Advances in Data Analysis and Classification (2023) 17:745–775
https://doi.org/10.1007/s11634-022-00523-5

REGULAR ARTICLE

Semiparametric finite mixture of regression models with Bayesian P-splines

Marco Berrettini[1] · **Giuliano Galimberti**[1] · **Saverio Ranciati**[1]

Received: 8 November 2021 / Revised: 21 September 2022 / Accepted: 2 October 2022 /
Published online: 18 October 2022

Abstract
Mixture models provide a useful tool to account for unobserved heterogeneity and are at the basis of many model-based clustering methods. To gain additional flexibility, some model parameters can be expressed as functions of concomitant covariates. In this Paper, a semiparametric finite mixture of regression models is defined, with concomitant information assumed to influence both the component weights and the conditional means. In particular, linear predictors are replaced with smooth functions of the covariate considered by resorting to cubic splines. An estimation procedure within the Bayesian paradigm is suggested, where smoothness of the covariate effects is controlled by suitable choices for the prior distributions of the spline coefficients. A data augmentation scheme based on difference random utility models is exploited to describe the mixture weights as functions of the covariate. The performance of the proposed methodology is investigated via simulation experiments and two real-world datasets, one about baseball salaries and the other concerning nitrogen oxide in engine exhaust.

Keywords Mixture of experts models · Gibbs sampling · Data augmentation

Mathematics Subject Classification 62H30

1 Introduction

Regression analysis represents one of the most popular tool to investigate the effect of a set of regressors/covariates on a dependent variable. In this context, a (possibly linear) regression model is usually specified to describe the conditional expected value of the dependent variable given the values of the regressors. When data come from different

✉ Marco Berrettini
marco.berrettini2@unibo.it

[1] Department of Statistical Sciences, University of Bologna, Bologna, Italy

subpopulations, it may be reasonable to assume that the unknown parameters of such model may vary across these subpopulations. Finite Mixtures of Regression (FMR) models deal with this kind of data, whenever the information about subpopulation membership is missing (i.e., when subpopulation membership is a source of unobserved heterogeneity). Since their introduction (Goldfeld and Quandt 1973), mixtures of regression models have been extensively employed in many research fields (see, for example, Wedel and DeSarbo 1993; Wang et al. 1996; Turner 2000; Green and Richardson 2002; Ding 2006; Tashman and Frey 2009; Dyer et al. 2012; Van Horn et al. 2015; McDonald et al. 2016).

According to their basic formulation, FMR models are characterised by the so-called assignment independence: namely, it is assumed that subpopulation membership does not depend on the regressors. Finite Mixtures of Regression models with Concomitant covariates (FMRC), also known as mixtures of experts models (Jacobs et al. 1991), overcome this limitation by specifying not only the component conditional expected values but also the component weights as functions of two (sub)sets of regressors, which can be disjoint, coinciding, or overlapping. In particular, a multinomial logistic regression structure is commonly chosen to link the component weights to the regressors. Applications of FMRC models are described in the statistical, econometric and machine learning literature (see, for example, Weigend and Shi 2000; Lu 2006; Gormley and Murphy 2008; Villani et al. 2009; Lê Cao et al. 2010; Li et al. 2010, 2011; Frühwirth-Schnatter et al. 2012; Gormley and Frühwirth-Schnatter 2019; Murphy and Murphy 2020). It is worth mentioning that some of these applications consider multivariate regressors and/or multivariate dependent variables. Alternatively, Xu et al. (1994) and Ingrassia et al. (2012) show how assignment dependence in the conditional distribution of the dependent variable can be achieved by resorting to cluster-weighted models (Gershenfeld 1997). This latter approach, however, requires the specification of the joint distribution of both dependent variable and regressors (which is typically assumed to be a mixture, whose components are represented as the product between a component conditional distribution for the dependent variable and a component marginal distribution for the regressors).

In order to enhance the flexibility of FMR/FMRC models, recently several authors have focused their attention on providing semiparametric or nonparametric extensions of such models (see Xiang et al. 2019, for a recent review). In the context of models with Gaussian components, Young and Hunter (2010) and Huang and Yao (2012) have suggested FMRC models where the component weights are assumed to be smooth functions of a univariate covariate, while retaining a linear structure for the conditional expected values. This latter assumption has been relaxed by Huang et al. (2013), who have considered models with conditional expected values and conditional variances allowed to vary smoothly according to the value of a covariate. Furthermore, Xiang and Yao (2018) and Zhang and Zheng (2018) have proposed a semiparametric representation of the conditional expected values. It is worth noting that in all the papers just mentioned, estimation has been carried out through modified versions of the Expectation-Maximization (EM) algorithm.

To the best of the Authors' knowledge, none of these flexible models have been examined from a Bayesian perspective, despite the fact that Bayesian algorithms to estimate their parametric counterparts have been extensively studied in the literature

(see, for example, Frühwirth-Schnatter 2006; Gormley and Frühwirth-Schnatter 2019). The aim of this Paper is to fill this gap by considering semiparametric FMRC models within the Bayesian framework. In particular, this Paper focuses on models where the log-odds of component weights and the conditional means are smooth functions of a univariate covariate. Following the approach detailed in Berrettini et al. (2021), Bayesian P-splines (Lang and Brezger 2004) are exploited to obtain a parsimonious representation of these smooth functions, and a new Gibbs sampler algorithm is developed to perform inference based on: (i) an adaption of a data augmentation scheme with a (partial) difference Random Utility Model (dRUM) representation; (ii) an approximation of the logistic distribution through a Gaussian mixture (Frühwirth-Schnatter et al. 2012). Differently from Berrettini et al. (2021), where mixtures of multinomials are considered and smoothness is only allowed on the component weights, whereas the other parameters are assumed constant, in this Paper:

- models for the conditional distribution of continuous dependent variables are examined,
- and the component means are also assumed to be a function of the covariate.

The remainder of the Paper is organised as follows. Model specification is provided in Sect. 2, while in Sect. 3 the associated Bayesian inference procedure is elicited; results from simulation studies are presented in Sect. 4, while the ones about real data applications are reported in Sect. 5. Finally, Sect. 6 is devoted to discussion and conclusions.

