Published in final edited form as:

Stat Methods Med Res. 2017 April ; 26(2): 880–897. doi:10.1177/0962280214561093.

Augmented mixed models for clustered proportion data

Dipankar Bandyopadhyay¹, Diana M Galvis², and Victor H Lachos²

¹Division of Biostatistics, University of Minnesota School of Public Health, Minneapolis, USA

²Departamento de Estatística, Universidade Estadual de Campinas, Campinas, SP, Brazil

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 to the presence of zeros and/or ones. To evade this, we introduce a class of general proportion density, and further augment the probabilities of zero and one to this general proportion density, 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 Markov chain Monte Carlo 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

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 model³ with Gaussian assumptions for logit-transformed proportion responses. However, covariate effects' interpretation is 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 distributions⁷ (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 density⁸ can represent a variety of shapes, accounting for uncorrectable non-normality and skewness⁹ 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 density⁶ and associated regression models¹⁰ were considered under a Bayesian framework. Note, the BRe regression includes the (constant dispersion) BR⁵ and the variable dispersion BR⁹ as special cases. The simplex regression¹ is based on the simplex distribution from the dispersion family.¹¹ It assumes constant dispersion and uses extended generalized estimating equations for inference connecting the mean to the covariates via the logit link. Subsequently, frameworks with heterogeneous dispersion¹² and for mixed-effects models¹³ were explored. Yet, their potential was 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 six sites of a subject's tooth, and we were interested in assessing covariate-response relationships on "proportion of diseased sites specific to a tooth-type," such as incisors, canines, pre-molars, and molars. 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 transformation⁹ of the response from [0, 1] to the interval (0, 1). These ad hoc re-scalings might work out for small proportions of zeros and ones, 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.

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 diagnostics¹⁷ 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 densities and presents some useful statistical properties. Section 3 develops the Bayesian estimation framework for the AugGPD-RE regression model and related diagnostics. Application to the motivating proportion density (PD) data appears 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 are then presented. Finally, some concluding statements appear in Section 6.

2 General proportion density

We start with the definition of PD models and then proceed to establish the GPD 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 its PDF can be expressed as

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

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 $\xi \sim \text{PD}(\lambda, \phi)$ to represent ξ a member of the PD class defined in equation (2.1). If $a_3(\xi, \lambda)$ in equation (2.1) is continuous and twice differentiable function with respect to ξ and λ and is non-zero, the variance function¹¹ is $V(\lambda) = -(\partial^2 a_3(\xi, \lambda) / \partial \lambda^2 \partial \xi) |_{\xi=\lambda}$.

Next, consider the density of the RV X following the two-component mixture $X = \eta U + (1 - \eta)\xi$, where $\eta \in [0, 1]$ is a mixture parameter and U a Uniform(0, 1) random variable distributed independently of ξ with PDF in equation (2.1). Then, X follows the GPD, i.e., $X \sim \text{GPD}(\eta, \lambda, \phi)$ with the PDF given by

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

where g_1 is as defined in equation (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 $\mu = E[X] = \eta/2 + (1 - \eta)E[\xi]$ and $\sigma^2 = \text{Var}(X) = \frac{\eta}{12} + (1 - \eta)^2 \text{Var}(\xi)$, 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 PDFs presented in online Appendix A (Available at <http://smm.sagepub.com/>).

Proposition 1—The beta density⁵ reparametrized in terms of μ (the mean) and ϕ (the precision parameter) belongs to the GPD class of distributions with its variance function given by $V(\mu) = \mu(1 - \mu)$.

