



Bayesian flexible beta regression model with functional covariate

Agnese Maria Di Brisco¹ · Enea Giuseppe Bongiorno¹ · Aldo Goia¹ · Sonia Migliorati²

Received: 15 November 2021 / Accepted: 20 May 2022 / Published online: 9 June 2022
© The Author(s) 2022

Abstract

Standard parametric regression models are unsuitable when the aim is to predict a bounded continuous response, such as a proportion/percentage or a rate. A possible solution is the flexible beta regression model which is based on a special mixture of betas designed to cope with (though not limited to) bimodality, heavy tails, and outlying observations. This work introduces such a model in the case of a functional covariate, motivated by a spectrometric analysis on milk specimens. Estimation issues are dealt with through a combination of standard basis expansion and Markov chains Monte Carlo techniques. Specifically, the selection of the most significant coefficients of the expansion is done through Bayesian variable selection methods that take advantage of shrinkage priors. The effectiveness of the proposal is illustrated with simulation studies and the application on spectrometric data.

Keywords Bayesian variable selection · Beta mixture model · Bounded response · Functional data · MCMC

1 Introduction

Determining the composition of milk is critical to establish the nutritional values of this staple of the human diet. Chromatographic techniques are very accurate in performing this task but expensive, time consuming, and require the use of well-equipped laboratories. A valid alternative is the near-infrared spectroscopy that measures the intensity of light reflected by matter as a function of wavelength in the near-infrared region of the electromagnetic spectrum. In recent years, technological

✉ Agnese Maria Di Brisco
agnese.dibrisco@uniupo.it

¹ Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale, Novara, Italy

² Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

innovation in this area has enabled the development of portable and low-cost instruments. Among them, two low-priced and portable devices are the NeoSpectra Micro Development Kit (Si-Ware), NeoSpectra for short, and the SCiO (Consumer Physics). An interesting study that motivated this work was conducted by Riu et al. (2020) on 45 commercial milk specimens: the entire sample was analysed with the NeoSpectra device whereas a subsample of size 39 with the SCiO device. Note that the SCiO sample originally included an additional observation characterized by a spectrometric curve with strongly dissimilar shape compared to the other curves. Since no additional information is available to account for this anomalous behaviour, the potential leverage specimen has been excluded from further analyses. The spectrometric curves were observed over a grid of 134 points for the NeoSpectra sample and over a grid of 331 points for the SCiO sample. Raw data have been smoothed by adopting a system of 20 cubic splines. Figure 1 shows the smoothed spectra which were suitably collected using NeoSpectra (left-hand panel) and SCiO (right-hand panel) devices. The wavelength signal recorded by the NeoSpectra device ranges between 1350 and 2558 nm, whereas the signal recorded by the SCiO one ranges between 740 and 1070 nm.

Such a collection of observations of discretised curves presents the typical features of functional data. The branch of statistics that deals with curves (but also with surfaces, images, or other objects) is known as Functional Data Analysis. In the last two decades this discipline has witnessed a strong growth thanks to the technological development of data collection devices, the increase in computational powers, and the theoretical refinement of ad-hoc mathematical tools and statistical methodologies. The interest of researchers toward this field is evidenced by a conspicuous literature (see e.g. the monographs by Ferraty and Vieu 2006; Horvath and Kokoszka 2012; Kokoszka and Reimherr 2017; Ramsay and Silverman 2005).

Along with the spectrometric analysis, the milk composition, i.e., the quantity of fat, protein, and carbohydrate, measured in g/100 mL, was detected from the food labels. Of particular interest is the analysis of the amount of fat that is evaluated as

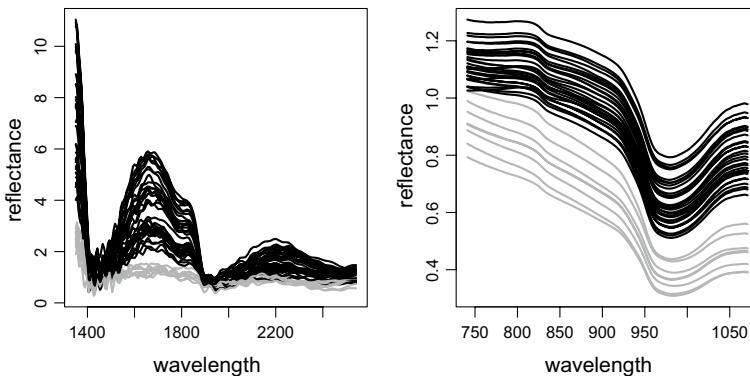


Fig. 1 Milk spectra recorded using NeoSpectra (left-hand panel) and SCiO (right-hand panel) devices. Curves are coloured according to the classification of milk specimens into “skimmed” (grey curves) and “not skimmed” (black curves)

a proportion over the remaining macronutrients. Figure 2 shows the kernel density estimates of the fat proportion in the two samples. The corresponding boxplots are provided with respect to the classification of milk specimens into “skimmed” (fat content lower than 0.5 g/100 mL) and “not skimmed”. Note that 9 out of 45 milk specimens are classified as “skimmed”. Of these, 8 are included in the subsample of 39 milk specimens analysed with the SCiO device. It is worth noting that the presence of a heavy left-tail is possibly induced by a mixture structure in the bounded responses. Interestingly, this behaviour can not be encompassed by a standard distribution for proportions such as the beta one, which is unable to give rise to unimodal shapes with heavy tails.

The problem of investigating the relationship between the fat proportion and the near-infrared spectra in milk specimens translates, from a statistical perspective, into a regression model between a real bounded response (the fat proportion) and a functional covariate (a spectrometric curve observed over a discrete mesh of an interval). This kind of scalar-on-function regression model has generated a wide range of statistical procedures both in the parametric case and in the semi and nonparametric ones (see e.g. Aneiros and Vieu 2016; Ferraty and Vieu 2002; Goia and Vieu 2015; Ling and Vieu 2018 and references therein). To the best of the authors’ knowledge, the issue of modelling a bounded response has not been thoroughly addressed in that literature and it is mostly tackled by resorting to a generalized functional linear model (GFLM) (Goldsmith et al. 2011; Greven and Scheipl 2017). The latter requires a transformation of the response that makes the interpretation of the estimated functional regression parameter more involved as it moves it from the original support to the transformed one.

Because of the boundedness of the response, outside the functional data context a widespread approach is the beta regression model. It has been introduced in the standard framework of multivariate covariates by Ferrari and Cribari-Neto (2004) and it is based on the assumption that the response variable is beta distributed. This

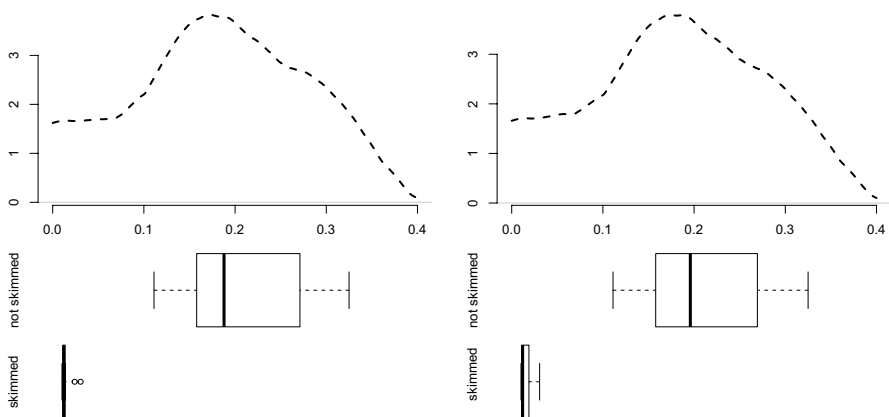


Fig. 2 Estimated densities of fat proportion (based on the Epanechnikov kernel with bandwidth equals 0.04) and their corresponding boxplots from NeoSpectra (left-hand panel) and SCiO (right-hand panel) samples

model has among its strengths the easiness of interpretation of the regression coefficients but it presents some inconveniences, shared also with GFLMs, due to the limited shapes of the beta when bimodality, outlying observations, and heavy tails appear. The latter often occur in real datasets when a relatively large number of observations take values close to the support boundaries, so that distributions which exhibit one positive finite tail limit (or even two positive but not necessarily equal tail limits) are needed. To overcome these drawbacks, more flexible distributions can be adopted. Among them, a regression model based on a special mixture of two betas, called flexible beta (FB) distribution (Migliorati et al. 2018), proved to fit well a large variety of data patterns and still being computationally tractable.