2 Model specification

Suppose $\{y_i\}, i = 1, \ldots, n$, is a random sample from a population clustered into G components, and that each observation i has an associated quantitative covariate x_i. For simplicity, both y_i and x_i are assumed univariate throughout this Paper. Let $c_i \in \{1, \ldots, G\}$ be the component indicator for the i-th unit having discrete distribution $\Pr(c_i = g|x_i) = p_g(x_i) > 0$, for $g = 1, \ldots, G$, such that $\sum_{g=1}^{G} p_g(x_i) = 1$, for $i = 1, \ldots, n$. In addition, suppose that, conditioning on c_i and x_i, y_i follows a Gaussian distribution with mean $\mu_{c_i}(x_i)$ and variance $\sigma^2_{c_i}$. It is further assumed that each $\mu_g(\cdot)$ is an unknown smooth function of the covariate x. Hence, given x_i, the random variable y_i follows a finite mixture of Gaussian components:

$$f(y_i|x_i) = \sum_{g=1}^{G} p_g(x_i) f_{\mathcal{N}}\left(\mu_g(x_i), \sigma^2_g\right), \tag{1}$$

where $f_{\mathcal{N}}(\cdot)$ denotes the Gaussian density function, and $p_1(\cdot), \ldots, p_G(\cdot)$ can be referred to as the component (or mixture) weights. Conditions for identifiability of Model (1), whose corresponding graphical representation is reported in Fig. 1, can be derived by Theorem 1 in Huang et al. (2013) by taking into account that each variance σ^2_g is assumed independent – and, thus, constant—with respect to covariate x:

- $p_g(x) > 0$ are continuous functions and $\mu_g(x)$ are differentiable functions, for $g = 1, \ldots, G$;

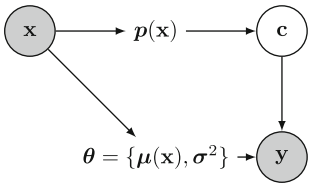


Fig. 1 Graphical model representation of the FMRC model in Eq. (1); grey-colored circle represent observed quantities

- for any $h = j$, $\sigma_h^2 \neq \sigma_j^2$, or, if there exists $h = j$ such that $\sigma_h^2 = \sigma_j^2$, then $||\mu_h(x) - \mu_j(x)|| + ||\mu_h'(x) - \mu_j'(x)|| \neq 0$ for any value of x;
- the domain $\mathcal{X}$ of x is an interval in $\mathbb{R}$.

Jacobs et al. (1991) model the component weights $p_g(x_i)$ using a multinomial logistic regression model, expressing the log-odds of these probabilities, with respect to the reference one (e.g., the G-th), as linear functions of the covariate x. In this Paper, similarly to Berrettini et al. (2021), each of these $G - 1$ linear predictors is replaced with a smooth function of x, represented by a linear combination of m cubic B-spline bases $B_\rho(\cdot)$ and coefficients $\gamma_{g\rho}$:

$$\log \frac{p_g(x_i)}{p_G(x_i)} = \eta_g(x_i) = \sum_{\rho=1}^{m} B_\rho(x_i)\gamma_{g\rho}, \quad \text{for } i = 1, \ldots, n. \tag{2}$$

By defining the $n \times m$ design matrix $\mathbf{B}$, where the element in row i and column ρ is given by $B_\rho(x_i)$, and after some algebra, Eq. (2) can be rewritten as:

$$\boldsymbol{p}_g(\mathbf{x}) = \frac{\exp(\mathbf{B}\boldsymbol{\gamma}_g)}{\sum_{g=1}^{G} \exp(\mathbf{B}\boldsymbol{\gamma}_g)}, \tag{3}$$

where $\boldsymbol{\gamma}_g = (\gamma_{g1}, \ldots, \gamma_{gm})'$ corresponds to the vector of unknown regression coefficients, where the exponential is applied elementwise. To guarantee identifiability, the vector of coefficients corresponding to the reference group G are all set equal to 0.

Regarding the components' normal densities, each mean $\mu_g(\cdot)$ is also assumed to be an unknown smooth function of covariate x, represented through B-splines:

$$\boldsymbol{\mu}_g(\mathbf{x}) = \mathbf{B}\boldsymbol{\beta}_g \tag{4}$$

with $\boldsymbol{\beta}_g = (\beta_{g1}, \ldots, \beta_{gm})'$ vector of unknown regression coefficients.

3 Bayesian inference

Adopting B-splines to represent a smooth function requires the specification of what is known as the number of knots (or, equivalently, the number of B-spline bases), which governs how the bases behave and the flexibility of the resulting function. In the Bayesian framework, Lang and Brezger (2004) suggest a large number of knots (between 20 and 40) to ensure enough flexibility; additionally, they show how to

define priors for the regression parameters $\gamma_{g1}, \ldots, \gamma_{gm}$ and $\beta_{g1}, \ldots, \beta_{gm}$ in terms of a random walk:

$$\gamma_{g\rho} = \gamma_{g,\rho-1} + w_{g\rho}, \quad w_{g\rho} \sim N(0, \delta_g^2); \tag{5}$$

$$\beta_{g\rho} = \beta_{g,\rho-1} + u_{g\rho}, \quad u_{g\rho} \sim N(0, \tau_g^2). \tag{6}$$

Using this representation is equivalent to inducing a penalisation, based on differences of adjacent B-spline coefficients, and leads to the definition of "penalised" B-splines, commonly abbreviated to "P-splines". Through this approach, the amount of smoothness is controlled by the variance parameters δ_g^2 and τ_g^2: their presence protects against possible overfitting when a larger than needed number of knots is chosen. In particular, small values for δ_g^2 and τ_g^2 lead to approximately constant log-odds and conditional mean, respectively. Hyperpriors are assigned to the variances τ_g^2, δ_g^2 and σ_g^2, selecting Inverse Gamma distributions IG(a, b), with $a = 1$ and a small value for b, for example $b = 0.005$, leading to almost diffuse priors. The priors in (5) and (6) can be equivalently written in the form of global smoothness priors:

$$\boldsymbol{\gamma}_g|\delta_g^2 \propto \exp\left(-\frac{1}{2\tau_g^2}\boldsymbol{\gamma}_g'\mathbf{K}\boldsymbol{\gamma}_g\right), \quad \boldsymbol{\beta}_g|\tau_g^2 \propto \exp\left(-\frac{1}{2\tau_g^2}\boldsymbol{\beta}_g'\mathbf{K}\boldsymbol{\beta}_g\right)$$

where the penalty matrix $\mathbf{K}$ is given by $\mathbf{K} = \boldsymbol{\Delta}_1'\boldsymbol{\Delta}_1$, with $\boldsymbol{\Delta}_1$ being the first order difference matrix (Rue and Held 2005, Chapter 2, p. 52). Because $\mathbf{K}$ is rank deficient with rank$(\mathbf{K}) = m - 1$ for a first-order random walk, the priors are improper. It is worth mentioning that in the literature these kind of priors are usually referred to as intrinsic Gaussian Markov random fields (Rue and Held 2005).

