel), the data are continuous on [0,1], with non-negligible proportions of of 0's (114, 9.8%) and 1's (94, 8.1%). Avoiding transformation, modeling via one of the members of the GPD class might not be feasible. Hence, we proceed using the AugGPD-RE model, adjusted for subject-level clustering. From Equations (3.1), (3.2) and (3.3), we have

$$\begin{aligned}
\text{logit}(\mu_{ij}) &= \mathbf{X}_{\mu_{ij}}^{\top}\boldsymbol{\beta} + b_i, \qquad (4.1)\\
\text{logit}(p_{0_{ij}}) &= \mathbf{X}_{0_{ij}}^{\top}\boldsymbol{\psi},\\
\text{logit}(p_{1_{ij}}) &= \mathbf{X}_{1_{ij}}^{\top}\boldsymbol{\rho},
\end{aligned}$$

where $\mathbf{X}_{\mu_{ij}} = (1, \text{Gender}_{ij}, \text{Age}_{ij}, \text{HbA1c}_{ij}, \text{Smoker}_{ij}, \text{Incisor}_{ij}, \text{Premolar}_{ij}, \text{Molar}_{ij})^{\top}$, $\mathbf{X}_{\mu_{ij}} = \mathbf{X}_{0ij} = \mathbf{X}_{1ij}$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_7)^{\top}$, $\boldsymbol{\psi} = (\psi_0, \ldots, \psi_7)^{\top}$ and $\boldsymbol{\rho} = (\rho_0, \ldots, \rho_7)^{\top}$ are the vectors of regression parameters, and b_i is the subject-level random effect. The examination age was standardized (subtracting the mean and dividing by its standard deviation) to achieve better convergence. We have 6 competing models, varying with the densities in the GPD class and the regression over p_0 and p_1, as follows:

Model 1 $Y_{ij} \sim \text{ZOAS-RE}(\mu_{ij}, \phi, p_{0_{ij}}, p_{1_{ij}})$.

Model 1a $Y_{ij} \sim \text{ZOAS-RE}(\mu_{ij}, \phi, p_0, p_1)$.

Model 2 $Y_{ij} \sim \text{ZOAB-RE}(\mu_{ij}, \phi, p_{0ij}, p_{1ij})$.

Model 2a $Y_{ij} \sim \text{ZOAB-RE}(\mu_{ij}, \phi, p_0, p_1)$.

Model 3 $Y_{ij} \sim \text{ZOABRe-RE}(\alpha, \mu_{ij}, \phi, p_{0ij}, p_{1ij})$.

Model 3a $Y_{ij} \sim \text{ZOABRe-RE}(\alpha, \mu_{ij}, \phi, p_0, p_1)$.

Note that the parameter α is specific to the ZOABRe model only. In addition, we also fit the LS-simplex model (or **Model 4**) by transforming the response from y to y' via the Lemon-squeezer (LS) transformation[9] given by $y' = [y(N-1) + 1/2]/N$, where N is the number total of observations, with the regression on μ as (4.1). Although models 1, 1a, 2, 2a, 3 and 3a can be compared using standard model choice criteria described in Subsection 3.4 because they fit the same dataset, this is not the case for the LS-simplex model which fits a transformed dataset. Thus, we assess its fit visually via the empirical cumulative distribution functions (ecdfs) of the fitted values. Table 1 presents the DIC$_3$, LPML, EAIC and EBIC values for the 6 competing models. Notice that Model 1 (ZOAS-RE model) provides the best fit uniformly across all criteria. Also, the fit for models with constant p_0 and p_1 are worser than the corresponding ones with regression on p_0 and p_1. The right panel of Figure 1 clear tells us that the ecdf from the fitted values using Model 1 represent the true data much closely as compared to Model 4. Hence, we select Model 1 as our best model and proceed with inference.

Table 1: Model comparison using DIC$_3$, LPML, EAIC and EBIC criteria.

Criterion	Model					
	1	1a	2	2a	3	3a
DIC$_3$	**915.3**	1165.3	993	1243.5	1001.3	1253.4
LPML	**-461.1**	-584.1	-500.5	-623.7	-503.8	-627.8
EAIC	**917.8**	1154.9	992.7	1231	967.4	1210.4
EBIC	**1047.2**	1210.5	1124.2	1286.6	1103.9	1281.2

Plots of the means of the posterior parameter estimates and their 95% CIs for the regression onto μ (left panel), p_0 (middle panel) and p_1 (right panel) for Models 1-4 are presented

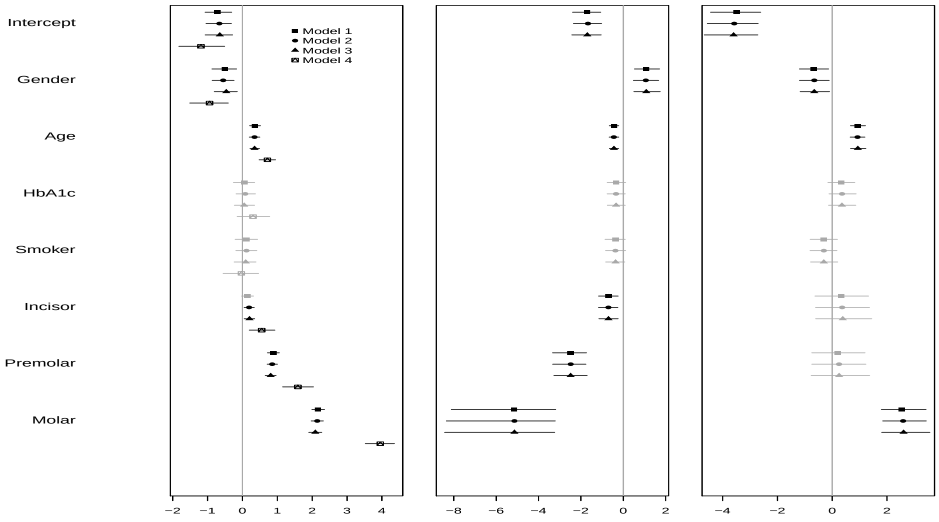

Figure 2: Posterior mean and 95% credible intervals (CI) of parameter estimates from Models 1-4 for μ (left pannel), for p_0 (middle pannel) and for p_1 (right pannel). CIs that include zero are gray, those that does not include zero are black.