Proof: In equation (2.2), consider $\eta = 0$, $\lambda = \mu$, and

$$g_1(x; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1} (1-x)^{(1-\mu)\phi-1}$$

such that

$$a_1(\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}, a_2(x, \phi) = (x(1-x)^{1-\phi})^{-1}$$

and $a_3(x, \mu) = \mu \log \frac{1-x}{x}$. Then, the variance function⁵ (from Definition 1) is

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

Proposition 2—The simplex distribution⁷ 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 equation (2.2), consider $\eta = 0$, $\lambda = \mu$, and

$$g_1(x; \mu, \phi) = \frac{\sqrt{\phi}}{\sqrt{2\pi}(x(1-x))^{3/2}} \times \exp \left\{ -\phi \frac{(x-\mu)^2}{2x(1-x)\mu^2(1-\mu)^2} \right\}$$

such that

$$a_1(\mu, \phi) = 1, a_2(x, \phi) = \frac{\phi}{\sqrt{2\pi}(x(1-x))^{3/2}}$$

and $a_3(x, \mu) = \frac{(x-\mu)^2}{2x(1-x)\mu^2(1-\mu)^2}$. Then, the variance function¹¹ 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 density⁶ with parameters η , λ , and ϕ belongs to the GPD class of distributions.

Proof: The proof follows from equation (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 online Appendix A), and a closed form expression of the variance function is not available.

For a more appealing pictorial comparison, Figure 2 plots the simplex, beta, and the BRe densities for various choices of λ and ϕ . Note that λ close to zero (one) leads to a large mass in the left (right) tails for all cases. The simplex density is relatively smooth for $\phi = 1$ and becomes more spiked for $\phi = 4$. The beta and the BRe shapes are very similar for all panels when η is moderate ($= 0.3$), as in our case. However, one observes tail behavior for the BRe compared to the beta when η gets closer to 1 (plots not shown here). From the plots, it is clear that the simplex density is more flexible than the two competitors. It is capable of capturing various shapes of the underlying proportion data density in $(0, 1)$, even in situations (say, small ϕ) where the popular beta density may be far from the ground truth. However, 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

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

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

From equation (2.3), the expectation and variance of Y are, respectively, $E[Y] = p_1 + (1 - p_0 - p_1)\mu$ 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 Definition (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 $g_1(\cdot)$ the beta density

Zero-one augmented simplex (ZOAS) density, if $\eta = 0$ and $g_1(\cdot)$ the simplex density

Zero-one augmented beta rectangular (ZOABRe) density, if $\eta > 0$ and $g_1(\cdot)$ the beta density

3 Model development and Bayesian inference

3.1 GPD regression model

Let $Y_1, \dots, Y_n$ be n independent RVs such that $Y_i \sim \text{GPD}(\eta_i, \lambda_i, \phi_i)$. Here, μ_i is directly modeled through covariates as $h_1(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, where h_1 is an adequate link function with counterdomain the real line, and $\boldsymbol{\beta}$ 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_i , which leads to a restricted parametric space of η_i defined as $0 < \eta_i < 1 - |2\mu_i - 1|$ 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 Bazan and coworkers,¹⁰ and define $\alpha_i \in [0, 1]$

such that $\alpha_i = \frac{\eta_i}{1 - (1 - \eta_i)|2\lambda_i - 1|}$. 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 equation (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 AugGPD-RE model. Let $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ be n independent continuous random vectors, where $\mathbf{Y}_i = (y_{i1}, \dots, y_{in_i})^\top$ 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 to covariates through suitable link functions as

$$h_1(\mu_{ij}) = \mathbf{X}_{\mu ij}^\top \boldsymbol{\beta} + \mathbf{X}_{b ij}^\top \mathbf{b}_i \quad (3.1)$$

$$h_2(p_{0ij}) = \mathbf{X}_{0 ij}^\top \boldsymbol{\psi} \quad (3.2)$$

$$h_3(p_{1ij}) = \mathbf{X}_{1 ij}^\top \boldsymbol{\rho} \quad (3.3)$$

where $\mathbf{X}_{\mu ij}$, $\mathbf{X}_{0 ij}$, and $\mathbf{X}_{1 ij}$ are the j th columns of the design matrices $\mathbf{X}_{\mu i}$, $\mathbf{X}_{0 i}$, and $\mathbf{X}_{1 i}$ of dimension $p \times n_i$, $r \times n_i$, and $s \times n_i$ corresponding to the vectors of fixed effects