The multinomial model in Equation (2) can be conveniently represented as a binary formulation in the partial dRUM representation proposed by Frühwirth-Schnatter et al. (2012). Conditional on each $\lambda_g(x_i) = \exp(\eta_g(x_i))$, the random utilities are defined as

$$z_{gi} = \eta_g(x_i) - \log\left(\sum_{l \neq g} \lambda_l(x_i)\right) + \epsilon_{gi}, \quad D_{gi} = \mathbb{1}(z_{gi} > 0); \tag{7}$$

where z_{gi} are latent variables, $D_{gi} = \mathbb{1}(c_i = g)$ are the allocation indicators and ϵ_{gi} are i.i.d. errors following a Logistic distribution ($g = 1, \ldots, G, i = 1, \ldots, n$). Given $\lambda_1(x_i), \ldots, \lambda_G(x_i)$ and the latent indicator variables $D_{1i}, \ldots, D_{Gi}$, the latent variables $z_{1i}, \ldots, z_{Gi}$ are distributed according to an Exponential distribution and can be easily sampled in a data augmented implementation. To avoid any Metropolis-Hastings step, Frühwirth-Schnatter and Frühwirth (2010) approximate, for each ϵ_{gi}, the Logistic distribution by a finite scale mixture of H Gaussian distributions, with zero means and variances $\{s_1^2, \ldots, s_H^2\}$ drawn with fixed probabilities $\{w_1, \ldots, w_H\}$. The same authors obtained their finite scale mixture approximation by minimizing the Kullback-Leibler divergence between the densities, and recommend choosing $H = 3$ in larger applications, where computing time matters, and to work with $H = 6$

whenever possible. In a second step of data augmentation, the component indicators r_{gi} $(g = 1, \ldots, G-1, i = 1, \ldots, n)$, each taking value $h = 1, \ldots, H$, are introduced as yet another level of latent variables. Conditional on $\mathbf{z}_g = (z_{g1}, \ldots, z_{gn})'$ and $\mathbf{r}_g = (r_{g1}, \ldots, r_{gn})'$, the binary logit regression model (7) reduces to a Gaussian regression model.

3.1 MCMC algorithm

Based on the representation of Sect. 3, a new MCMC algorithm is implemented, for a fixed G, by integrating the scheme proposed by Frühwirth-Schnatter et al. (2012) with the Bayesian P-spline approach by Lang and Brezger (2004), similarly to Berrettini et al. (2021). A sketch of the algorithm is comprised of the following steps:

1. Sample the regression coefficients' vector $\boldsymbol{\gamma}_g$ conditional on $\mathbf{z}_g$ and $\mathbf{r}_g$, $g = 1, \ldots, G-1$. Using the prior in Equation (5), the full conditional of $\boldsymbol{\gamma}_g$ is given by a multivariate Gaussian density. Straightforward calculations (Brezger and Lang 2006) show that the precision matrix $\mathbf{P}_g$ and the mean $\mathbf{m}_g$ of $\boldsymbol{\gamma}_g|\cdot$ are given by

$$\begin{aligned} \mathbf{P}_g &= \mathbf{B}'\mathbf{W}_g^{-1}\mathbf{B} + \frac{1}{\delta_g^2}\mathbf{K}, \\ \mathbf{m}_g &= \mathbf{P}_g^{-1}\mathbf{B}'\mathbf{W}_g^{-1}\left(\mathbf{z}_g + \log \boldsymbol{\lambda}_{-g}(\mathbf{x})\right), \end{aligned} \tag{8}$$

   respectively, where $\mathbf{W}_g$ is a $n \times n$ diagonal matrix with nonzero elements equal to the randomly drawn variances $(\omega_{1g} = s^2_{r_{g1}}, \ldots, \omega_{ng} = s^2_{r_{gn}})$ for the g-th group, the i-th element of $\boldsymbol{\lambda}_{-g}(\mathbf{x})$ is $\sum_{l \neq g} \lambda_l(x_i)$ and the logarithm is applied elementwise;
2. Sample the $G-1$ variance parameters δ_g^2 conditional on $\boldsymbol{\gamma}_g$:

$$\delta_g^2|\boldsymbol{\gamma}_g \sim IG\left(a + \frac{\text{rank}(\mathbf{K})}{2}, b + \frac{1}{2}\boldsymbol{\gamma}_g'\mathbf{K}\boldsymbol{\gamma}_g\right); \tag{9}$$

3. For each unit $i = 1, \ldots, n$, sample all the (partial) differences of utilities $z_{1i}, \ldots, z_{G-1,i}$ simultaneously from:

$$z_{gi} = \log\left(\frac{\lambda_g(x_i)}{\sum_{l \neq g}\lambda_l(x_i)}U_{gi} + D_{gi}\right) - \log\left(1 - U_{gi} + \frac{\lambda_g(x_i)}{\sum_{l \neq g}\lambda_l(x_i)}D_{gi}\right), \tag{10}$$

   with $U_{gi} \sim \text{Unif}(0, 1)$;
4. For $g = 1, \ldots, G-1$ and $i = 1, \ldots, n$, sample the component indicators r_{gi} conditional on z_{gi} from:

$$\Pr(r_{gi} = h|z_{gi}, \boldsymbol{\gamma}_g) \propto \frac{w_h}{\sqrt{s_h^2}} \exp\left[-\frac{1}{2}\left(\frac{z_{gi} - \eta_g(x_i) + \log\sum_{l \neq g}\lambda_l(x_i)}{\sqrt{s_h^2}}\right)^2\right]; \tag{11}$$

5. Sample the regression coefficients' vector $\boldsymbol{\beta}_g$, $g = 1, \ldots, G$ from a multivariate Gaussian density with covariance matrix $\mathbf{V}_g$ and mean $\boldsymbol{v}_g$

$$\mathbf{V}_g = \left(\frac{1}{\sigma_g^2} \mathbf{B}^{(g)'} \mathbf{B}^{(g)} + \frac{1}{\tau_{gj}^2} \mathbf{K}_j \right)^{-1}, \quad \boldsymbol{v}_g = \mathbf{V}_g \mathbf{B}^{(g)'} \mathbf{y}^{(g)}, \tag{12}$$

where the superscript $^{(g)}$ is applied throughout this section to any matrix or vector to indicate the rows of that matrix (or the elements of that vector) corresponding to the units allocated to the g-th group;

6. Sample the G variance parameters τ_g^2 conditional on $\boldsymbol{\beta}_g$:

$$\tau_g^2 | \boldsymbol{\beta}_g \sim IG \left(a + \frac{\text{rank}(\mathbf{K})}{2}, b + \frac{1}{2} \boldsymbol{\beta}_g' \mathbf{K} \boldsymbol{\beta}_g \right) \tag{13}$$

7. Sample the G variance parameters σ_g^2 conditional on $\boldsymbol{\mu}_g(\mathbf{x}) = \mathbf{B}\boldsymbol{\beta}_g$:

$$\sigma_g^2 | \boldsymbol{\mu}_g(\mathbf{x}), \sim IG \Big(a + \frac{\sum_{i=1}^n D_{gi}}{2}, b + \frac{1}{2} \left(\mathbf{y}^{(g)} - \boldsymbol{\mu}_g^{(g)}(\mathbf{x}) \right)' \left(\mathbf{y}^{(g)} - \boldsymbol{\mu}_g^{(g)}(\mathbf{x}) \right) \Big). \tag{14}$$

8. Classify each unit i according to Bayes' rule: draw D_{gi} ($g = 1, \ldots, G$, $i = 1, \ldots, n$) from the following discrete probability distribution which combines the likelihood and the prior:

$$\Pr(D_{gi} = 1 | y_i, x_i, \boldsymbol{\gamma}, \boldsymbol{\beta}_g, \sigma_g^2) \propto \frac{\lambda_g(x_i)}{\sum_{g=1}^G \lambda_g(x_i)} f_{\mathcal{N}}(y_i | \mu_g(x_i), \sigma_g^2). \tag{15}$$

It is worth mentioning that Steps 1 to 4 of this algorithm are similar to those proposed by Berrettini et al. (2021) to sample the parameters related to the component weights, while Steps 3 to 5 are specific to the models considered in this Paper, and are needed to sample the parameters associated with the component means.