in Figure 6. We do not report the estimates from the models that consider p_0 and p_1 as constants (i.e., Models 1a, 2a, and 3a). In this Figure, the gray intervals contain zero (non-significant covariates), while the black intervals do not include zero and are considered significant at 5% level. From the left panel (regression onto μ_{ij}), the covariates gender, age and tooth-types significantly explain the proportion responses mostly for Models 1-4, with the exception of Incisor for Model 1 where it is non-significant. Parameter interpretation can be expressed in terms of its effect directly on μ_{ij}, specifically $\frac{\mu_{ij}}{1-\mu_{ij}}$, conditional on the set of other covariates and REs[16]. Here, μ_{ij} is the 'expected proportion of diseased sites, and $1-\mu_{ij}$ is the complement, i.e., the 'expected remaining proportion to being completely diseased', both conditional on μ_{ij} not being zero or one. These results are interpreted in terms of the number of times the ratio is higher/lower with every unit increase (for a continuous covariate, such as age), or a change in category say from 0 to 1 (for a discrete covariate, say gender). For example, for age (a strong predictor of PrD), this ratio is 1.43 ($\exp(0.36) = 1.43$, 95% CI$=[1.23, 1.66]$) times higher for every unit increase in Age. For Gender, this ratio is 40% lower for males as compared to females, which might be influenced by the lower participation of males common in this population[27]. Similarly, this ratio is 8.7 times higher for molars as compared to the canines (the baseline), which confirms that the posteriorly placed molars typically experience a higher PrD status than the anterior canines. From the plots in the middle and right panels of Figure 6, we identify gender, age and tooth-types to be significant in explaining absence of PrD, while gender, age and molar significantly explaining the completely diseased category. Once again, we have similar odds-ratio explanation as earlier. For example, the odds of a tooth type free of PrD are 3 times greater for men than for women, while the odds of a completely diseased molar are about 13 times than of a (baseline) canine. Rest of the parameters can be interpreted similarly.

The mean estimates (standard deviations) of ϕ from Models 1-4 are 0.14 (0.007), 7.6 (0.43), 10.6 (1.56) and 0.002 (< 0.0001), and of σ_b^2 are 1.3 (0.13), 1.2 (0.13), 1.2 (0.13) and 2.6 (0.34), respectively. Due to parametrization involved, these estimates of ϕ are not comparable across Models 1-3. However, the effect of the LS transformation is evident while comparing the estimates between Models 1 and 4. Additionally, the estimates of σ_b^2 reveal that the transformation in Model 4 leads to a higher (estimated) variance of the response Y than the Models 1-3.

The adequacy of the logit link is assessed via plots of the linear predictor versus the predicted probability[28] as depicted in the Figure in Appendix C. Considering $\text{logit}^{-1}(\mu_{ij})$ from Model 1, we divided it into 10 intervals containing roughly an equal number of observations, and plot the distribution of the inverse-logit transformed linear predictors (denoted by the black box-plots) that represents the fitted mean μ_{ij} of the non-zero-one responses. Next, we overlay the empirical distributions of the observed non-zero-one responses represented by the gray box-plots. There seem to be no evidence of model misspecification, i.e., the shapes of the fitted and observed trends are similar, as revealed from Figure C in the Appendix.

In addition, we conduct sensitivity analysis on the prior assumptions for the random effects precision ($1/\sigma_b^2$) and the fixed effects precision parameters on $\boldsymbol{\beta}$ by changing one parameter at a time and refitting Model 1, as in Galvis et al.[16]. In particular, we allowed $\sigma_b \sim \text{Uniform}(0, k)$, where $k \in \{10, 50\}$, and also the typical Inverse-gamma choice on the precision $1/\sigma_b^2 \sim \text{Gamma}(k, k)$, where $k \in \{0.001, 0.1\}$. We also chose the normal precision on the fixed effects to be 0.1, 0.25 (which reflects an odds-ratio in between e^{-4} to e^4) and 0.001. There were slight changes observed in parameter estimates and model comparison values, however, that did not change our conclusions regarding the best model, inference (and sign) of the fixed-effects, and the influential observations.