Motivated by the spectrometric problem described above, the aim of the paper is to define a functional FB regression (fFBreg) model when the covariate is functional and the response is bounded between 0 and 1. In this setting, the conditional mean of the response is bounded as well, and it is transformed so that it takes values on the real line instead of the unit interval (0, 1). This transformation is linked to the functional covariate by using a linear specification, and the estimation step is done with a Bayesian strategy. This approach to inference, which has only recently received attention in the functional data field (Crainiceanu and Goldsmith 2010; Reiss et al. 2017), has the advantage of being able to cope with complex models such as mixtures. To operationalise the latter, a truncated basis expansion of the functional terms involved in the model is adopted, allowing to write the likelihood expression in a parametric form. Among all possible strategies, the one induced by the functional principal component analysis (fPCA) is chosen. Here the preference goes to the PCs approach that has the characteristics of being estimated from data, and of acting globally, that is, on the whole support of the curves. In that framework, an important task is the selection of the principal components (PCs) to be included in the model. A standard practice is to take the first k PCs, ordered according to the explained variability, with k chosen by using the fraction of explained variance. Here, the choice of the most significant PCs is performed through Bayesian variable selection techniques that take advantage of shrinkage priors on the coefficients in the expansion (Crainiceanu and Goldsmith 2010; Malloy et al. 2010). Thanks to the combination of the FB specification and Bayesian thought, the obtained model is suitable for bounded responses when a mixture structure appears and a proper variable selection approach is needed.

The effectiveness of the model and of the Bayesian estimation are illustrated by simulated numerical experiments, that allow also to evaluate the abilities of the variable selection method. Finally, the proposed approach is applied to the prediction of the proportion of fat of milk specimens by using the spectrometric curves illustrated above. Since the NeoSpectra and SCiO spectra are materially different, in the sense that they differ in terms of wavelengths and reflectance ranges (see Fig. 1), the corresponding models are analysed separately.

The outline of the paper is as follows. Section 2 is devoted to the introduction of the main notations and the regression model. Section 3 describes the Bayesian approach to inference focusing on the priors choice and the Markov chain Monte Carlo (MCMC) techniques. Section 4 illustrates and discusses the results from some numerical studies. Section 5 shows the performances of the proposal on the

spectrometric data. Finally, in Sect. 6 some conclusions are drawn about the methodology illustrated in the paper.

2 Notation and model

Consider a real random variable Y taking values on the unit interval $(0, 1)$, and a functional random curve X valued in $\mathcal{L}^2(\mathcal{T})$, the space of square integrable real functions defined over a compact interval \mathcal{T} , equipped with its natural inner product $\langle f, g \rangle = \int_{\mathcal{T}} f(t)g(t)dt$ and induced norm $\|f\| = \langle f, f \rangle^{1/2}$.

Assume that the conditional distribution of Y with respect to X is a special mixture of two betas, namely an FB distribution with the following conditional pdf (in the notation, the dependence on x is dropped for simplicity):

$$f_{FB}(y|\lambda_1, \lambda_2, \phi, p) = pf_B(y|\lambda_1, \phi) + (1 - p)f_B(y|\lambda_2, \phi), \quad 0 < y < 1, \quad (1)$$

where λ_1 and λ_2 are dependent on x , with the constraint $0 < \lambda_2 < \lambda_1 < 1$ to ensure identifiability, $0 < p < 1$, $\phi > 0$, and $f_B(y|\lambda_j, \phi)$ denotes the pdf of a beta random variable with parameters λ_j and ϕ , corresponding to the mean and the precision, respectively. Hence, the parameters λ_1 and λ_2 represent the (conditional) means of the first and second mixture components, respectively, whereas ϕ is the precision parameter, and p is a mixing proportion parameter (see Migliorati et al. 2018 for further details on the FB distribution). It is worth noting that the FB distribution contains the beta distribution as an inner point. Indeed, fixing $\phi = 1/(\lambda_1 - \lambda_2)$ and $p = \lambda_2/(1 - \lambda_1 + \lambda_2)$ in (1), it is possible to show that the FB coincides with a beta with mean $\lambda_2/(1 - \lambda_1 + \lambda_2)$ and precision equal to $(1 - \lambda_1 + \lambda_2)/(\lambda_1 - \lambda_2)$.

In order to define the fFBreg model, the conditional mean of the response must be linked to the functional covariate. Here, the following linear specification is adopted:

$$g(\mu) = \alpha + \langle \beta, x \rangle, \quad (2)$$

where $\alpha \in \mathbb{R}$, $\beta \in \mathcal{L}^2(\mathcal{T})$ is the functional regression coefficient, $g(\cdot)$ is a monotone and twice differentiable link function (a common choice being the function $logit(u) = \log(u/(1 - u))$ for ease of interpretation), and μ is the overall conditional mean of the mixture defined in (1):

$$\mu = \mathbb{E}[Y|x] = p\lambda_1 + (1 - p)\lambda_2.$$

Since the original parametrization of the FB does not explicitly include μ , a reparametrisation is needed. In particular, the set μ , ϕ , and p must be complemented with $w = (\lambda_1 - \lambda_2)/\min\{\mu/p, (1 - \mu)/(1 - p)\}$, a standardized measure of the distance between the two mixture components. The final parametric space of $f_{FB}(y|\mu, \phi, w, p)$, where $0 < \mu < 1$, $0 < w < 1$, $0 < p < 1$, and $\phi > 0$, is variation independent, i.e., no additional constraints on the parametric space exist (Migliorati et al. 2018).

Following similar arguments, it is possible to define a simplified version of the model, namely the functional beta regression (fBreg) model. Such a model

postulates that $Y \sim \text{Beta}(\mu, \phi)$, and the mean parameter μ is regressed onto functional covariates as in Eq. (2). It is worth noting that both the fFBreg and fBreg models differ from standard GLM since the related distributions do not belong to the exponential family.

Consider now a sample $\{(X_i, Y_i), i = 1, \dots, n\}$, $n \geq 1$, of independent copies of (X, Y) . A possible strategy to deal with the linear specification in (2), that involves a functional parameter β belonging to an infinite dimensional space, is to approximate it over a finite dimensional subspace of $\mathcal{L}^2(\mathcal{T})$. In particular, given an orthogonal basis $(\psi_k)_{k=1}^\infty$ of $\mathcal{L}^2(\mathcal{T})$, the approximated sample version of Eq. (2) can be written as:

$$\tilde{g}(\mu_i) = \alpha + \sum_{k=1}^K b_k \chi_{ik}, \quad i = 1, \dots, n,$$

where $b_k = \langle \beta, \psi_k \rangle$ are unknown real parameters, $\chi_{ik} = \langle X_i, \psi_k \rangle$ are real random elements, and K is the dimension of the approximating subspace (with $K \geq 1$).

The choice of the basis and the dimension K become, in this context, relevant steps of the analysis. Here the preference goes to the PCs bases because of their data-driven and global nature. For what concerns K , a standard practice is to choose it according to the fraction of explained variance. Since this approach does not guarantee to select the PCs that better explain the response (see e.g. Jolliffe 1982), in order to choose how many (and which ones) components of the approximation are needed in the regression framework, a Bayesian approach is exploited, in particular by taking advantage of Bayesian techniques of variable selection (O'Hara and Silanpää 2009). Moreover, the Bayesian framework is also convenient to cope with complex models such as the mixture model at hand.