3.2 Label switching, posterior inference and model selection

The MCMC algorithm described in the previous section can be prone to label switching (see Frühwirth-Schnatter 2006, Section 3.5 for a review). A possible solution to deal with this problem, which exploits k-means clustering (with G clusters) of the posterior draws to identify a unique labeling, has been proposed by Frühwirth-Schnatter et al. (2012). This solution is readily available to the semiparametric FMRC models introduced in this Paper.

Posterior inference is carried out after completing the prefixed number T of iterations. As usual, parameter's posterior mean can be computed considering averages of the last $T - T_0$ draws of the chains, with T_0 defining the burn-in phase. As far

as the smooth functions are concerned, posterior quantities are obtained by exploiting their representation as linear combinations of spline bases and the corresponding regression coefficients' estimates. Pointwise percentiles (usually 2.5–97.5 or 5–95) computed over the last $T - T_0$ posterior draws can be used to quantify uncertainty associated to the estimated smooth functions.

The Maximum-A-Posteriori (MAP) rule is adopted to partition the observations into G groups, by allocating them to the G components. In particular, each unit $i = 1, \ldots, n$ is assigned to the component $\hat{c}_i$ such that

$$\hat{c}_i = \arg\max_g \left(\sum_{t=T_0}^{T} D_{1i}^{(t)}, \ldots, \sum_{t=T_0}^{T} D_{gi}^{(t)}, \ldots, \sum_{t=T_0}^{T} D_{Gi}^{(t)} \right). \tag{16}$$

Occasionally, the use of the MAP rule can lead to empty groups, when one or more components could have no units assigned to them. In such situations, it might be worth distinguishing between the number of components G and the number of nonempty components, denoted as

$$\tilde{G} = \sum_{g=1}^{G} \mathbb{1}\left(\hat{n}_g > 0\right), \tag{17}$$

where $\hat{n}_g = \sum_{i=1}^{n} \mathbb{1}(\hat{c}_i = g)$ is the number of observations assigned to group g $(g = 1, \ldots, G)$.

A relevant issue related to mixture models is the choice of the number of components, which originated many efforts in the statistical literature. The proposed MCMC algorithm requires the value of G to be fixed in advance. Thus, the algorithm should be run for different values of G and the obtained results should be compared in order to select the optimal number of components. Several model selection criteria are available to perform these comparisons (see Celeux et al. 2019, for a recent review). Many of these criteria require the determination of the number of free parameters of each candidate model. However, the quantification of this number for the semiparametric FMRC models described in this Paper can be difficult due to the regularisation induced by the prior distributions on the spline coefficients. A solution to circumvent this problem is proposed by Raftery et al. (2007). They suggest the use of $2s_l^2$ as an estimate of this unknown quantity, where s_l^2 is the sample variance of the log-likelihoods computed as $l^{(t)} = \sum_{i=1}^{n} \log f(\mathbf{y}_i | \boldsymbol{\theta}_{\mathbf{D}_i}^{(t)})$, with $\boldsymbol{\theta}_{\mathbf{D}_i}^{(t)}$ denoting the vector of estimated parameters for the component unit i is allocated to, at iterations $t = T_0, \ldots, T$, after the burn-in. Using this estimate, they derive two model selection criteria, whose values depend only on the log-likelihoods from the posterior simulation, that are readily available:

$$\text{AICM} = -2(\bar{l} - s_l^2), \tag{18}$$

$$\text{BICM} = -2\left[\bar{l} - s_l^2(\log(n) - 1)\right], \tag{19}$$

where $\bar{l}$ is the sample mean of the sequence of log-likelihoods $l^{(t)}$, for each iteration $t = T_0, \ldots, T$, after the burn-in. As pointed out by Raftery et al. (2007), the AICM is connected to the DIC criterion (Spiegelhalter et al. 2002). More specifically, it coincides with the DIC definition provided by Gelman et al. (2004, Sect. 6.7). Concerning the BICM, it can be related to an approximation of the log-marginal likelihood. Successful applications of the AICM in the mixture modelling context are described, for example, by Erosheva et al. (2007), Gormley and Murphy (2010), Gormley and Murphy (2011), and Mollica and Tardella (2017). The BICM is exploited, for example, by Ranciati et al. (2017), Murphy et al. (2020) and Redivo et al. (2020).

4 Simulation study

The performance of the proposed approach is investigated in a simulated environment, considering two scenarios that differ in terms of the true number of components and the distribution of the manifest variable. In both scenarios, the manifest variable y and the concomitant covariate x are assumed to be univariate, for simplicity.

The quality of the estimates for the covariate effects on the conditional means is evaluated through a comparison between the true effects and their posterior estimates, after fitting each of the following mixture of regression models:

- Semiparametric Finite Mixture of Regressions models with Concomitants (SFMRC), with flexible specification of both the mixture weights $\pi_g(x)$ and the conditional means $\mu_g(x)$, $g = 1, \ldots, G$;
- Semiparametric Finite Mixture of Regression (SFMR) models, with constant mixture weights π_g and flexible specification of the conditional means $\mu_g(x)$, $g = 1, \ldots, G$;
- (parametric) FMRC, with linearity assumption for the effect of x on both the log-odds of the mixture weights $\log(\pi_g(x)/\pi_G(x)) = \eta_g(x)$ and the conditional means $\mu_g(x)$, $g = 1, \ldots, G$;
- (parametric) FMR, with constant mixture weights π_g and linearity assumption for the effect of x on the conditional means $\mu_g(x)$, $g = 1, \ldots, G$.

Additionally, by adapting to the univariate Gaussian case the models discussed in Berrettini et al. (2021), two mixture models with concomitants are also considered:

- Semiparametric Finite Mixture models with Concomitants (SFMC), with flexible specification of the mixture weights $\pi_g(x)$ and constant conditional means μ_g, $g = 1, \ldots, G$;
- (parametric) FMC, with linearity assumption for the effect of x on the log-odds of the mixture weights $\log(\pi_g(x)/\pi_G(x)) = \eta_g(x)$ and constant conditional means μ_g, $g = 1, \ldots, G$;

For every class of models, G is initially set equal to the true number of components. In particular, the pointwise means of the estimated $\mu_g(x^*)$, denoted $\hat{\mu}_g(x^*)$, are plotted together with the pointwise 2.5 and 97.5 percentiles among all samples, where $\{x_i^*\}$, $i = 1, \ldots, n$, are grid points taken evenly in the range of covariate x. To quantitatively assess the performance of the estimators of the unknown regression functions

$\mu_g(x)$, the same measure employed in Huang and Yao (2012), Huang et al. (2013) and Xiang and Yao (2018) is adopted, that is the square Root of the Average Squared Errors (RASE), computed as

$$\text{RASE}_{\mu_g} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_g(x_i^*) - \mu_g(x_i^*)\right)^2}, \quad g = 1, \ldots, G; \tag{20}$$

in practice, the RASE measures the (Euclidean) pointwise distance between the "true" curve and the estimated one. The same graphical and quantitative evaluations are carried out for the covariate effects on the mixture weights, this time by restricting the analysis to the class of semiparametric and parametric models with concomitants.