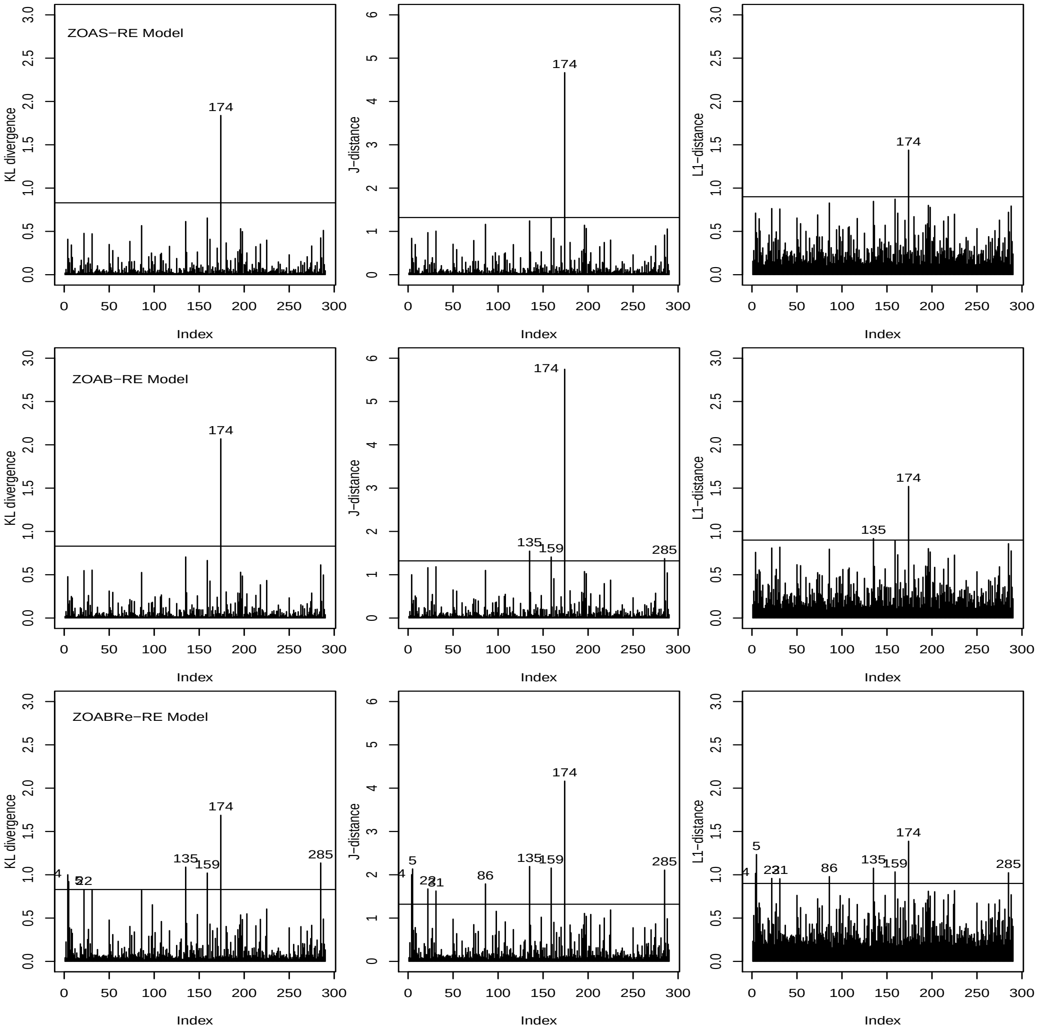

Figure 3: K-L, J and L1 divergences from the ZOAS-RE (upper panel), ZOAB-RE (middle panel) and ZOABRe-RE (lower panel) models for the PrD dataset.

Finally, we detect outlying observations via the q-divergence measures for the augmented models using the cut-offs described in Subsection 3.4. These plots are presented in Figure 3, where the upper, middle and lower panels represent the ZOAS-RE, ZOAB-RE and ZOABRe-

RE models, respectively. Interestingly, we find that the ZOABRe-RE model produces several outlying observations exceeding the threshold, whereas the best-fitting model (ZOAS-RE) produces only one such observation (subject id # 174). To quantify the impact of this observation, we refit the model by removing it. The covariate 'Molar' in the regression onto p_{0ij} is impacted by this observation, perhaps due to this subject is free of PrD for all tooth types. However, the parameter significance and sign of the coefficients remained the same. Henceforth, we stick to the estimates obtained from fitting Model 1 to the full data, without removing this particular subject.

5 Simulation Studies

In order to assess the finite sample performance of the class of AugGPD-RE mixed regression models, we conduct two simulation studies. First (Scheme 1), we assess the impact of model misspecification on the parameters for the ZOAS-RE, ZOAB-RE and ZOABRe-RE models when the data in (0,1) are generated from a logistic normal model[29]. Next (Scheme 2), we analyze the impact of the LS transformation on the parameter estimates in presence of various proportions of zeros and ones. In both studies, we generate data with various sample sizes, and compare the mean squared error (MSE), absolute relative bias (Abs.RelBias), and coverage probability (CP) of the regression parameters across the various models.

Initially, we generate y_{ij} for both schemes and sample sizes $n = 50, 100, 150, 200$ as $y_{ij} = \text{logit}^{-1}(Z_{ij})$, $i = 1, \ldots, n$ (the number of subjects), $j = 1, \ldots, 5$ (indicating cluster of size 5 for each subject), with $Z_{ij} \sim \text{Normal}(\mu_{ij}, 1)$ and the location parameter μ_{ij} modeled as $\mu_{ij} = \beta_0 + \beta_1 x_{ij} + b_i$, with $b_i \sim N(0, \sigma^2)$. The explanatory variables x_{ij} are generated as independent draws from a Uniform$(0, 1)$, with the regression parameters fixed at $\beta_0 = -0.5$, and $\beta_1 = 0.5$, variance component $\sigma^2 = 2$, and constant proportions $p_0 = 0.1$ and $p_1 = 0.1$. Thus, $y_{ij} \in (0, 1)$ are draws from a logistic-normal model. Finally, via multinomial sampling, we allocate the 0's, 1's, and the $y_{ij} \in (0, 1)$ with probabilities p_0, p_1 and $(1 - p_0 - p_1)$

respectively. No regression onto p_0 and p_1 are considered.