3 Bayesian inference

Bayesian inference on the unknown parameters requires the computation of the posterior distribution, which can be accomplished starting from the prior distribution and the likelihood function. This computation admits no analytical solution in the given framework, therefore MCMC techniques are resorted to, and specifically a Metropolis within Gibbs algorithm. The latter is based on the traditional Gibbs sampling (i.e. the iterative sampling from the full conditional distributions), complemented with the implementation of a Metropolis step when the full conditionals can not be determined analytically. The computation of the likelihood is straightforward:

$$L(\mathbf{y}|\boldsymbol{\eta}) = \prod_{i=1}^n f_{FB}(y_i|\boldsymbol{\eta}), \quad (3)$$

where $\mathbf{y} \in \mathbb{R}^n$ is the vector of the observations, $\boldsymbol{\eta} = (\alpha, \mathbf{b}, \phi, p, w) \in \mathbb{R}^{K+4}$ is the vector of the unknown parameters, and f_{FB} is given by (1). It is worth noting that the membership of each i -th observation to either component of the mixture is unknown and this makes the (mixture) likelihood a tricky function of the unknown parameters. A possible solution to make the sampling from the likelihood, and thus

from the posterior, more feasible is to adopt the data augmentation strategy (Tanner and Wong 1987) that consists in associating each observation with a latent variable \mathbf{v} identifying the allocation to the mixture components. The resulting probabilistic scheme is thus the following:

$$\begin{cases} v_i \sim \text{Bernoulli}(p), \\ y_i | v_i = 1 \sim \text{Beta}(\lambda_{1i}, \phi), \\ y_i | v_i = 0 \sim \text{Beta}(\lambda_{2i}, \phi), \end{cases} \tag{4}$$

$i = 1, \dots, n$. Thanks to the above notation, the complete-data likelihood writes:

$$L(\mathbf{y}, \mathbf{v} | \boldsymbol{\eta}) = \prod_{i=1}^n [pf_B(y_i | \lambda_{1i}, \phi)]^{v_i} [(1 - p)f_B(y_i | \lambda_{2i}, \phi)]^{1-v_i}, \tag{5}$$

which is much more tractable. Note that the (mixture) likelihood defined in (3) can be restored from (5) by marginalizing out the latent variable \mathbf{v} .

With respect to the a priori information, non-informative priors are selected with the scope of inducing the least impact on the posterior distributions. The coefficients b_k are associated with spike-and-slab priors showing a mixture structure:

$$b_k | I_k, \tau_{b_k} \sim I_k \mathcal{N}(0, \tau_{b_k}^{-1}) + (1 - I_k) \delta(0), \quad k = 1, \dots, K.$$

The first component of the mixture is a diffuse normal (the slab) with a precision hyper-parameter τ_{b_k} that is gamma distributed with both its parameters small enough to induce non-informativeness. Whereas, the second component of the mixture $\delta(0)$ indicates a discrete measure concentrated at zero (the spike). The mixture weights of the spike-and-slab priors I_k are Bernoulli distributed with hyper-parameter q_k following a uniform prior over $(0, 1)$. Moreover, for the intercept term α a diffuse normal prior is chosen, and for the precision parameter ϕ a non-informative gamma distribution is taken. Finally, uniform priors over $(0, 1)$ are chosen for the additional parameters w, p of the FB distribution.

Denoting by $\pi(\cdot)$ a generic prior distribution, $\pi(\boldsymbol{\eta})$ is the joint prior of the unknown parameters. As observed above, without elaborating further, the parametric space of the FB is variation independent. This characteristic is particularly helpful in a Bayesian framework since it allows inference to be made under the hypothesis of a priori independence of the parameters, and this results into the factorization of the joint prior as follows:

$$\pi(\boldsymbol{\eta}) = \pi(\alpha)\pi(\mathbf{b})\pi(\phi)\pi(p)\pi(w). \tag{6}$$

Given the likelihood (5) and the prior (6) as specified above, the posterior distribution is:

$$f(\boldsymbol{\eta} | \mathbf{y}) \propto \pi(\boldsymbol{\eta})L(\mathbf{y} | \boldsymbol{\eta}),$$

which is analytically intractable. In order to get samples from the posterior distribution, one takes advantage of the Gibbs sampling algorithm that is briefly described

in what follows. The main rationale is that given some initial values, that are set at random, the sampling from the full conditional distributions is repeated iteratively, always conditioning on the most recent values of the other variables. Note that the algorithm takes advantage of the complete likelihood (5), as it emerges from its description that follows:

1. Initialize, at random, the unknown vector of parameters

$$\boldsymbol{\eta}^{(0)} = (\boldsymbol{\alpha}^{(0)}, \mathbf{b}^{(0)}, \boldsymbol{\phi}^{(0)}, p^{(0)}, w^{(0)}).$$

2. Repeat for $s = 1, \dots, S$, (until convergence):

- (a) For $i = 1, \dots, n$, generate $v_i^{(s)}$ from a Bernoulli distribution such that

$$\mathbb{P}\left(v_i^{(s)} = 1 \mid p^{(s-1)}, \lambda_{1i}^{(s-1)}, \lambda_{2i}^{(s-1)}, \boldsymbol{\phi}^{(s-1)}, y_i\right) = \frac{p^{(s-1)} f_B\left(y_i \mid \lambda_{1i}^{(s-1)}, \boldsymbol{\phi}^{(s-1)}\right)}{p^{(s-1)} f_B\left(y_i \mid \lambda_{1i}^{(s-1)}, \boldsymbol{\phi}^{(s-1)}\right) + (1 - p^{(s-1)}) f_B\left(y_i \mid \lambda_{2i}^{(s-1)}, \boldsymbol{\phi}^{(s-1)}\right)}.$$

- (b) Generate $p^{(s)}$ from a $Beta\left(\left(\sum_{i=1}^n v_i^{(s)} + 1\right)/(n + 2), n + 2\right)$.
- (c) Generate the remaining parameters $\boldsymbol{\alpha}^{(s)}, \mathbf{b}^{(s)}, \boldsymbol{\phi}^{(s)}, p^{(s)}$, and $w^{(s)}$ from their full conditional distributions. Since the full conditionals can not be written in closed form, a direct sampling from each of them is not feasible. Because of this, a Metropolis–Hasting algorithm (Robert and Casella 1999) is performed within this step of the Gibbs sampling.

The implementation of the estimation algorithm is done through the software OpenBUGS and R (Lunn et al. 2009).

4 Numerical studies

This section illustrates some numerical studies to comparatively assess the performances of the fFBreg and fBreg models: Sect. 4.1 describes the simulative settings whereas results are discussed in Sect. 4.2.

4.1 Simulation settings

The data generating mechanism is configured coherently with the functional regression framework described in Sect. 2. The functional covariate is a Wiener process. In particular, data are simulated as follows:

$$x_i(t) = \sum_{k=1}^M \varepsilon_{ki} \lambda_k \psi_k(t) \quad i = 1, \dots, n,$$

where $\varepsilon_{ki} \sim N(0, 1)$ iid, $\lambda_k = \sqrt{2}/((k - 1/2)\pi)$, $\psi_k(t) = \sin((k - 1/2)\pi t)$, $t \in \{t_0 = 0, \dots, t_{100} = 1\}$ equispaced, $M = 100$ to mimic the infinite dimension of the process, and $n = 50, 100, 200$.

Two alternative scenarios are set up that differ with respect to the selected elements of the basis of the functional regression coefficient, namely: $\beta(t) = 0.5\psi_2(t)$ in scenario A and $\beta(t) = -0.3\psi_1(t) - 0.7\psi_2(t) + 0.5\psi_4(t)$ in scenario B. These two scenarios are designed to account for different levels of complexity of the approximating subspace. Moreover, with the purpose of trying out the Bayesian variable selection strategy, the elements of the basis are chosen non-consecutively.

For each of the two outlined scenarios, three cases that differ with respect to the response-generating mechanism are considered:

- case (1): the observed responses are simulated from a beta with the mean parameter that is linked to the functional covariates according to Eq. (2), the intercept term is equal to $\alpha = -0.5$, and the precision parameter is equal to $\phi = 50$;
- case (2): the observed responses are simulated from a FB with the mean and precision parameters as in case (1) and the additional parameters are equal to $p = 0.3$, and $w = 0.4$;
- case (3): the observed responses are simulated from a FB with the mean and precision parameters as in case (1) and the additional parameters are equal to $p = 0.3$, and $w = 0.8$.

The substantial difference between cases (2) and (3), where the response is of FB type, concerns the value of the parameter w which governs the distance between the component means. The chosen values lead to a mixture with either overlapped (case (2)) or well separated (case (3)) components.

For the purpose of this numerical study, 100 Monte Carlo replications are performed. The estimate of the unknown parameters of fFBreg and fBreg models is dealt with MCMC methods based on the Gibbs sampling algorithm illustrated in Sect. 3. For both scenarios, in all cases, and for all sample sizes, chains at least of length 10,000 are simulated and the first half of the values are discarded. Convergence to the equilibrium distribution was assessed through analytical and graphical tools. In particular, the Geweke and Heidel diagnostics were used to assess stationarity, and the Raftery diagnostic was useful in ascertaining the level of autocorrelation (Gelman et al. 2014). As for the graphical tools, traceplots, density plots, and autocorrelation plots were inspected.