$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top, \boldsymbol{\psi} = (\psi_1, \dots, \psi_r)^\top, \text{ and } \boldsymbol{\rho} = (\rho_1, \dots, \rho_s)^\top$$

respectively, and $\mathbf{X}_{bij}$ is the j th column of the design matrix $\mathbf{X}_{bi}$ of dimension $q \times n_i$ corresponding to the vector of REs $\mathbf{b}_i = (b_{i1}, \dots, b_{iq})^\top$. Choice of the link functions for h_1 , h_2 , and h_3 remains 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 paper, 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 constants across subjects. Finally, we denote our AugGPD-RE model as

$$Y_{ij} \sim \text{AugGPD} - \text{RE}(p_{0ij}, p_{1ij}, \mu_{ij}, \alpha, \phi)_{i=1, \dots, n, j=1, \dots, n_i}$$

Let $\mathcal{D} = (\mathbf{X}_\mu, \mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_b, \mathbf{y})^\top$ be the full observed data and $\mathbf{\Omega} = (\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\rho}, \phi, \alpha)^\top$ be the parameter vector in the AugGPD-RE model. The joint data likelihood $L(\mathbf{\Omega}; \mathcal{D}, \mathbf{b})$, conditional on the random effects $\mathbf{b}_i$, is given by

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

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

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

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

Although ML estimation of $\mathbf{\Omega}$ is certainly feasible using standard softwares (e.g., SAS, R, etc.), we seek a Bayesian treatment here. The Bayesian approach accommodates full parameter uncertainty through appropriate choice of priors choices and proper sensitivity investigations and provides direct probability statement about a parameter through credible intervals (CI).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 $\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\rho}, \phi$ (dispersion parameter), α , and the random effects $\mathbf{b}_i$. In general, for the regression components, we can assume $\boldsymbol{\beta} \sim \text{Normal}_p(\mathbf{0}, \Sigma_\beta^{-1})$, $\boldsymbol{\psi} \sim \text{Normal}_r(\mathbf{0}, \Sigma_\psi^{-1})$, $\boldsymbol{\rho} \sim \text{Normal}_s(\mathbf{0}, \Sigma_\rho^{-1})$. A Uniform (0, 1) density¹⁰ was adopted as prior for α . Prior on each element of $\mathbf{b}_i$ are $N(0, \sigma_b^2)$, where $\sigma_b \sim \text{Uniform}(0, c_1)$, the usual Gelman specification.¹⁹ The prior on ϕ for the specific models in subsection 2.1 were chosen as follows:

- i. Beta and BRe models: $\phi \sim \text{Gamma}(a, c)$, with small positive values for a and c ($c \ll a$).
- ii. (ii) Simplex model: $\phi^{-1/2} \sim \text{Uniform}(0, a_1)$, 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 equation (3.4), and the joint prior densities given by

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

where $\pi(\boldsymbol{\Omega}, \mathbf{b}, \sigma_b) = \pi_0(\boldsymbol{\beta})\pi_1(\boldsymbol{\Psi})\pi_2(\boldsymbol{\rho})\pi_3(\boldsymbol{\alpha})\pi_4(\boldsymbol{\phi})\pi_5(\mathbf{b}|\sigma_b)\pi_6(\sigma_b)$ and $\pi_j(\cdot), j=0, \dots, 6$ denote the prior/hyperprior distributions on the model parameters as described above. The full conditional distributions necessary to implement the MCMC algorithm in the AugGPD-RE model are given below. Here, the notation $\Omega_{(-\cdot)}$ stands for the parameter vector Ω with the element $(\cdot)$ excluded.

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

$$\exp \left\{ -\frac{1}{2} \boldsymbol{\Psi}^\top \Sigma_{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Psi} \right\} \prod_{i=1}^n \prod_{j=1}^{n_i} p_{0ij}^{I_{y_{ij}=0}} (1 - p_{0ij} - p_{1ij})^{I_{y_{ij} \in (0,1)}}$$

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

$$\exp \left\{ -\frac{1}{2} \boldsymbol{\rho}^\top \Sigma_{\boldsymbol{\rho}}^{-1} \boldsymbol{\rho} \right\} \prod_{i=1}^n \prod_{j=1}^{n_i} p_{1ij}^{I_{y_{ij}=1}} (1 - p_{0ij} - p_{1ij})^{I_{y_{ij} \in (0,1)}}$$

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

$$\exp \left\{ -\frac{1}{2} \boldsymbol{\beta}^\top \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta} \right\} \prod_{i=1}^n \prod_{j=1}^{n_i} g(y_{ij}; \boldsymbol{\alpha}, \mu_{ij}, \phi)^{I_{y_{ij} \in (0,1)}}$$

with $g(y_{ij}; \boldsymbol{\alpha}, \mu_{ij}, \phi)$ given by equation (3.5).