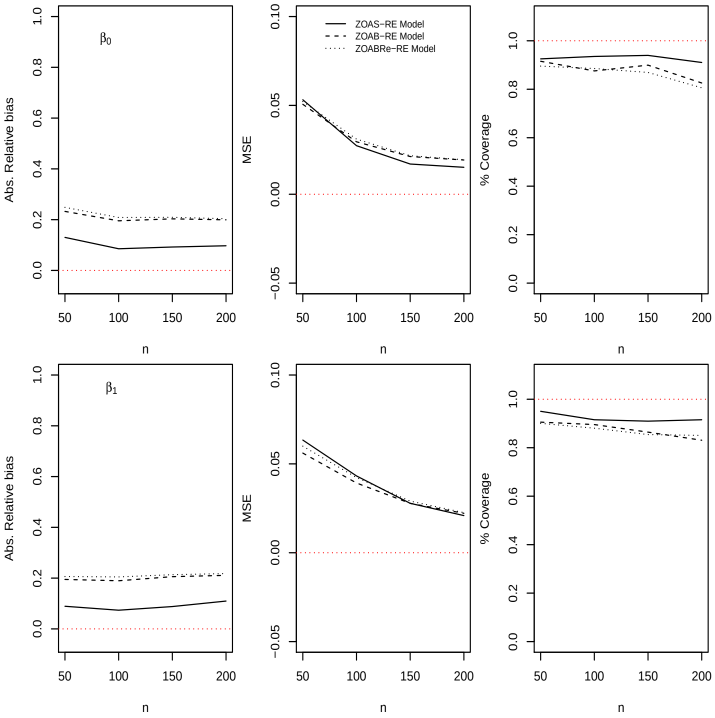

Figure 4: Absolute relative bias, MSE and coverage probability of β_0 and β_1 after fitting ZOAS-RE (continuous), ZOAB-RE (dashed) and ZOABRe-RE (dotted) models.

After simulating 200 such datasets, we fitted the ZOAS-RE, ZOAB-RE and ZOABRe-RE models with similar prior choices as in the data analysis. With our parameter space $\boldsymbol{\theta} = \{\beta_0, \beta_1, p_0, p_1, \sigma_b^2\}$, and θ_s being an element of $\boldsymbol{\theta}$, we calculate the MSE as $\text{MSE}(\hat{\theta_s}) = \frac{1}{200}\sum_{i=1}^{200}(\hat{\theta}_{is} - \theta_s)^2$, the absolute relative bias as Abs.RelBias $(\hat{\theta_s}) = \frac{1}{200}\sum_{i=1}^{200}|\frac{\hat{\theta}_{is}}{\theta_s} - 1|$, and the 95% coverage probability (CP) as $\text{CP}(\hat{\theta_s}) = \frac{1}{200}\sum_{i=1}^{200} I(\theta_s \in [\hat{\theta}_{s,LCL}, \hat{\theta}_{s,UCL}])$, where I is the indicator function such that θ_s lies in the interval $[\hat{\theta}_{s,LCL}, \hat{\theta}_{s,UCL}]$, with $\hat{\theta}_{s,LCL}$ and $\hat{\theta}_{s,UCL}$ as the estimated lower and upper bounds of the 95% CIs, respectively. The results from this study for varying sample sizes are presented in Figure 4 and Table 1 (Appendix D). Figure 4 presents a visual comparison of the models (bold line for the ZOAS-RE model, dashed line for the ZOAB-RE model and dotted line for the ZOABRe-RE model) for β_0 (upper panel) and β_1 (lower panel). For the sake of brevity, we do not produce plots for p_0, p_1 and σ_b^2.

We observe that the Abs.RelBias of both β_0, β_1 and σ_b^2 are much smaller for the ZOAS-RE model as compared to the ZOAB-RE model and the ZOABRe-RE models, while those for p_0 and p_1 are comparable. The MSEs of the parameters other than σ_b^2 are comparable. For σ_b^2, the ZOAS-RE performs better (MSE is lower) than the other two. CP remains higher for the ZOAS-RE as compared to the other two models across all parameters. Interestingly, for σ_b^2, the CP is estimated close to zero for higher n ($n = 150, 200$)