The goodness of fit of the models is compared through the wide applicable information criterion (WAIC), a fully Bayesian criterion that penalizes the goodness of fit for an estimate of the effective numbers of parameters (Gelman et al. 2014). The criterion is computed as $-2(\widehat{LPPD} - \widehat{p}_{\text{WAIC}})$ where:

$$\widehat{LPPD} = \sum_{i=1}^n \log \left(\frac{1}{S} \sum_{s=1}^S f_{\cdot}(y_i | \boldsymbol{\eta}^{(s)}) \right),$$

$$\widehat{p}_{\text{WAIC}} = 2 \sum_{i=1}^n \left[\log \left(\frac{1}{S} \sum_{s=1}^S f_{\cdot}(y_i | \boldsymbol{\eta}^{(s)}) \right) - \frac{1}{S} \sum_{s=1}^S \log(f_{\cdot}(y_i | \boldsymbol{\eta}^{(s)})) \right],$$

where $f_{\cdot}(\cdot)$ denotes either $f_{FB}(\cdot)$ or $f_B(\cdot)$, respectively when evaluating the criterion for an fFBreg or an fBreg model. The lower the estimate the better the fit of the model. Moreover, the quadratic L^2 distance between the true and the estimated functional regression coefficients is evaluated. The calculation is done through numerical approximation.

4.2 Results and discussion

The results for the simulative scenarios outlined above are reported in Tables 1 (scenario A) and 2 (scenario B). Both the Monte Carlo posterior mean and standard deviation (SD) and some performance measures are shown.

A first analysis concerns the estimates of the additional parameters of the fBreg (i.e. the precision ϕ) and fFBreg models (i.e., the precision ϕ , the mixing proportion p , and the distance measure w). In case (1) of both scenarios, when the response comes from a beta distribution, the fBreg model provides unbiased estimates of the precision parameter with an SD that decreases as the sample size increases. Differently, the fFBreg model adapts to the beta distributed data by slightly overestimating the precision parameter, but the bias reduces as the sample size increases. Moreover, the fFBreg model identifies two equally weighted components that are almost overlapping and being centred around almost the same mean. This is because the estimate of the mixing proportion p is around 0.5 (stable as the sample size increases) and the estimate of the cluster distance w is low (tending to decrease as the sample size increases). In cases (2) and (3) of both scenarios, the fFBreg model provides unbiased estimates for all the additional parameters ϕ , p , and w , with SDs that decrease as the sample size increases, as expected. On the contrary, the fBreg struggles to adapt to the bimodal structure, more or less evident (cases (2) and (3), respectively), from the data; in the light of the possible shapes of the beta distribution, its only way of flexibility lies in tuning the precision parameter. As a result, in both scenarios the precision parameter is underestimated by the fBreg model when the data generating mechanism is of FB type. The estimates appear to be stable as the sample size increases, whereas SDs decrease. The underestimation is more severe in case (3) than in (2), since in the latter case the bimodality is less accentuated. Moreover, the underestimation is more evident in scenario A than in B. The reason for this result is that the variability of the FB distributed response is greater in scenario A than in B in light of the combination of the parameters. It is worth mentioning, in fact, that the observed variability does not depend only on the precision parameter ϕ but also on the other ones.

As for the goodness of fit, it is worth noting that in both scenarios, in all cases, and for both the considered models, the Monte Carlo means of the WAIC estimates

Table 1 Scenario A: Monte Carlo posterior means and SDs (in parenthesis)

			<i>n</i> = 50	<i>n</i> = 100	<i>n</i> = 200
Case (1)	fBreg	$\phi = 50$	53.260 (11.631)	51.344 (7.074)	50.835 (5.041)
		WAIC	- 125.108 (10.634)	- 255.255 (13.436)	- 516.936 (19.527)
		R^2	0.766 (0.060)	0.744 (0.050)	0.744 (0.032)
		L^2	2.870 (0.635)	2.737 (0.503)	2.551 (0.316)
	fFBreg	ϕ	70.458 (21.514)	62.278 (11.972)	57.851 (7.299)
		p	0.509 (0.131)	0.518 (0.139)	0.572 (0.156)
		w	0.187 (0.043)	0.151 (0.028)	0.137 (0.047)
		WAIC	- 124.530 (10.729)	- 254.645 (13.284)	- 516.310 (19.560)
		R^2	0.841 (0.052)	0.812 (0.042)	0.794 (0.033)
		L^2	2.890 (0.650)	2.747 (0.509)	2.559 (0.314)
Case (2)	fBreg	ϕ	8.666 (1.284)	8.519 (0.999)	8.352 (0.675)
		WAIC	- 43.181 (7.310)	- 92.901 (10.857)	- 187.416 (15.638)
		R^2	0.381 (0.097)	0.345 (0.070)	0.323 (0.054)
		L^2	3.493 (1.421)	2.897 (0.763)	2.740 (0.496)
	fFBreg	$\phi = 50$	52.638 (13.742)	51.296 (8.054)	49.685 (4.797)
		$p = 0.3$	0.298 (0.048)	0.298 (0.035)	0.302 (0.019)
		$w = 0.4$	0.402 (0.031)	0.402 (0.018)	0.400 (0.013)
		WAIC	- 70.743 (11.037)	- 151.769 (15.648)	- 305.092 (20.122)
		R^2	0.928 (0.028)	0.923 (0.012)	0.918 (0.008)
		L^2	2.919 (0.811)	2.646 (0.479)	2.579 (0.338)
Case (3)	fBreg	ϕ	2.055 (0.280)	2.057 (0.223)	2.039 (0.126)
		WAIC	- 10.975 (8.293)	- 29.619 (11.903)	- 66.880 (17.266)
		R^2	0.159 (0.078)	0.135 (0.054)	0.119 (0.039)
		L^2	5.175 (2.575)	3.524 (1.357)	2.957 (0.779)
	fFBreg	$\phi = 50$	54.980 (12.566)	51.342 (8.151)	50.575 (4.516)
		$p = 0.3$	0.299 (0.044)	0.295 (0.034)	0.297 (0.021)
		$w = 0.8$	0.798 (0.016)	0.800 (0.010)	0.800 (0.008)
		WAIC	- 99.031 (13.751)	- 204.258 (17.706)	- 419.327 (24.821)
		R^2	0.979 (0.007)	0.977 (0.005)	0.976 (0.003)
		L^2	2.795 (0.616)	2.665 (0.468)	2.534 (0.325)

decrease as the sample size increases. The increase in the corresponding SDs is only apparent and due to a change in the magnitude of the estimates, as it can be immediately shown by computing any measure of relative variability. By looking at the results from both scenarios, it is evident that when the response variable is of beta type, i.e., case (1), the goodness of fit of the fFBreg model is completely equivalent to the one of the fBreg model, despite the latter model being the response-generating mechanism. In case (2) of both scenarios, the response variable shows a mild bimodality since, by setting $w = 0.4$, the two component means are not so far from each other. The fBreg behaves acceptably in terms of fit, but despite this, the goodness of fit of the fFBreg model is better. In case (3) of both scenarios, the parameter w is

Table 2 Scenario B: Monte Carlo posterior means and SDs (in parenthesis)