Regarding the clustering performance, a comparison between true allocations and inferred ones is made in terms of Adjusted Rand Index (ARI) (Hubert and Arabie 1985) and soft ARI (sARI) (Flynt et al. 2019). For each method and each value of G, 4000 MCMC draws are simulated after a burn-in of as many draws. Both AICM and BICM are considered to select the optimal number of components, and the number of nonempty components $\tilde{G}$ is computed according to Equation (17). For each of the competing classes of models, a proper MCMC algorithm has been implemented in R (R Core Team 2020). The R codes for the four algorithms are available on GitHub at the following link: github.com/MarcoBerrettini/sMoE.

4.1 First simulation experiment: G=2

A batch of 100 independent datasets is generated with $n = 1000$ from a two-component mixture of regression models with weights

$$\pi_1(x) = 0.1 + 0.85\sin(\pi x), \quad \pi_2(x) = 1 - \pi_1(x),$$

where x is the only covariate, sampled from a standard uniform distribution: $x_i \sim \text{Unif}(0, 1)$, $i = 1, \ldots, 1000$. The functional form of $\eta_1(x) = \log\frac{\pi_1(x)}{1-\pi_2(x)}$, coupled with the specific range of values for x_i, leads to a nonmonotonic concave log-odds. Conditional on x and the component indicators, each component density is a Gaussian distribution, with means $\mu_1(x)$, $\mu_2(x)$, and variances σ_1^2, σ_2^2 given by:

$$\begin{aligned} \mu_1(x) &= 15(x - 0.5)^2 + 1, & \sigma_1^2 &= 0.09; \\ \mu_2(x) &= 5(x - 0.5)^2, & \sigma_2^2 &= 0.0625. \end{aligned}$$

Figure 2 shows one of the 100 independent samples. Figure 3 highlights the limits of the parametric approach when a nonmonotonic function, symmetric about $x = 0.5$, has to be approximated. In particular, for fixed number of components $G = 2$, both FMC and FMRC tend to fit a constant function with an associated RASE_η (and its standard deviation), averaged over the 100 simulations, equal to 2.402 (0.209) and 10.091 (31.020), respectively. On the other hand, SFMRC seems to catch the underlying trend even though some oversmoothing is present around the peak of the function.

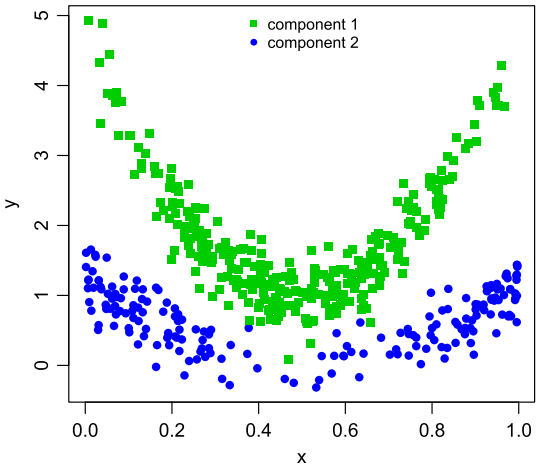


Fig. 2 First simulation experiment: example of a simulated dataset

For this model, the average RASE_η drops to 0.209, with standard deviation of 0.665. A peculiar behaviour emerges when examining the estimates for $\eta_1(x)$ obtained with SFMC, characterised by an average RASE_η equal to 4.646 (and a standard deviation equal to 0.776). This might be caused by an unsuccessful attempt to counterbalance the evident model misspecification, related to the assumption of constant conditional means, by using the flexible mixture weights' specification.

Regarding the estimated conditional means, the SFMRC shows good performance in the left panel of Fig. 4, apart from some oversmoothing in the lower component $\mu_2(x)$, for central values of x. In this area, the probability of observing units from component 2 reaches its minimum, as previously shown in Fig. 3. Around this region, most of the observations come from component 1, with only few observations from component 2. This disproportion, coupled with a certain degree of overlap of the two components, seems to have led the MCMC algorithm to assign erroneously some units from component 1 to component 2, with a consequent slight upward bias in $\hat{\mu}_2(x)$. This explains also the oversmoothing observed when estimating the effect of the covariate x on the log-odds $\eta_1(x)$ of the mixture weights. This issue becomes way more evident if constant weights are assumed without considering the effects of the concomitant covariate x, as for the SFMR; see the second plot in Fig. 4. Again, the main problem regards mostly the lower component, whose true mean is not fully included in the bands, even though they widen considerably in the overlap region.

No assumption of constant weights is made when fitting the FMRC, but, as previously shown in Fig. 3, this model estimates a constant effect of the covariate, making it practically equivalent to a FMR. This is evident in Fig. 5, where the conditional means estimated by the two models are compared. Since these two functions are generated to be quadratic and symmetric about $x = 0.5$, both parametric regression models fit horizontal lines, effectively collapsing to a simple mixture of Gaussians not involving the effect of the covariate for the conditional distribution of the dependent variable, thus leading to estimated conditional means very close to those obtained with FMC and SFMC (Fig. 6). It is worth mentioning that the fitted constant means obtained by