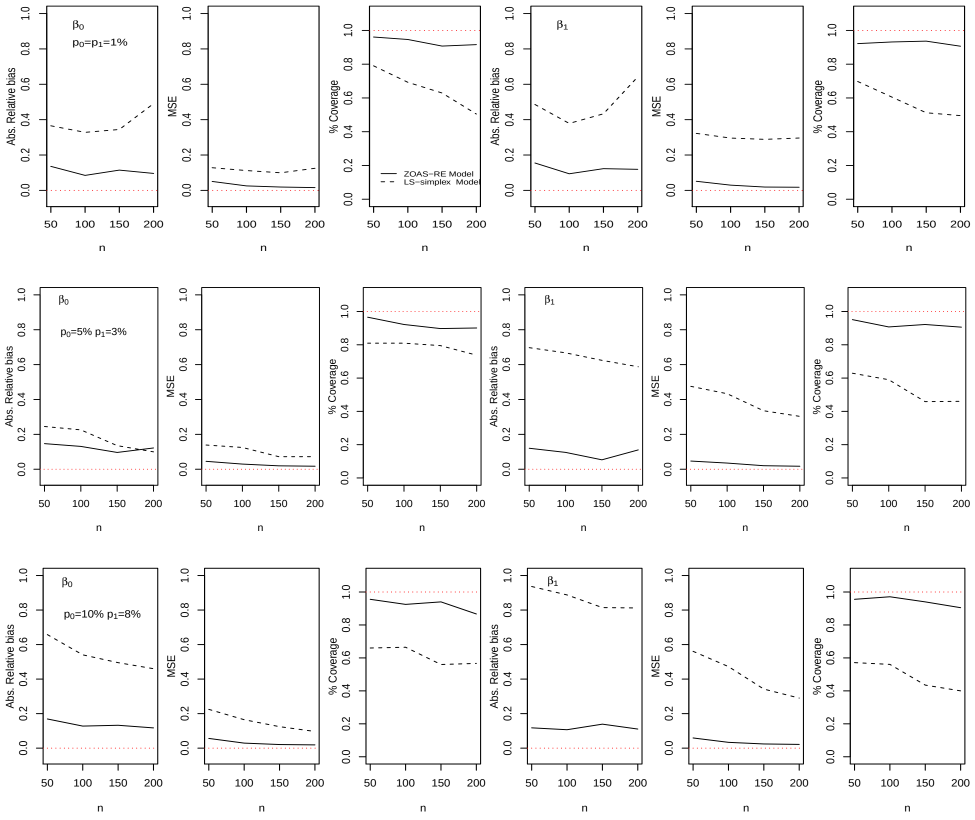

Figure 5: Absolute relative bias, MSE and coverage probability of β_0 and β_1 after fitting ZOAS-RE (continuous)and LS-simplex (dashed) models, for $p_0 = p_1 = 1\%$ (upper panel), $p_0 = 5\%$, $p_1 = 3\%$ (middle panel) and $p_0 = 10\%$, $p_1 = 8\%$ (lower panel).

In Scheme 2, we compare the performance of the ZOAS-RE and LS-simplex models for three scenarios of p_0 and p_1, namely (a): $p_0 = p_1 = 1\%$, (b) $p_0 = 3\%, p_1 = 5\%$, and (c) $p_0 = 10\%, p_1 = 8\%$ (that represents the real data). Figure 5 present the plots for MSE, Abs.RelBias and CP. The ZOAS-RE outperforms the LS-simplex model with lower MSE and Abs.RelBias, and higher CP across all scenarios, with the performance of the simplex model getting worser with increase in the proportion of 0's and 1's.

6 Conclusions

Motivated by the presence of extreme proportion responses, we develop a class of (parametric) augmented proportion density models under a Bayesian framework, and demonstrate its application to a PrD dataset. As a byproduct of the MCMC output, we also develop tools for outlier detection using results from q-divergence measures. Both simulation and real data analysis reveal the importance of utilizing an appropriate theoretical model over ad hoc data transformations.

Note that in our model development, we regress the covariates onto μ_{ij} as in Definition 2. For a direct interpretation of the covariate effect on the response Y, one might consider regressing onto δ_{ij} (the conditional expectation of the true AugGPD response) via. some link functions. However, on applying this to our dataset, we experienced problems with MCMC convergence. Hence, we did not pursue it any further, although it may be appropriate for other datasets.

The current clustered setup can be extended to a longitudinal, or a clustered-longitudinal framework (often found in dental clinical trials). In addition, the current development explores a simple parametric framework with ease in implementation. Certainly, the shape of the proportion data can also be adequately captured via some (flexible) nonparametric specification of the density. However, the Bayesian implementation may not be automatic, and would require developing customized MCMC algorithms. All these remain viable components of future research.

Acknowledgements

We thank the Center for Oral Health Research at MUSC for providing the motivating data, and Prof. Elizabeth Slate for interesting insights on clinical interpretations. Bandyopadhyay acknowledges support from the US National Institutes of Health grants UL1TR000114 (CTSA award), P30 CA77598 (University of Minnesota Masonic Cancer Center grant), R03DE021762 and R03DE023372. Galvis acknowledges support from CAPES/CNPq- IEL Nacional- Brasil. Lachos was supported by grants 305054/2011-2 from CNPq-Brazil, and 2011/17400-6 from FAPESP-Brazil.

APPENDIX

APPENDIX A: Some densities in the GPD class

- *The beta distribution*

The density of a r.v Y following the beta distribution with mean μ and precision parameter ϕ is given by

$$f(y|\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{A1}$$

with $0<E[Y]=\mu<1$, $\mathrm{Var}(Y)=\frac{\mu(1-\mu)}{1+\phi}$ and $\phi>0$.

- *The simplex distribution*

A r.v Y follows a simplex distribution with parameters μ and ϕ if its pdf is given by

$$f(y|\mu,\phi)=\frac{\sqrt{\phi}}{\left(\pi\left[y(1-y)\right]^3\right)^{1/2}}\exp\left\{-\phi\frac{(y-\mu)^2}{2y(1-y)\mu^2(1-\mu)^2}\right\}, \tag{A2}$$

with $0<E[Y]=\mu<1$, $\mathrm{Var}(Y)=\phi\mu^3(1-\mu)^3$ and $\phi>0$.

- *The beta rectangular*

A r.v Y is distribuited according to beta rectangular distribution with parameters η, λ and ϕ if its pdf is given by

$$f(y|\eta,\lambda,\phi)=\eta+(1-\eta)\frac{\Gamma(\phi)}{\Gamma(\lambda\phi)\Gamma((1-\lambda)\phi)}y^{\lambda\phi-1}(1-y)^{(1-\lambda)\phi-1}, \tag{A3}$$

with $0\leq\eta\leq 1$, $0<\lambda<1$, $\phi>0$, $E[Y]=\eta/2+(1-\eta)\lambda$ and $\mathrm{Var}(Y)=\frac{\lambda(1-\lambda)}{1+\phi}(1-\eta)(1+$

$\eta(1+\phi)) + \frac{\eta}{12}(4-3\eta)$.