			$n = 50$	$n = 100$	$n = 200$
Case (1)	fBreg	$\phi = 50$	51.421 (11.353)	50.416 (7.259)	50.086 (4.672)
		WAIC	- 127.475 (10.930)	- 265.215 (15.122)	- 539.224 (19.212)
		R^2	0.907 (0.027)	0.901 (0.022)	0.901 (0.013)
		L^2	2.385 (0.575)	2.221 (0.439)	2.007(0.352)
	fFBreg	ϕ	67.288 (22.645)	60.456 (11.572)	56.601 (8.043)
		p	0.460 (0.115)	0.480 (0.137)	0.473 (0.157)
		w	0.187 (0.032)	0.148 (0.030)	0.128 (0.030)
		WAIC	- 127.138 (11.102)	- 264.750 (14.987)	- 538.685 (19.251)
		R^2	0.936 (0.022)	0.927 (0.019)	0.920 (0.014)
		L^2	2.374 (0.571)	2.235 (0.463)	2.014 (0.352)
Case (2)	fBreg	ϕ	12.751(2.125)	13.692 (1.946)	13.357 (1.262)
		WAIC	- 62.780 (8.995)	- 143.302 (13.969)	- 293.044 (19.652)
		R^2	0.699 (0.071)	0.706 (0.057)	0.696 (0.041)
		L^2	3.080 (0.696)	2.968 (0.624)	2.548 (0.587)
	fFBreg	$\phi = 50$	51.337 (19.039)	48.138 (8.884)	49.575 (6.295)
		$p = 0.3$	0.313 (0.063)	0.295 (0.052)	0.296 (0.029)
		$w = 0.4$	0.386 (0.048)	0.392 (0.036)	0.399 (0.021)
		WAIC	- 81.616 (12.457)	- 176.557 (13.568)	- 367.126 (21.992)
		R^2	0.951 (0.026)	0.949 (0.012)	0.948 (0.006)
		L^2	2.523 (0.607)	2.159 (0.480)	1.954 (0.314)
Case (3)	fBreg	ϕ	3.055 (0.755)	3.277 (0.437)	3.172 (0.266)
		WAIC	- 18.978 (9.600)	- 55.455 (13.466)	- 116.627 (16.907)
		R^2	0.384 (0.125)	0.430 (0.084)	0.419 (0.060)
		L^2	3.508 (0.922)	3.238 (0.485)	2.983 (0.510)
	fFBreg	$\phi = 50$	51.753 (11.668)	51.080 (7.207)	49.958 (5.506)
		$p = 0.3$	0.311 (0.053)	0.300 (0.034)	0.304 (0.025)
		$w = 0.8$	0.795 (0.025)	0.798 (0.019)	0.801 (0.010)
		WAIC	- 88.766 (10.854)	- 192.647 (15.689)	- 392.493 (24.139)
		R^2	0.970 (0.007)	0.967 (0.007)	0.966 (0.005)
		L^2	2.179 (0.516)	2.043 (0.350)	1.839 (0.222)

set to 0.8, and this implies a bimodal response variable characterized by two well-separated groups. Therefore, the goodness of fit of the fFBreg model is far better than the one of the fBreg model, the latter model being not designed to cope with latent structures in data.

Moreover, the inspection of the R^2 allows to evaluate the goodness of fit of the models in terms of the proportion of variability of the response that is explained by the functional covariate. Overall, the results are stable as the sample size varies and consistent with those observed for the WAIC.

Finally, the results concerning the L^2 distance give an insight into the goodness of the estimates of the functional regression coefficients, that depends both on the

ability of the model in identifying the truly significant elements of the basis and on the goodness of the estimates of the real parameters $b_k, k = 1, \dots, K$. Interestingly, the findings are quite similar in the two scenarios. It emerges that in case (1) the MC means of L^2 for the fBreg and fFBreg models are similar, thus confirming the ability of the latter model to well adapt to beta-distributed responses. Conversely, in cases (2) and (3) the fFBreg model always provides better estimates of the functional coefficient, thus resulting in lower MC means of the L^2 distance. This is particularly evident in case (3) where the simulated response comes from a mixture with well-separated components.

To complete the analysis, the performances of the used selection criterium are discussed. For what concerns scenario A, the performance of the variable selection strategy is very good. In particular, the only issues concern the fBreg model in cases (2) and (3) when $n = 50$ with a correct selection rate equals 83% within the whole Monte Carlo simulations. For the more challenging scenario B, Table 3 reports the percentage of times the first five PCs are selected in the Monte Carlo replicates. Those PCs explain more than 95% of the overall variability of the functional data: a wrong selection of a PC from the sixth onwards has thus a negligible impact on the overall estimate of the functional regression coefficient $\beta(t)$.

One can observe that the parameters associated with the third and fifth PCs, that are not significant, are wrongly selected in rather few cases whereas, the first, second, and forth PCs are correctly selected with very high rates. It is worth noticing that for the fFBreg model the higher the coefficient magnitude is, the better the selection approach behaves and, in any ways, performances improve

Table 3 Scenario B: Percentage of times the PCs associated with coefficients $b_k, k = 1, \dots, 5$ are selected in the Monte Carlo replicates

Coefficients	Sample Size	Case (1)		Case (2)		Case (3)	
		fBreg	fFBreg	fBreg	fFBreg	fBreg	fFBreg
$b_1 = -0.3$	$n = 50$	61	58	40	52	40	66
	$n = 100$	69	70	50	72	45	89
	$n = 200$	90	90	50	97	57	99
$b_2 = -0.7$	$n = 50$	97	96	77	94	73	100
	$n = 100$	100	100	86	100	78	100
	$n = 200$	100	100	96	100	83	100
$b_3 = 0$	$n = 50$	11	11	8	14	6	19
	$n = 100$	9	9	5	15	3	15
	$n = 200$	12	14	6	11	4	19
$b_4 = 0.5$	$n = 50$	86	82	56	75	62	95
	$n = 100$	98	97	68	97	66	100
	$n = 200$	100	100	82	100	76	100
$b_5 = 0$	$n = 50$	10	10	5	7	3	11
	$n = 100$	9	10	4	5	6	14
	$n = 200$	8	6	3	8	5	16

as the sample size increases. On the other hand, the fBreg produces poor performances coherently with the overall poor fit of the model in terms of L^2 distance.

5 Results on milk data

The present section provides some key findings from the analysis on the samples of milk specimens, and it is divided into two sub-sections, coherent with a two-fold objective. On the one hand, the interest is in assessing the goodness of fit of the fFBreg and fBreg models and in evaluating how well they adapt to the data. This step of analysis is done on the entire samples and it is referred to as “in-sample”. On the other hand, the focus is on the evaluation of the prediction accuracy of the proposed model, comparing it also with alternative regression models. This step of analysis is based on cross-validation and it is referred to as “out-of-sample”.

The spectrometric curves illustrated in Fig. 1 have an inherently functional nature and therefore it is possible to evaluate the different impact of their derivatives. To this purpose, three configurations were examined depending on whether the observed data (model 0), the first derivative (model 1), or the second derivative (model 2) is included as a functional covariate in Eq. (2).

5.1 In-sample results

The fFBreg model (and the fBreg one for comparison purpose) has been specified according to Eq. (2) by choosing the logit transformation as a link function, and has been estimated by means of the Bayesian procedure described in Sect. 3. Tables 4 and 5 synthesize the main results referred to the estimates of the coefficients b_k , $k = 1, \dots, K$, and of the additional parameters, as well as the goodness of fit among competing models.

Looking at the WAIC estimates from both samples (Table 5), the fFBreg model always provides a better fit than the fBreg one, the advantage in fit being particularly pronounced especially for the NeoSpectra sample. Focusing on the NeoSpectra sample, it emerges that the model with the observed functional covariate is to be preferred, in terms of goodness of fit, with respect to competing models with either the first or the second derivatives. Differently, as for the SCiO sample, the fit of the models in the different configurations appears to be comparable although model 2 (i.e., with second derivative as functional covariate) is slightly better.

The graphical representation of the estimated functional parameter $\beta(t)$ is illustrated in Figs. 3 (NeoSpectra sample) and 4 (SCiO sample). The estimated curves provided by the models under examination differ little for the SCiO data, whereas they differ slightly more for the NeoSpectra data. The major differences among the estimated curves are observed for model 0 for the NeoSpectra sample and for model 2 for the SCiO sample. Interestingly, these two configurations are indeed the ones where the models ensure the best fit.