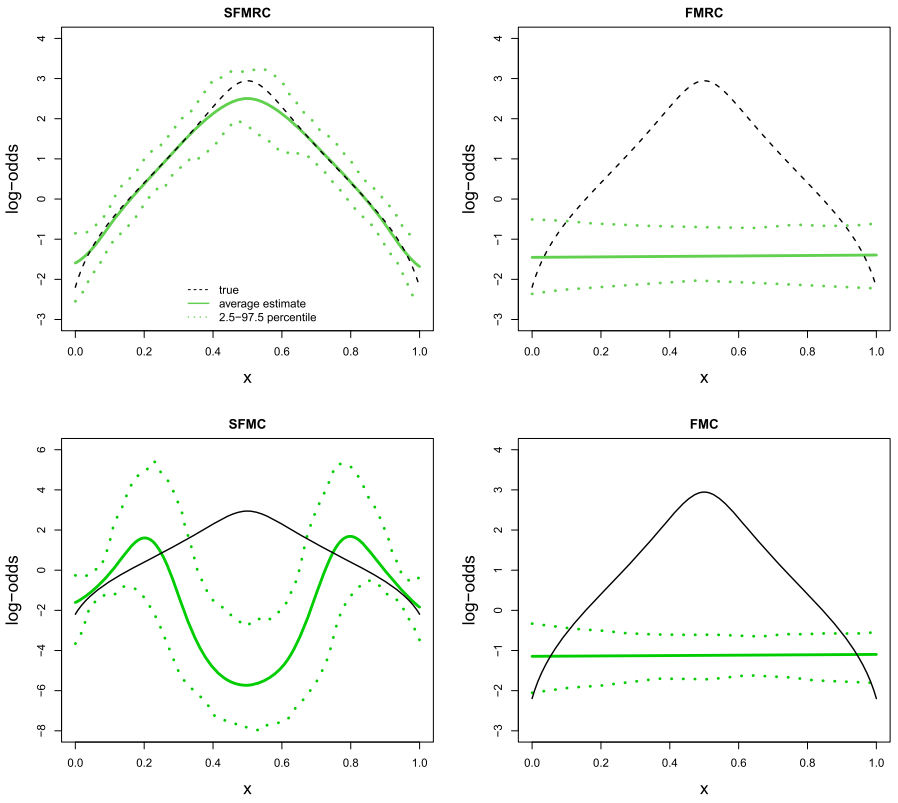


Fig. 3 First simulation experiment: pointwise average and 2.5–97.5 percentiles of the log-odds of the mixture weight η_1 estimated by the four mixture models with concomitants over 100 simulated datasets, for fixed $G = 2$

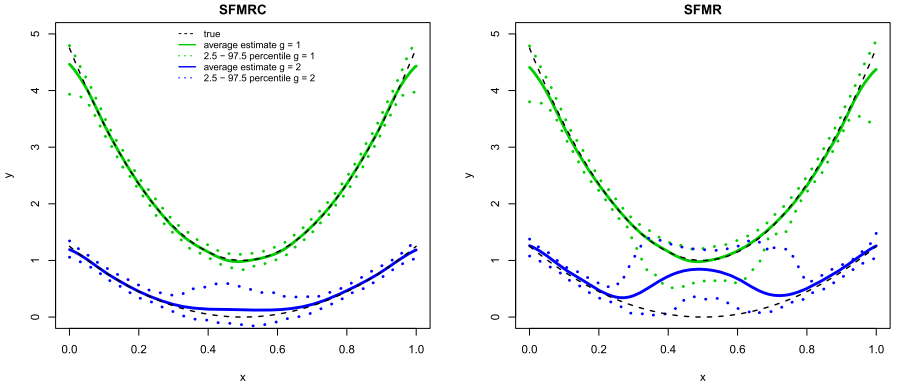


Fig. 4 First simulation experiment: pointwise average and 2.5–97.5 percentiles of the conditional means estimated with both semiparametric regression approaches over 100 simulated datasets, for fixed $G = 2$

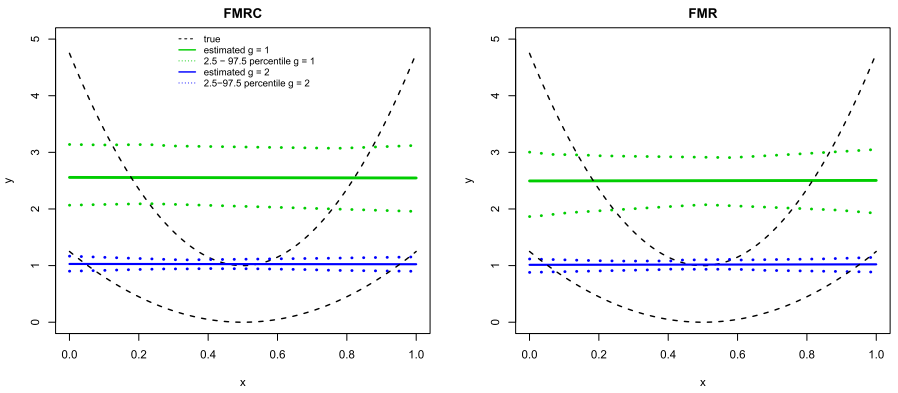


Fig. 5 First simulation experiment: pointwise average and 2.5–97.5 percentiles of the conditional means estimated with both parametric regression approaches over 100 simulated datasets, for fixed $G = 2$

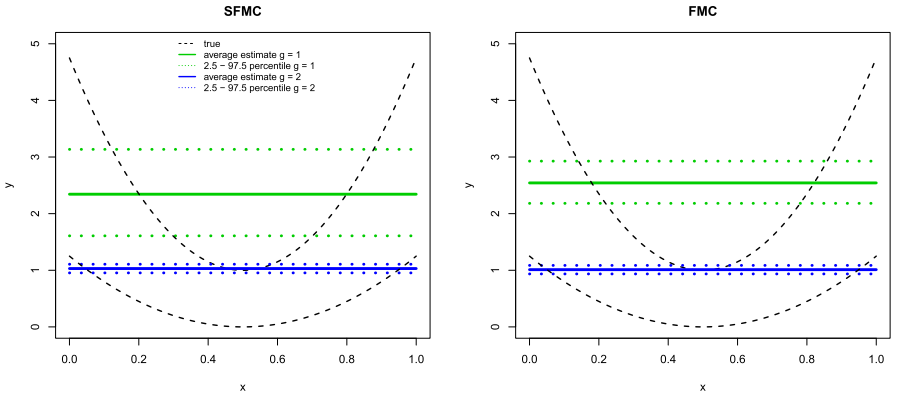


Fig. 6 First simulation experiment: pointwise average and 2.5–97.5 percentiles of the conditional means estimated with both mixture approaches with concomitants over 100 simulated datasets, for fixed $G = 2$

these four class of models are not even centered around the true average group means. Table 1 summarises this comparison among the conditional mean functions estimated by the six competing models from a quantitative point of view, by displaying, for each combination of method and component $g = 1, 2$, the average RASE_{μ_g} and the standard deviation over 100 simulated datasets. Quality of the estimates are strictly related to the quality of the allocations, as Table 2 confirms. The SFMRC, in fact, outperforms its competitors in terms of AICM, BICM, ARI and sARI for fixed number of components $G = 2$, followed by the SFMR. Given the previous considerations about the results, both the parametric approaches and those assuming constant conditional means prove to be not satisfactory in this simulation setting.