APPENDIX B: Full conditional distributions from models ZOAS-RE, ZOAB-RE and ZOABRe-RE in the augmented GPD class

The full conditional distributions of the parameters $\boldsymbol{\psi}$, $\boldsymbol{\rho}$ and σ_b necessary for the MCMC algorithm in the three models above remain equal to presented for the augmented-GPD class. For the other parameters, the full conditional distributions are obtained for every model as follows.

ZOAS-RE model

- The full conditional density for $\boldsymbol{\beta}|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\boldsymbol{\beta})}$, $\pi\left(\boldsymbol{\beta}|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\boldsymbol{\beta})}\right)$ is proportional to
$$\exp\left\{-\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \phi\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\left[y_{ij}-(1-y_{ij})A_{ij}\right]^2(1+A_{ij})^2}{2y_{ij}(1-y_{ij})A_{ij}^2} I_{y_{ij}\in(0,1)}\right\},$$
where $A_{ij} = \exp\{\mathbf{X}_{ij}^\top\boldsymbol{\beta} + \mathbf{Z}_{ij}^\top\mathbf{b}_i\}$.

- The full conditional density for $\phi|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\phi)}$, $\pi(\phi|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\phi)})$ is a left truncated gamma with left truncation point a^{-2}. That is $\phi|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\phi)} \sim TGamma^{-}(a^{-2},(n-1)/2,\sum_{i=1}^{n}\sum_{j=1}^{n_i}a_3(y_{ij},\mu_{ij}))$ where $a_3(y_{ij},\mu_{ij}) = \frac{(y_{ij}-\mu_{ij})^2}{2y_{ij}(1-y_{ij})\mu_{ij}^2(1-\mu_{ij})^2}$ and $\mu_{ij} = \frac{A_{ij}}{1+A_{ij}}$.

- The full conditional density for $\mathbf{b}_i|\mathbf{y},\sigma_b,\boldsymbol{\Omega}$, $\pi(\mathbf{b}_i|\mathbf{y},\sigma_b,\boldsymbol{\Omega})$ is proportional to
$$\exp\left\{\tfrac{-1}{2\sigma_b^2}\textstyle\sum_{k=1}^{q}b_{ik}^2 - \phi\sum_{j=1}^{n_i}\frac{[y_{ij}-(1-y_{ij})A_{ij}]^2(1+A_{ij})^2}{2y_{ij}(1-y_{ij})A_{ij}^2}I_{y_{ij}\in(0,1)}\right\} I_{\{b_{ik}\in\mathbb{R}\}}.$$

ZOAB-RE model

- The full conditional density for $\boldsymbol{\beta}|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\boldsymbol{\beta})}$, $\pi\left(\boldsymbol{\beta}|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\boldsymbol{\beta})}\right)$ is proportional to

$$\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(\mu_{ij}\phi\log\frac{y_{ij}}{1-y_{ij}} - B_{ij}\right)I_{y_{ij}\in(0,1)}\right\},$$

where $B_{ij} = \log \Gamma(\mu_{ij}\phi) + \log[\Gamma(1-\mu_{ij})\phi]$, $\mu_{ij} = \frac{A_{ij}}{1+A_{ij}}$.

- The full conditional density for $\phi | \mathbf{y}, \mathbf{b}, \sigma_b^2, \mathbf{\Omega}_{(-\phi)}$, $\pi(\phi | \mathbf{y}, \mathbf{b}, \sigma_b^2, \mathbf{\Omega}_{(-\phi)})$ is proportional to

$$\phi^{a-1} \exp\left\{ -\phi \left(c - \sum_{i=1}^{n} \sum_{j=1}^{n_i} C_{ij} I_{y_{ij} \in (0,1)} \right) \right\},$$

where $C_{ij} = \mu_{ij} \log \frac{y_{ij}}{1-y_{ij}} + (1-\mu_{ij}) \log(1-y_{ij}) + \log(\phi) - B_{ij}$ and $\phi > 0$.

- The full conditional density for $\mathbf{b}_i | \mathbf{y}, \sigma_b^2, \mathbf{\Omega}$, $\pi(\mathbf{b}_i | \mathbf{y}, \sigma_b^2, \mathbf{\Omega})$ is proportional to

$$\exp\left\{ \frac{-1}{2\sigma_b^2} \sum_{k=1}^{q} b_{ik}^2 - \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(\mu_{ij}\phi \log \frac{1-y_{ij}}{y_{ij}} - B_{ij} \right) I_{y_{ij} \in (0,1)} \right\},$$

with $b_{ik} \in \mathbb{R}$.

ZOABRe-RE model

- The full conditional density for $\boldsymbol{\beta} | \mathbf{y}, \mathbf{b}, \sigma_b^2, \mathbf{\Omega}_{(-\boldsymbol{\beta})}$, $\pi\left(\boldsymbol{\beta} | \mathbf{y}, \mathbf{b}, \sigma_b^2, \mathbf{\Omega}_{(-\boldsymbol{\beta})}\right)$ is proportional to

$\exp\left\{ -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \Sigma^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \sum_{i=1}^{n} \sum_{j=1}^{n_i} I_{\{y_{ij} \in (0,1)\}} \log\left[\eta_{ij} + (1-\eta_{ij}) M_{ij}\right] \right\},$

where $M_{ij} = \frac{\Gamma(\phi)}{\Gamma(\lambda_{ij}\phi)\Gamma((1-\lambda_{ij})\phi)} y_{ij}^{\lambda_{ij}\phi - 1} (1-y_{ij})^{(1-\lambda_{ij})\phi-1}$, $\eta_{ij} = \alpha(1 - 2|\mu_{ij} - \frac{1}{2}|)$, $\lambda_{ij} = \frac{\mu_{ij} - \frac{\eta_{ij}}{2}}{1-\eta_{ij}}$ and $\mu_{ij} = \frac{A_{ij}}{1+A_{ij}}$.