The analysis of the curves can be deepened in order to better understand the impact of the functional parameter in the regression model. By way of example,

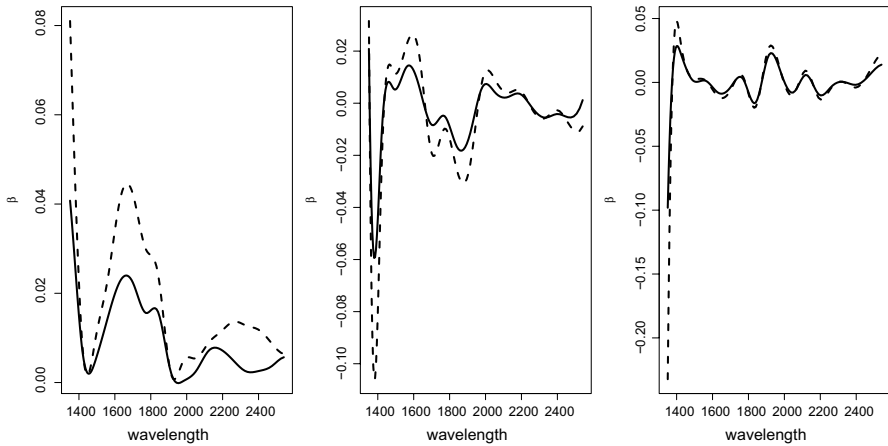


Fig. 3 NeoSpectra data: estimated functional regression coefficients $\beta(t)$ for model 0 (left-hand side panel), model 1 (middle panel), and model 2 (right-hand side panel), and for fBreg (dashed lines) and fFBreg (solid lines) models

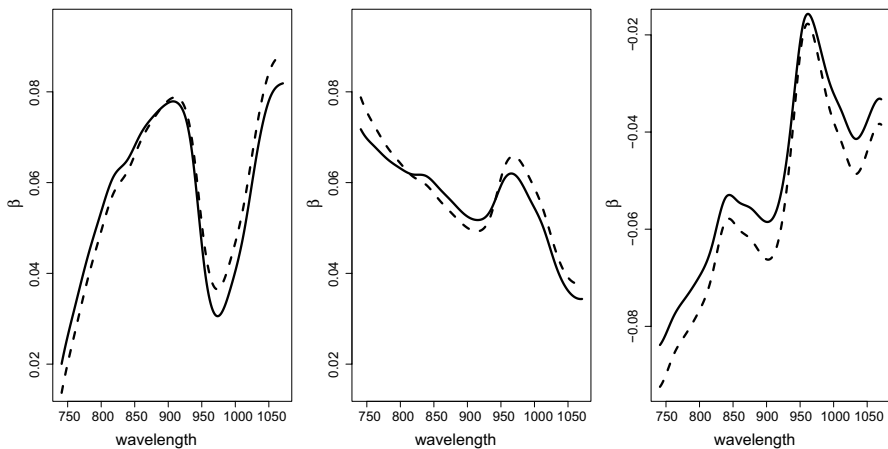


Fig. 4 SCiO data: estimated functional regression coefficient $\beta(t)$ for model 0 (left-hand side panel), model 1 (middle panel), and model 2 (right-hand side panel), and for fBreg (dashed lines) and fFBreg (solid lines) models

focusing on the NeoSpectra data in the model 0 configuration (left-hand side panel of Fig. 3), it emerges that the functional parameters of the competing models have their peaks, and thus the greatest impact on the response, over the wavelengths intervals [1350,1400] and [1600, 1700] nm. In correspondence to these intervals, the observed spectra are indeed more spread, and the partition between skimmed and not skimmed specimens is particularly accentuated, as can be seen in Fig. 1. Similar comments can be made for the other configurations and for the results related to SCiO data as well.

Table 4 Milk data: posterior means and SDs (in parenthesis) of the significant b_k coefficients

	NeoSpectra data		SCiO data	
	fBreg	fFBreg	fBreg	fFBreg
Model 0	$\alpha = -1.774(0.072)$ $b_1 = 0.764(0.064)$ $b_2 = 0.103(0.079)$	$\alpha = -1.766(0.084)$ $b_1 = 0.413(0.033)$ $b_5 = 0.067(0.035)$	$\alpha = -1.907(0.049)$ $b_1 = 1.087(0.077)$ $b_3 = -0.223(0.047)$	$\alpha = -1.922(0.047)$ $b_1 = 1.065(0.066)$ $b_3 = -0.231(0.039)$
Model 1	$\alpha = -1.786(0.066)$ $b_1 = -0.778(0.063)$ $b_6 = -0.098(0.092)$	$\alpha = -1.771(0.068)$ $b_1 = -0.437(0.04)$	$\alpha = -1.883(0.045)$ $b_1 = 1.018(0.066)$ $b_3 = 0.203(0.038)$	$\alpha = -1.82(0.19)$ $b_1 = 0.998(0.17)$ $b_2 = 0.188(0.039)$ $b_3 = 0.160(0.049)$
Model 2	$\alpha = -1.769(0.08)$ $b_1 = -0.733(0.075)$ $b_2 = 0.135(0.093)$	$\alpha = -1.738(0.072)$ $b_1 = -0.366(0.041)$ $b_2 = 0.169(0.029)$ $b_3 = -0.062(0.034)$	$\alpha = -1.893(0.043)$ $b_1 = -0.992(0.066)$ $b_2 = -0.377(0.037)$	$\alpha = -1.77(0.179)$ $b_1 = -0.883(0.158)$ $b_2 = -0.357(0.042)$

Table 5 Milk data: posterior means and SDs (in parenthesis) of the additional parameters and WAIC measures

		NeoSpectra data		SCiO data	
		fBreg	fFBreg	fBreg	fFBreg
Model 0	ϕ	43.106 (9.937)	294.804 (70.298)	170.384 (47.158)	245.513 (66.935)
	p	—	0.782 (0.058)	—	0.637 (0.423)
	w	—	0.822 (0.031)	—	0.655 (0.217)
	WAIC	- 143.574	- 177.424	- 167.51	- 174.097
Model 1	ϕ	44.453 (10.649)	196.065 (43.866)	170.123(46.192)	331.987 (137.5)
	p	—	0.793 (0.06)	—	0.67 (0.337)
	w	—	0.801 (0.038)	—	0.364 (0.241)
	WAIC	- 146.456	- 161.898	- 171.005	- 174.696
Model 2	ϕ	34.223 (8.141)	220.701 (54.066)	196.221 (56.292)	237.295 (77.043)
	p	—	0.796 (0.053)	—	0.43 (0.401)
	w	—	0.822 (0.034)	—	0.441 (0.322)
	WAIC	- 131.563	- 159.542	- 176.602	- 178.15

A closer look into the behaviour of the estimated functional parameter is provided by looking at the posterior estimates of the coefficients in Table 4. Here, only the posterior estimates of the significant coefficient b_k were reported, i.e. those for which the posterior probability of inclusion of that coefficient is greater than 0.5 meaning that $\frac{1}{S} \sum_{s=1}^S I_k^{(s)} > 0.5$. The analysis of the results from Table 4 is of interest because it shows that the most significant estimated PCs in the regression framework are not necessarily the first ones once they have been ordered with respect to the fractional explained variability. Let us focus on the results from the NeoSpectra sample. In model 0, the fFBreg includes the fifth coefficient

in the final model, despite its corresponding PC accounts for only 0.18% of variability of the functional data. Similarly, in model 1 the fBreg includes the sixth coefficient in the final model, despite its corresponding PC accounts for only 1.04% of variability of the first derivative. An analogous phenomenon can be observed for the SCiO sample. In model 0, the third coefficient is selected, corresponding to a PC that only explains 0.11% of the total variability of the covariate, by both the fFBreg and fBreg models. Conversely, the second coefficient is excluded by both of them despite the fact that it accounts for a slightly greater amount of variability. All these findings confirm the soundness of relying on shrinkage prior-based Bayesian variable selection techniques in order to choose the most significant PCs.

To enhance the comparison between the regression models, one could directly compute, from the results reported in Table 4, the ratio of the SD to the absolute value of the corresponding posterior mean as a relative measure of variability. By doing this, the only remarkable finding that emerges is that the lower the fraction of variability explained by a PC, the higher the relative variability of the associated coefficient estimate.

To deepen the analysis, it is of interest to look at the posterior estimates of the additional parameters of the fFBreg (see Table 5). In particular, the estimate of the mixing proportion p and of the normalized distance w allow to understand how the special mixture structure of the FB fits data. Focusing on the NeoSpectra sample, it is worth noting that the fFBreg recognizes two latent groups in all models. The first component has an estimated weight p of around 0.8, and the component means of the two latent groups are quite far from each other (estimated w larger than 0.8). Interestingly, the model dedicates the second component of its mixture to model the group of milk specimens with the lowest level of fat, the one that is classified as “skimmed”.

Differently, by looking at results from the SCiO data, it emerges the presence of two almost equally weighted components with not so far cluster means and quite large SDs. Moreover, the mixture structure identified by the fFBreg model does not find a match into the classification of data into “skimmed” and “not skimmed”.