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

$$\pi(\phi) \prod_{i=1}^n \prod_{j=1}^{n_i} g(y_{ij}; \boldsymbol{\alpha}, \mu_{ij}, \phi)^{I_{y_{ij} \in (0,1)}}$$

- The full conditional density for $\boldsymbol{\alpha}|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\boldsymbol{\alpha})}, \pi(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\boldsymbol{\alpha})})$ 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 online Appendix B. Clearly, sampling from these conditional posteriors can be achieved via the popular Metropolis-within-Gibbs algorithm.²⁰ 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 independent and identically distributed Normal(0, 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 $\alpha_s \sim \text{Gamma}(1, 0.01)$, $s = 1, 2, 3$. After discarding the first 50,000 burn-in samples, we used 50,000 more samples (with a spacing of 10) from two independent chains with widely dispersed starting values for posterior summaries. Convergence was monitored via MCMC trace plots, autocorrelation plots, and the Brooks-Gelman-Rubin $\hat{R}$ statistics.²¹ Associated winBUGS code for implementing the regressions within the GPD class is available via the weblink: <http://www.biostat.umn.edu/~dipankar/ZOAS.txt>.

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) statistic²² 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.²³ 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 Akaike information criterion (EAIC), expected Bayesian information criterion (EBIC),²² and the deviance information criterion (DIC₃)²⁴ criteria. The DIC₃ was used as an alternative to the usual DIC²⁵ 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,²⁶ and the q -divergence measures^{27,28} 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 method¹⁷ to obtain the cutoff 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 ≥ 3 mm) per tooth-type as the response for each subject. Hence, this clustered data framework has four observations (corresponding to the four 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 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 zeros (114, 9.8%) and ones (94, 8.1%). 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) to (3.3), we have

$$\text{logit}(\mu_{ij}) = \mathbf{X}_{\mu ij}^\top \boldsymbol{\beta} + b_i \quad \text{logit}(p_{0ij}) = \mathbf{X}_{0ij}^\top \boldsymbol{\psi} \quad \text{logit}(p_{1ij}) = \mathbf{X}_{1ij}^\top \boldsymbol{\rho} \quad (4.1)$$

where

$$\mathbf{X}_{\mu ij} = (1, \text{Gender}_{ij}, \text{Age}_{ij}, \text{HbA1c}_{ij}, \text{Smoker}_{ij}, \text{Incisor}_{ij}, \text{Premolar}_{ij}, \text{Molar}_{ij})^\top, \\ \mathbf{X}_{\mu ij} = \mathbf{X}_{0ij} = \mathbf{X}_{1ij}, \boldsymbol{\beta} = (\beta_0, \dots, \beta_7)^\top, \boldsymbol{\psi} = (\psi_0, \dots, \psi_7)^\top$$

and $\boldsymbol{\rho} = (\rho_0, \dots, \rho_7)^\top$ 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 six competing models, varying with the densities in the GPD class and the regression over p_0 and p_1 , as follows:

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

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

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

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

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

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

Note that the parameter α is specific to the ZOABRe model only. In addition, we also fit the Lemon-squeezer (LS)-simplex model (Model 4) by transforming the response from y to y' via the LS transformation⁹ given by $y' = [y(N-1) + 1/2]/N$, where N is the number total of observations with the regression on μ as equation (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 DIC₃, LPML, EAIC, and EBIC values for the six 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 is worse than the corresponding ones with regression on p_0 and p_1 . The right panel of Figure 1 clearly tells us that the ECDF from the fitted values using Model 1 represents the true data much closely as compared to Model 4. Hence, we select Model 1 as our best model and proceed with inference.

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 in Figure 3. 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 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.²⁹ 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 3, 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 three times

greater for males than for females, 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 probability³⁰ as depicted in the Figure S1 in online Appendix C. We divided $\text{logit}(\mu_{ij})$ from Model 1 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.