- The full conditional density for $\phi | \mathbf{y}, \mathbf{b}, \sigma_b^2, \mathbf{\Omega}_{(-\phi)}$, $\pi(\phi | \mathbf{y}, \mathbf{b}, \sigma_b^2, \mathbf{\Omega}_{(-\phi)})$ is proportional to

$\phi^{a-1} \exp\left\{ -\phi c + \sum_{i=1}^{n} \sum_{j=1}^{n^i} I_{\{y_{ij} \in (0,1)\}} \log\left[\eta_{ij} + (1-\eta_{ij}) M_{ij}\right] \right\},$

with $\phi > 0$.

- The full conditional density for $\mathbf{b}_i | \mathbf{y}, \sigma_b^2, \mathbf{\Omega}$, $\pi(\mathbf{b}_i | \mathbf{y}, \sigma_b^2, \mathbf{\Omega})$ is proportional to

$\exp\left\{ \frac{-1}{2\sigma_b^2} \sum_{k=1}^{n_i} b_{ik}^2 + \sum_{i=1}^{n} \sum_{j=1}^{n_i} I_{\{y_{ij} \in (0,1)\}} \log\left[\eta_{ij} + (1-\eta_{ij}) M_{ij}\right] \right\},$

with $b_{ij} \in \mathbb{R}$.

- The full conditional density for $\alpha|\mathbf{y}, \sigma_b^2, \mathbf{\Omega}$, $\pi(\alpha|\mathbf{y}, \sigma_b^2, \mathbf{\Omega})$ is proportional to $\exp\left\{\sum_{i=1}^{n} \sum_{j=1}^{n_i} I_{\{y_{ij} \in (0,1)\}} \log\left[\eta_{ij} + (1 - \eta_{ij}) M_{ij}\right]\right\} I_{\{\alpha \in [0,1]\}}$.

APPENDIX C: Adequacy of the logit link

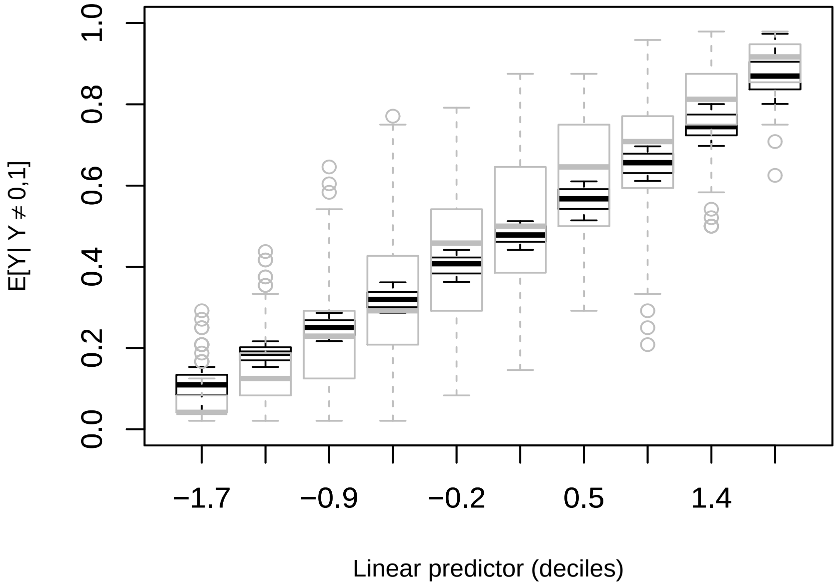

Figure 6: Observed and fitted relationship between the linear predictor and the (conditional) non-zero-one mean μ_{ij}. Modeled logit relationships are represented by black box-plots, while the empirical proportions by gray box-plots.

APPENDIX D: Simulation Results from Scheme 1

Table 2: Absolute Relative bias (Abs.RelBias), mean squared error (MSE), and coverage probabilities (CP) of the the parameter estimates after fitting the ZOAS-RE, ZOAB-RE, and ZOABRe-RE models to simulated data for various sample sizes.