To better understand the behaviour of the fFBreg model, it is enlightening to look at posterior predictive distributions (Gelman et al. 2014). Generally speaking, the posterior predictive distribution is the distribution of a replicated response \tilde{y}_i under the fitted model, conditional on the observed data. Having simulated S draws from the posterior distributions of $\boldsymbol{\eta}$, sampling from the posterior predictive distribution for each \tilde{y}_i , $i = 1, \dots, n$, is straightforward and works as follows:

- simulate $v^{(s)}$ from a Bernoulli distribution as in 2.(a) of the algorithm in Sect. 3;
- if $v^{(s)} = 1$, simulate $\tilde{y}_i^{(s)} \sim \text{Beta}(\lambda_{1i}^{(s)}, \phi^{(s)})$, i.e., from the first component of the FB mixture;
- otherwise if $v^{(s)} = 0$, simulate $\tilde{y}_i^{(s)} \sim \text{Beta}(\lambda_{2i}^{(s)}, \phi^{(s)})$, i.e., from the second component of the FB mixture.

The vector $(\tilde{y}_i^{(1)}, \dots, \tilde{y}_i^{(S)})$ represents a sample from the posterior predictive distribution.

The 95% confidence bounds of the posterior predictives for the specimens under model 0 are reported in Fig. 5 (but similar results are obtained for models 1 and 2). Looking at the plots concerning the SCiO sample (bottom panels), it emerges how the confidence bounds of the fBreg and fFBreg models are almost coincident. Thus, this is a typical scenario where the latent mixture structure observed in the bounded response is fully explained by the covariate and, as a result, the fBreg and fFBreg perform similarly. In contrast, the scenario that emerges from the NeoSpectra sample is completely different. The confidence bounds of the fBreg model (top left-hand panel) are very wide, showing the worst posterior predictions for the skimmed specimens. Instead, posterior predictive bounds of the fFBreg model (top right-hand

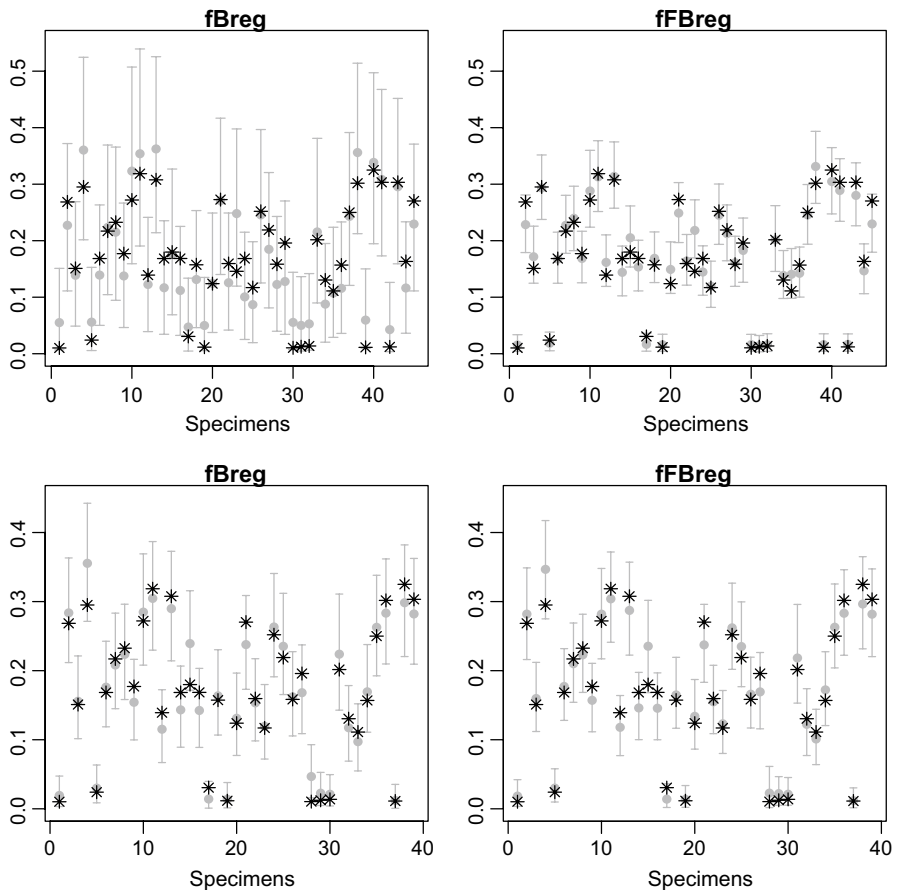


Fig. 5 Posterior predictive confidence bounds at 95% for fBreg (left-hand panels) and fFBreg (right-hand panels) models from NeoSpectra data (top panels) and SCiO data (bottom panels). The asterisks represent the observed response values whereas the filled points are the corresponding average posterior predictive values

panel) are tight and well-centred around the observed values, particularly for the skimmed specimens. Thus, this evidence suggests that the covariate alone is not capable of fully accounting for the latent mixture structure of the bounded response. The choice of the fFBreg model, in this case, is fundamental since it allows to recover the latent mixture structure, greatly increasing the overall fit with respect to non-mixture models such as beta-type ones.

Finally, to appreciate the ability of the proposed methodology in recovering the special features of the densities of the responses (see Fig. 2), the estimated densities of the average posterior predictives are plotted in Fig. 6 and superimposed on the original ones. All the estimates are computed by using the same kernel and bandwidth. Concerning the NeoSpectra sample (left-hand side panel), it is worth noting that the entire original density is fully captured by the fFBreg model, whereas the fBreg model identifies neither the left heavy tail nor the mode. Conversely, in the SCiO sample (right-hand side panel) both models are able to reproduce these patterns.

5.2 Out-of-sample results

In this applicative context, a key feature of the model to be preferred concerns its predictive ability. For this reason, a cross-validation analysis was performed. In particular, since the sample size is small, a leave-one-out cross-validation LOO-CV procedure was implemented, meaning that at each iteration of the cross-validation algorithm one specimen of milk was used as test set while all other specimens were used as training set.

First, it is natural to compare the predictive ability of the fFBreg model to the one provided by the fBreg. As it has been already noted above, to the best of the authors' knowledge, there is a lack of regression models specifically designed to cope simultaneously with a bounded continuous response and functional covariate.

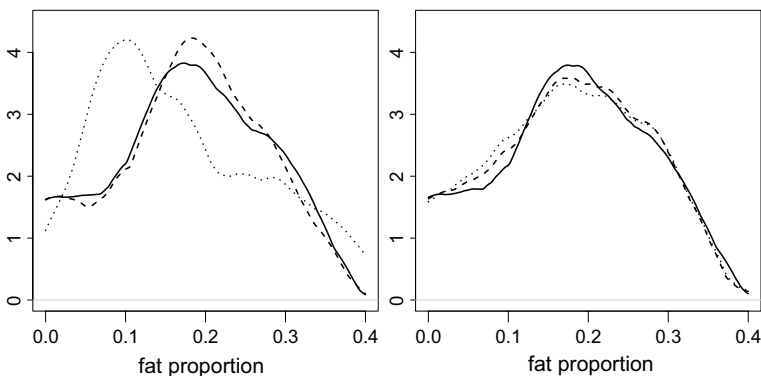


Fig. 6 Estimated densities (based on the Epanechnikov kernel with bandwidth equals 0.04) of the observed fat proportion (solid lines) and of the average posterior predictives for fBreg (dotted lines) and fFBreg (dashed lines) models from NeoSpectra data (left-hand side panel) and SCiO data (right-hand side panel)

Despite this, it is still of interest to try to make comparisons with existing models in terms of predictive ability. To this end, a GFLM (Ramsay and Silverman 2005) and a non parametric (NP) scalar-on-function regression model are evaluated (Ferraty and Vieu 2006), having preliminarily performed a *logit* transformation of the bounded response. The estimation of these models is done through the use of the *fda.usc* package in the R software.

The predictive accuracy of a model is measured by computing the inverted balance relative error (IBRE) (Tofallis 2015), a normalised index equal to:

$$IBRE = \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \hat{y}_i^{(-i)})^2}{\max(y_i, \hat{y}_i^{(-i)})}, \tag{7}$$

where $\hat{y}_i^{(-i)}$ is the predicted value of the i -th observation based on the estimates obtained when the i -th data point is removed.

Results are reported in Table 6: original values have been multiplied by 1000 to improve readability. Reading the table results can be twofold. On the one hand, it is possible to assess the best functional covariate (among the observed data, the first derivative, and the second derivative) in terms of accuracy of the prediction within each regression model. On the other hand, one can determine the best regression model, in terms of prediction accuracy, among the four alternatives at hand.