A comparison among the six competing models is performed also by examining the best models selected according to AICM and BICM when considering a number of components ranging from 1 to 4. Table 3 reports the distribution of the number of nonempty components $\tilde{G}$ resulting from the selection based on AICM. $\tilde{G} = 2$ is always the best choice according to the SFMRC, while 26 times out of 100 the SFMR provides

Table 1 First simulation experiment: mean (and standard deviation) of the RASE scores computed on the estimated conditional means over 100 simulated datasets

RASE_{μ_g}	$\mu_1(x)$	$\mu_2(x)$
SFMRC	0.102 (0.066)	0.099 (0.086)
SFMR	0.162 (0.232)	0.450 (0.237)
FMRC	1.222 (0.199)	0.716 (0.062)
FMR	1.179 (0.078)	0.719 (0.133)
SFMC	1.246 (0.092)	0.718 (0.035)
FMC	1.172 (0.055)	0.701 (0.033)

Table 2 First simulation experiment: average AICM, BICM, ARI and sARI (number of times each model ranks first) over 100 simulated datasets, for fixed $G = 2$

	AICM	(best)	BICM	(best)	ARI	(best)	sARI	(best)
SFMRC	232.0	(99)	309.2	(99)	0.960	(98)	0.940	(98)
SFMR	453.9	(1)	792.8	(1)	0.627	(2)	0.580	(2)
FMRC	1521.9	(0)	2020.8	(0)	−0.012	(0)	−0.000	(0)
FMR	1607.8	(0)	2247.0	(0)	−0.007	(0)	0.006	(0)
SFMC	1235.9	(0)	2956.7	(0)	−0.006	(0)	−0.009	(0)
FMC	1532.1	(0)	4743.9	(0)	−0.008	(0)	0.006	(0)

Table 3 First simulation experiment: number of nonempty component selected for each method, according to AICM, over 100 simulated datasets

	$\tilde{G} = 1$	$\tilde{G} = 2$	$\tilde{G} = 3$	$\tilde{G} = 4$
SFMRC	–	100	–	–
SFMR	–	74	26	–
FMRC	90	10	–	–
FMR	93	2	5	–
SFMC	–	1	20	79
FMC	4	–	7	89

a better AICM with an additional component. It is interesting to note that parametric regression approaches tend to show better values for AICM when considering a single component. Conversely, models characterised only by the presence of concomitants appear to improve if a larger number of components is considered. This again might be related to the fact that increasing the number of components might compensate for their intrinsic mispecification.

By comparing the results reported in Table 3 to those in Table 4, the tendency of the BICM to select models with fewer nonempty components is apparent. This seems to be consistent with what has been already reported in the literature (see, for example, Redivo et al. 2020). Only SFMRC models do not seem to suffer from this systematic underestimation.

Examining the best models (according to AICM) fitted with each method for each simulated dataset, rather than fixing the number of components, the conclusions does not seem to change. All AICM averaged values reported in Table 5 improve with

Table 4 First simulation experiment: number of nonempty component selected for each method, according to BICM, over 100 simulated datasets

	$\tilde{G}=1$	$\tilde{G}=2$	$\tilde{G}=3$	$\tilde{G}=4$
SFMRC	–	100	–	–
SFMR	66	34	–	–
FMRC	100	–	–	–
FMR	100	–	–	–
SFMC	96	–	4	–
FMC	100	–	–	–

Table 5 First simulation experiment: average AICM, ARI and sARI (number of times each model ranks) over 100 simulated datasets, for optimal G according to AICM

	AICM	(best)	ARI	(best)	sARI	(best)
SFMRC	228.4	(100)	0.975	(97)	0.955	(95)
SFMR	430.7	(0)	0.587	(3)	0.533	(5)
FMRC	1342.0	(0)	−0.003	(0)	−0.002	(0)
FMR	1344.2	(0)	−0.002	(0)	−0.001	(0)
SFMC	545.2	(0)	0.135	(0)	0.117	(0)
FMC	1015.8	(0)	0.072	(0)	0.062	(0)

respect to those in Table 2, also for the SFMRC, because sometimes having an extra component, even if it is emptied during the posterior allocation, slightly decreases AICM. For the same reason, both the average ARI and sARI appear to slightly improve for the SFMRC, while they worsen for models that tend to pick the wrong number of components; see Table 5. Similar conclusions can be drawn when considering the best models selected using BICM (data not shown).

4.2 Second simulation expertiment: G=3

A batch of 100 independent datasets is generated with $n = 1000$ from a three-component mixture of regression models, with log-odds of mixture weights $\eta_g(x) = \log \pi_g(x)/\pi_3(x)$, $g = 1, 2$, defined as:

$$\eta_1(x) = 3\,\frac{\exp(7.5 - 15x)}{1 + \exp(7.5 - 15x)} - 1.5,$$
$$\eta_2(x) = 3\,\frac{\exp(15x - 7.5)}{1 + \exp(15x - 7.5)} - 1.5;$$

where x is the only covariate, sampled from a uniform distribution: $x_i \sim \text{Unif}(0, 1)$, $i = 1, \ldots, 1000$. Conditional on x and the component indicators, y follows a univariate Gaussian distribution, with means $\mu_1(x), \mu_2(x), \mu_3(x)$, and variances $\sigma_1^2, \sigma_2^2, \sigma_3^2$,

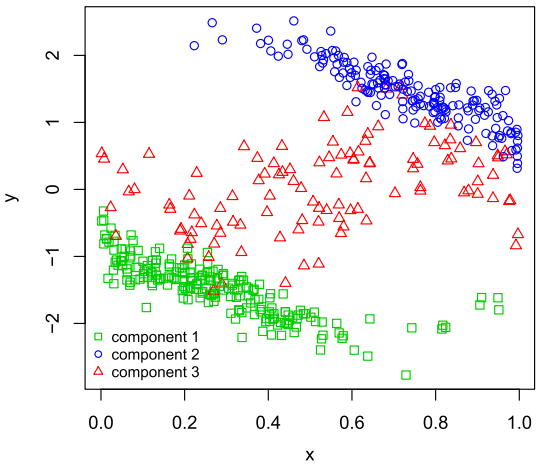


Fig. 7 Second simulation experiment: example of a simulated dataset

respectively, defined as follows:

$$\mu_1(x) = 0.5\sin(6x + 0.8) + \exp(-16(3x + 0.15)^2) - 1.75, \quad \sigma_1^2 = 0.04;$$
$$\mu_2(x) = 1.75 - 0.5\sin(6x + 0.8) + \exp(-16(3x + 0.15)^2) - 1.75), \quad \sigma_2^2 = 0.04;$$
$$\mu_3(x) = -0.5\sin(2\pi x), \quad \sigma_3^2 = 0.25.$$

Figure 7 shows one of the 100 independently generated samples for this second scenario.

Figure 8 shows that the SFMRC is able to catch almost perfectly the effects of the covariate x on both predictors η_1 and η_2. On the contrary, due to nonlinearity, the linear approximation by the FMRC is worse, so that the true effects exceed the bands at the boundaries of the range of x. As expected, the average RASE scores for the SFMRC (together with the associated standard deviations) reported in Table 6 are lower than those of the parametric competitor. Table 6 contains also information about the performance of SFMC and FMC. In this second simulation experiment, it is evident that imposing constant conditional means has a dramatic impact on the ability of these two class of models in recollecting the effect of the covariates on the mixture weights: the average RASE associated with these two classes of mixture models with concomitants are almost ten times larger than those obtained with SFMRC.

Regarding the estimates of the conditional means, the SFMRC seems to outperform the competitors, despite some overlap present between Cluster 1 and Cluster 3 for low values of x, and between Cluster 2 and Cluster 3 for high values of x (see Fig. 9). The FMRC is unable to properly approximate the nonlinear trends, especially where there are fewer observations (i.e. in Cluster 1 for high values of x, and in Cluster 2 for low values of x). Nevertheless, Fig. 10 shows that, thanks to the good estimates of the mixture weights, the FMRC discriminates almost perfectly among groups in the aforementioned overlapping areas.