	ZOAS-RE model				ZOAB-RE model				ZOABRe-RE Model			
Parameter	$n = 50$	$n = 100$	$n = 150$	$n = 200$	$n = 50$	$n = 100$	$n = 150$	$n = 200$	$n = 50$	$n = 100$	$n = 150$	$n = 200$
					Abs.RelBias							
β_0	0.13	0.08	0.09	0.10	0.23	0.19	0.20	0.20	0.25	0.21	0.21	0.20
β_1	0.09	0.07	0.09	0.11	0.19	0.19	0.20	0.21	0.21	0.20	0.21	0.22
p_0	0.02	0.02	0.00	0.0001	0.02	0.02	0.0003	0.00058	0.02	0.022	0.0009	0.0002
p_1	0.05	0.005	0.01	0.01	0.05	0.005	0.01	0.01	0.05	0.005	0.011	0.01
σ_b^2	0.19	0.20	0.22	0.226	0.37	0.38	0.40	0.40	0.38	0.38	0.40	0.40
					MSE							
β_0	0.05	0.03	0.02	0.01	0.05	0.03	0.02	0.02	0.05	0.03	0.02	0.19
β_1	0.06	0.04	0.03	0.02	0.06	0.04	0.03	0.02	0.06	0.04	0.03	0.02
p_0	0.0004	0.0002	0.0001	8e-05	0.0004	0.0002	0.0001	8e-05	0.0004	0.0002	0.0001	8e-0
p_1	0.0004	0.0001	0.0001	8e-05	0.0004	0.0001	0.0001	8e-05	0.0004	0.0001	0.0001	8e-0
σ_b^2	0.28	0.23	0.24	0.24	0.63	0.62	0.66	0.66	0.65	0.63	0.66	0.66
					CP							
β_0	92.5	93.5	94	91	91.5	87.6	90	82.6	89.6	88.6	87	80.6
β_1	95	91.5	91	91.5	90.5	89.6	86.4	83.1	90.0	88.1	85	85.1
p_0	94	92.5	93	96	94	93	93	97.5	94.5	91.5	92	95.5
p_1	94	95.5	94	96.5	92.5	95.5	94.5	96	93	95.5	94	96
σ^2	82.1	68.2	50.3	34.8	48.8	15.9	1.51	0.5	48.8	14.9	1.5	0

References

[1] Song PXK, Tan M. Marginal models for longitudinal continuous proportional data. Biometrics. 2000;56(2):496–502.

[2] Kieschnick R, McCullough BD. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. Statistical Modelling. 2003;3(3):193–213.

[3] Aitchison J. The Statistical Analysis of Compositional Data. London, U.K.: Chapman & Hall; 1986.

[4] Cepeda-Cuervo E. Modeling variability in generalized linear models. Mathematics Institute, Universidade Federal do Rio de Janeiro; 2001.

[5] Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. Journal of Applied Statistics. 2004;31(7):799–815.

[6] Hahn ED. Mixture densities for project management activity times: A robust approach to PERT. European Journal of Operational Research. 2008;188(2):450–459.

[7] Barndorff-Nielsen OE, Jørgensen B. Some parametric models on the simplex. Journal of Multivariate Analysis. 1991;39(1):106–116.

[8] Johnson N, Kotz S, Balakrishnan N. Continuous Univariate Distributions, Vol. 2. New York: John Wiley & Sons; 1994.

[9] Smithson M, Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods. 2006;11(1):54.

[10] Bayes CL, Bazan JL, Garcia C. A new robust regression model for proportions. Bayesian Analysis. 2012;7(4):841–866.

[11] Jørgensen B. The Theory of Dispersion Models. vol. 76. CRC Press; 1997.

[12] Song PXK, Qi Z, Tan M. Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. Biometrical Journal. 2004;46(5):540–553.

[13] Qiu Z, Song PXK, Tan M. Simplex mixed-effects models for longitudinal proportional data. Scandinavian Journal of Statistics. 2008;35(4):577–596.

[14] Fernandes J, Salinas C, London S, Wiegand R, Hill E, Slate E, et al. Prevalence of periodontal disease in Gullah African American diabetics. J Dent Res. 2006;85:997.

[15] Jara A, Quintana F, San Martín E. Linear mixed models with skew-elliptical distributions: A Bayesian approach. Computational Statistics & Data Analysis. 2008;52(11):5033–5045.

[16] Galvis DM, Bandyopadhyay D, Lachos VH. Augmented mixed beta regression models for periodontal proportion data. Statistics in Medicine. 2014;Available from: `http://onlinelibrary.wiley.com/doi/10.1002/sim.6179/pdf`.

[17] Peng F, Dey DK. Bayesian analysis of outlier problems using divergence measures. The Canadian Journal of Statistics. 1995;23:199–213.

[18] Dunson DB. Commentary: practical advantages of Bayesian analysis of epidemiologic data. American Journal of Epidemiology. 2001;153(12):1222.

[19] Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian analysis. 2006;1(3):515–534.

[20] Carlin BP, Louis TA. Bayesian Methods for Data Analysis (Texts in Statistical Science). Chapman and Hall/CRC, New York,; 2008.

[21] Raftery A, Newton M, Satagopan J, Krivitsky P. In: Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion). vol. 8. Oxford University Press; 2007. p. 1–45.

[22] Celeux G, Forbes F, Robert CP, Titterington DM. Deviance information criteria for missing data models. Bayesian Analysis. 2006;1(4):651–673.

[23] Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B. 2002;64(4):583–639.

[24] Cook RD, Weisberg S. Residuals and Influence in Regression. Boca Raton, FL: Chapman & Hall/CRC; 1982.

[25] Csisz I, et al. Information-type measures of difference of probability distributions and indirect observations. Studia Sci Math Hungar. 1967;2:299–318.

[26] Weiss R. An approach to Bayesian sensitivity analysis. Journal of the Royal Statistical Society, Series B. 1996;58(4):739–750.

[27] Johnson-Spruill I, Hammond P, Davis B, McGee Z, Louden D. Health of Gullah families in South Carolina with type-2 diabetes self-management analysis from Project SuGar. The Diabetes Educator. 2009;35(1):117–123.

[28] Hatfield LA, Boye ME, Hackshaw MD, Carlin BP. Multilevel Bayesian models for survival times and longitudinal patient-reported outcomes with many zeros. Journal of the American Statistical Association. 2012;107:875–885.

[29] Atchison J, Shen SM. Logistic-normal distributions: Some properties and uses. Biometrika. 1980;67(2):261–272.