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 β 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.

Finally, we detect outlying observations via the q -divergence measures for the augmented models using the cutoffs described in subsection 3.4. These plots are presented in Figure 4, 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.³¹ 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 $y_{ij} = \text{logit}^{-1}(Z_{ij})$, $i = 1, \dots, n$ (the number of subjects), $j = 1, \dots, 5$ (indicating cluster of size five 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, $y_{ij} \in (0, 1)$ is drawn from a logistic-normal model. Finally, via multinomial sampling, we allocate the zeros, ones, 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 is considered.

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

$\theta = \{\beta_0, \beta_1, p_0, p_1, \sigma_b^2\}$, and θ_s being an element of θ , we calculate the MSE as

$$\text{MSE}(\hat{\theta}_s) = 1/200 \sum_{i=1}^{200} (\hat{\theta}_{is} - \theta_s)^2, \text{ the Abs.RelBias as}$$

$$\text{Abs.RelBias}(\hat{\theta}_s) = \frac{1}{200} \sum_{i=1}^{200} \left| \frac{\hat{\theta}_{is}}{\theta_s} - 1 \right|, \text{ and the 95\% CP as}$$

$CP(\hat{\theta}_s) = \frac{1}{200} \sum_{i=1}^{200} I(\theta_s \in [\hat{\theta}_{is,LCL}, \hat{\theta}_{is,UCL}])$, where I is the indicator function such that θ_s lies in the interval $[\hat{\theta}_{is,LCL}, \hat{\theta}_{is,UCL}]$, with $\hat{\theta}_{is,LCL}$ and $\hat{\theta}_{is,UCL}$ as the estimated lower and upper bounds of the 95% CIs of θ_s at the i th iteration, respectively. The results from this study for varying sample sizes are presented in Figure 5 and Table S1 (online Appendix D). Figure 5 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 is 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 ($n = 150, 200$).

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) $p_0 = 10\%$, $p_1 = 8\%$ (representing the real data). Figure 6 presents the plots of 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 worse with increasing proportions of zeros and ones.

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 analyses reveal the importance of utilizing an appropriate theoretical model over ad hoc data transformations. Within the GPD class, the simplex density (from Figure 2) is more flexible than both its competitors. Hence, most likely the simplex regression (and its augmented counterpart) will outperform the beta and the BRe regressions for relatively non-smooth proportion data, such as, data with lots of spikes and structures, for support within $(0, 1)$ (and $[0, 1]$). However, we recommend a pragmatic modeling approach by fitting these three parametric densities successively to any dataset, and choosing the best one via popular model selection techniques.

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 editor, associate editor, and two referees whose constructive comments led to an improved presentation of the paper. We also thank the Center for Oral Health Research at MUSC for providing the motivating data, and Prof. Elizabeth Slate for interesting insights on clinical interpretations.

Funding

Bandyopadhyay acknowledges support from the US National Institutes of Health grants R03DE021762 and R03DE023372. Galvis acknowledges support from CAPES/CNPq- IEL National- Brazil. Lachos was supported by grants 305054/2011-2 from CNPq-Brazil, and 2014/02938-9 from FAPESP-Brazil.

References

1. Song PXX, Tan M. Marginal models for longitudinal continuous proportional data. *Biometrics*. 2000; 56:496–502. [PubMed: 10877309]