Focusing on the NeoSpectra sample, it is worth noting that the fBreg, GFLM, and NP models perform better with model 0, whereas fFBreg model predicts slightly better with model 1. By comparing the competing models, focusing on the results for model 0 it emerges that the worst model is the GFLM, followed closely by the fBreg one. Conversely, the NP is the best model in terms of predictive accuracy closely followed by the fFBreg model. Results for model 1 lead to analogous conclusions. Note also that for model 2, which is the best among the three models according to WAIC of both fBreg and fFBreg, the best accuracy in prediction is provided by far by the fFBreg model.

Focusing on the SciO sample, all regression models perform similarly whatever the functional covariate included in the model. The fBreg and NP models have a slightly better performance with model 0, whereas the fFBreg and GFLM perform slightly better with model 1. The comparison among competing models shows that

Table 6 Predictive accuracy of the competing models based on LOO-CV for the NeoSpectra and SciO samples

	NeoSpectra data				SciO data			
	fBreg	fFBreg	GFLM	NP	fBreg	fFBreg	GFLM	NP
Model 0	13.507	5.506	19.806	3.357	4.511	3.940	7.057	3.158
Model 1	15.539	4.154	24.603	3.809	6.029	3.160	6.103	7.024
Model 2	17.587	4.478	21.585	14.527	4.853	3.805	7.855	7.353

Results have been multiplied by 1000 to improve readability

the worst performance is always provided by the GFLM, whereas the fFBreg and NP models are preferable in terms of predictive accuracy. In particular, looking at results for model 0, the best model in terms of predictive accuracy is the NP one, while looking at results for model 1 and model 2 it is the fFBreg model that outperforms all competing models.

Overall, it should be noted that the fFBreg shows a predictive accuracy that is comparable to the one provided by the NP model, only sometimes being slightly worse. However, the tiny advantage in terms of prediction provided by a non parametric method comes with a great cost in terms of interpretability of the regression function.

Finally, it is worth noting that similar findings would have been found using alternative prediction measures such as the root mean square predictive error (MSPE) [that is the root square of the numerator of (7)], or its relative counterpart, that is the root of the MSPE divided by the sample deviance.

6 Concluding remarks

The fFBreg model is proposed to simultaneously handle a bounded response and a functional covariate in a regression framework. The estimation issue is performed according to a Bayesian rationale. Moreover, a bases representation strategy is adopted to operationalise the linear specification of the conditional mean of the response. This approach presents the problem of determining how many and which real coefficients are significant in the regression model. The adopted solution takes advantage of a Bayesian variable selection strategy, consistent with the Bayesian approach, that exploits spike-and-slab priors. Several Monte Carlo studies enabled the comparison between the proposed fFBreg model and a more standard fBreg one, as well as the inspection of the goodness of the estimates. The Metropolis within Gibbs sampling algorithm behaves well in all configurations, leading to chains that fulfill the convergence checks. Moreover, the fFBreg model shows a satisfactory behaviour in all the considered scenarios. Specifically, when the response generating mechanism is of beta-type, the fFBreg has a similar fit than the fBreg model. Conversely, when the response generating mechanism is of mixture-type, with groups that are from moderately to highly separated, the fFBreg model outperforms the fBreg one. Therefore, it emerges that the fFBreg model should always be the preferred, despite being more complex, even when the response does not show some typical features such as heavy tails and/or multimodality.

Finally, the proposed regression model is applied to a real spectrometric example that also motivated the work. Interestingly, the observations from the simulation studies find full application in the example. Indeed, in the SCiO sample the left heavy tail observed in the response is fully explained by the functional covariate, and hence the fFBreg and fBreg perform similarly and very well both in terms of fit and prediction accuracy [these results remind of case (1) of simulation study]. Conversely, in the NeoSpectra sample the latent mixture structure of the bounded response is not entirely explained by the functional covariate. As a result, the fBreg performs poorly, whereas the special mixture structure of the

fFBreg model proves to be essential in the goodness of fit of the model [these results resemble case (3) of simulation study].

It is worth noting that a generic mixture of beta distributions, despite being potentially more flexible than the FB one, would not be identifiable, thus resulting in severe computational issues.

The introduced model can be extended in different manners. One of the most interesting contemplates the possibility of using more real and/or functional covariates (e.g. one could consider the raw functional data and at the same time some of its derivatives). This type of extension can be managed by carefully redefining the covariates space that would leads to a change in the definition of the internal product appearing in (2) as well as the covariance operator and the associated principal components.

Acknowledgements The authors acknowledge Riu J., Gorla G., Chakif D., Boqué R., and, in particular, Giussani B. for providing access to the spectrometric dataset. The financial support of Università del Piemonte Orientale is acknowledged by the first three authors. The first three authors are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). The authors thank two anonymous reviewers and the associate editor for their constructive comments on an earlier version of the paper. The authors thank Lax Chan for a careful linguistic review.

Funding Open access funding provided by Università degli Studi del Piemonte Orientale Amedeo Avogadro within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aneiros G, Vieu P (2016) Sparse nonparametric model for regression with functional covariate. *J Nonparametr Stat* 28(4):839–859. <https://doi.org/10.1080/10485252.2016.1234050>
- Crainiceanu CM, Goldsmith AJ (2010) Bayesian functional data analysis using WinBUGS. *J Stat Softw* 32:1–33
- Ferrari S, Cribari-Neto F (2004) Beta regression for modelling rates and proportions. *J Appl Stat* 31(7):799–815. <https://doi.org/10.1080/0266476042000214501>
- Ferraty F, Vieu P (2002) The functional nonparametric model and application to spectrometric data. *Comput Stat* 17(4):545–564. <https://doi.org/10.1007/s001800200126>
- Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer series in statistics. Springer, New York
- Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D (2014) Bayesian data analysis. CRC Press, London
- Goia A, Vieu P (2015) A partitioned single functional index model. *Comput Stat* 30(3):673–692. <https://doi.org/10.1007/s00180-014-0530-1>

- Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, Reich D (2011) Penalized functional regression. *J Comput Graph Stat* 20(4):830–851. <https://doi.org/10.1198/jcgs.2010.10007>
- Greven S, Scheipl F (2017) A general framework for functional regression modelling. *Stat Model* 17(1–2):1–35. <https://doi.org/10.1177/1471082X16681317>
- Horvath L, Kokoszka P (2012) Inference for functional data with applications. Springer series in statistics. Springer, New York
- Jolliffe IT (1982) A note on the use of principal components in regression. *J R Stat Soc Ser C (Appl Stat)* 31(3):300–303. <https://doi.org/10.2307/2348005>
- Kokoszka P, Reimherr M (2017) Introduction to functional data analysis. Chapman and Hall, Boca Raton
- Ling N, Vieu P (2018) Nonparametric modelling for functional data: selected survey and tracks for future. *Statistics* 52(4):934–949. <https://doi.org/10.1080/02331888.2018.1487120>
- Lunn D, Spiegelhalter D, Thomas A, Best N (2009) The BUGS project: evolution, critique and future directions. *Stat Med* 28(25):3049–3067. <https://doi.org/10.1002/sim.3680>
- Malloy EJ, Morris JS, Adar SD, Suh H, Gold DR, Coull BA (2010) Wavelet-based functional linear mixed models: an application to measurement error-corrected distributed lag models. *Biostatistics* 11(3):432–452. <https://doi.org/10.1093/biostatistics/kxq003>
- Migliorati S, Di Brisco AM, Ongaro A (2018) A new regression model for bounded responses. *Bayesian Anal* 13(3):845–872. <https://doi.org/10.1214/17-BA1079>
- O'Hara RB, Sillanpää MJ (2009) A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal* 4(1):85–117. <https://doi.org/10.1214/09-BA403>
- Ramsay J, Silverman BW (2005) Functional data analysis. Springer series in statistics. Springer, New York
- Reiss PT, Goldsmith J, Shang HL, Ogden RT (2017) Methods for scalar-on-function regression. *Int Stat Rev* 85(2):228–249
- Riu J, Gorla G, Chakif D, Boqué R, Giussani B (2020) Rapid analysis of milk using low-cost pocket-size NIR spectrometers and multivariate analysis. *Foods*. <https://doi.org/10.3390/foods9081090>
- Robert C, Casella G (1999) Monte Carlo statistical methods. Springer texts in statistics. Springer, New York
- Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. *J Am Stat Assoc* 82(398):528–540. <https://doi.org/10.1080/01621459.1987.10478458>
- Tofallis C (2015) A better measure of relative prediction accuracy for model selection and model estimation. *J Oper Res Soc* 66:1352–1362. <https://doi.org/10.1057/jors.2014.103>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.