2. Kieschnick R, McCullough BD. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Stat Model*. 2003; 3:193–213.
3. Aitchison, J. The statistical analysis of compositional data. London, UK: Chapman & Hall; 1986.
4. Cepeda-Cuervo, E. Modeling variability in generalized linear models. Mathematics Institute, Universidade Federal do Rio de Janeiro; 2001. <http://www.bdigital.unal.edu.co/9394/> [accessed 19 July 2013]
5. Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. *J Appl Stat*. 2004; 31:799–815.
6. Hahn ED. Mixture densities for project management activity times: a robust approach to PERT. *Eur J Oper Res*. 2008; 188:450–459.
7. Barndorff-Nielsen OE, Jørgensen B. Some parametric models on the simplex. *J Multivariate Anal*. 1991; 39:106–116.
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. *Psychol Methods*. 2006; 11:54. [PubMed: 16594767]
10. Bayes CL, Bazan JL, Garcia C. A new robust regression model for proportions. *Bayesian Anal*. 2012; 7:841–866.
11. Jørgensen, B. The theory of dispersion models. Vol. 76. Boca Raton, FL: Chapman and Hall/CRC; 1997.
12. Song PXX, Qi Z, Tan M. Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometric J*. 2004; 46:540–553.
13. Qiu Z, Song PXX, Tan M. Simplex mixed-effects models for longitudinal proportional data. *Scand J Stat*. 2008; 35:577–596.
14. Fernandes J, Salinas C, London S, 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. *Comput Stat Data Anal*. 2008; 52:5033–5045.
16. Galvis DM, Bandyopadhyay D, Lachos VH. Augmented mixed beta regression models for periodontal proportion data. *Stat Med*. 2014; 33:3759–3771. [PubMed: 24764045]
17. Peng F, Dey DK. Bayesian analysis of outlier problems using divergence measures. *Can J Stat*. 1995; 23:199–213.
18. Dunson DB. Commentary: practical advantages of Bayesian analysis of epidemiologic data. *Am J Epidemiol*. 2001; 153:1222. [PubMed: 11415958]
19. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal*. 2006; 1:515–534.
20. Tierney L. Markov chains for exploring posterior distributions. *Ann Stat*. 1994; 22:1701–1762.
21. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat*. 1998; 7:434–455.
22. Carlin, BP., Louis, TA. Bayesian methods for data analysis (texts in statistical science). New York: Chapman and Hall/CRC; 2008.
23. Raftery, AE., Newton, MA., Satagopan, JM., et al. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In: Bernardo, JM., Bayarri, MJ., Berger, JO., et al., editors. *Bayesian Statistics*. Vol. 8. Oxford, New York: Oxford University Press; 2007. p. 371–416.
24. Celeux G, Forbes F, Robert CP, et al. Deviance information criteria for missing data models. *Bayesian Anal*. 2006; 1:651–673.
25. Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit. *J Roy Stat Soc B*. 2002; 64:583–639.
26. Cook, RD., Weisberg, S. Residuals and influence in regression. Boca Raton, FL: Chapman & Hall/CRC; 1982.
27. Csisz I, et al. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci Math Hungar*. 1967; 2:299–318.

28. Weiss R. An approach to Bayesian sensitivity analysis. *J Roy Stat Soc B*. 1996; 58:739–750.
29. Johnson-Spruill I, Hammond P, Davis B, et al. Health of Gullah families in South Carolina with type-2 diabetes self-management analysis from project SuGar. *Diabetes Educ*. 2009; 35:117–123. [PubMed: 19244567]
30. Hatfield LA, Boye ME, Hackshaw MD, et al. Multilevel Bayesian models for survival times and longitudinal patient-reported outcomes with many zeros. *J Am Stat Assoc*. 2012; 107:875–885.
31. Atchison J, Shen SM. Logistic-normal distributions: some properties and uses. *Biometrika*. 1980; 67:261–272.

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.

Figure 2.
Plots of the simplex, beta and the beta rectangular densities for various choices of λ and ϕ .
For the beta rectangular density, we choose $\eta = 0.3$.

Figure 3.

Posterior mean and 95% credible intervals (CI) of parameter estimates from Models 1 to 4 for the regression on μ (left panel), p_0 (middle panel), and p_1 (right panel). CIs that include zero are gray and those that do not include zero are black.

Figure 4. KL, J , and L_1 divergences from the ZOAS-RE (upper panel), ZOAB-RE (middle panel), and ZOABRe-RE (lower panel) models for the PrD dataset.

Figure 5. 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 in Scheme 1. MSE: mean squared error.

Figure 6.

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. MSE: mean squared error.

Table 1

Model comparison using DIC₃, LPML, EAIC, and EBIC criteria.

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

LPML: log pseudo-marginal likelihood; EAIC: expected Akaike information criterion; EBIC: expected Bayesian information criterion; DIC₃: deviance information criterion.