Results for SFMR are reported, in detail, in Fig. 11. The flexibility allowed for the estimates of the conditional means, combined with the impossibility to include the

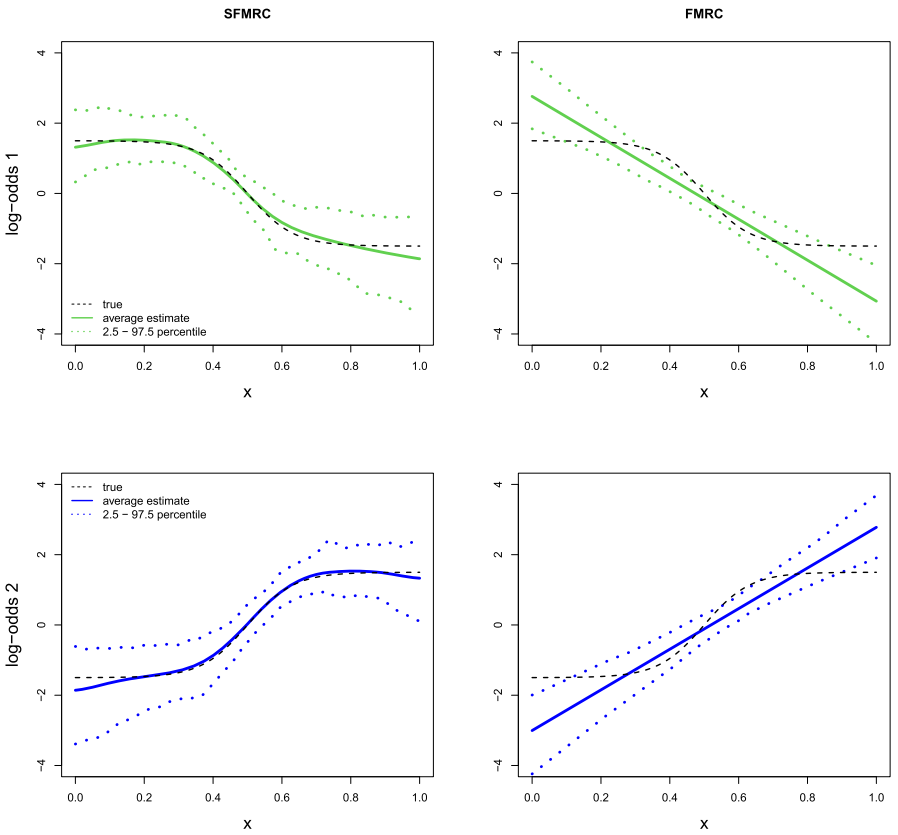


Fig. 8 Second simulation experiment: comparison between the log-odds of the mixture weights estimated by the semiparametric (left) and parametric (right) mixture of regressions with concomitants over 100 simulated datasets, for fixed $G = 3$

Table 6 Second simulation experiment: mean (and standard deviation) of the RASE scores computed on the estimated log-odds of the mixture weights over 100 simulated datasets, for fixed $G = 3$

RASE_{η_g}	$\eta_1(x)$	$\eta_2(x)$
SFMRC	0.421 (0.153)	0.431 (0.169)
FMRC	0.679 (0.210)	0.663 (0.213)
SFMC	3.372 (3.662)	2.789 (1.216)
FMC	4.079 (2.206)	3.963 (2.146)

effect of covariate x into the estimates of the mixture weights, results into overlapping estimated functions and wide bands. The performance of the FMR is just slightly worse with respect to the ones obtained by the FMRC. The main differences can be observed in the overlapping regions of Fig. 12, where the estimated conditional means intersect each other. As far as SFMC and FMC are concerned, these two class of models show a similar behaviour in the estimated (constant) conditional means (Figs. 13 and 14). Due to the specific settings considered in this second experiment, the impact of the mispecification error seems less severe. However, it is apparent that both models

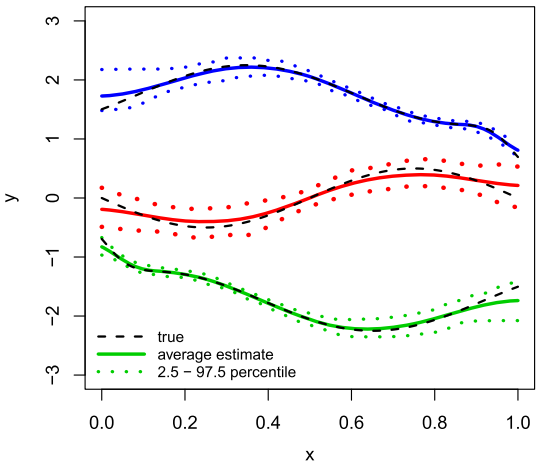


Fig. 9 Second simulation experiment: conditional means estimated by the SFMRC over 100 simulated datasets, for fixed $G = 3$

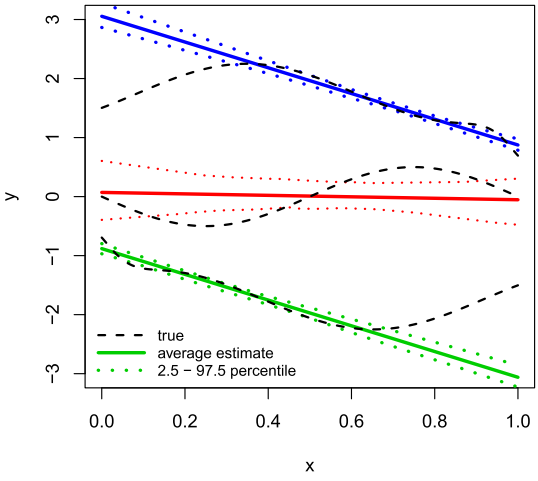


Fig. 10 Second simulation experiment: conditional means estimated by the FMRC over 100 simulated datasets, for fixed $G = 3$

suffers from a large sampling variability in the estimated conditional means. This is particularly evident in the estimates for component 3. It is worth noting that component 3 is used as baseline to define the log-odds. Thus, the extremely large variability in the estimates for the conditional mean of this component can be connected with the previously mentioned poor performance of SFMC and FMC in estimating the log-odds. All the conclusions drawn from a graphical point of view are confirmed by the quantitative results in terms of RASE_μ reported in Table 7.

Table 8 shows that, in terms of both AICM and BICM, the SFMRC is evidently better than its competitors with $G = 3$. However, this result does not correspond to an equal gap in the quality of the allocations, expressed in terms of both ARI and sARI. Indeed, either mixture of regression models with concomitants perform well, even though the semiparametric one slightly prevails.

Regarding model selection, for each competing method and each simulated dataset, the best model is considered among different mixture models with $G = 1, \ldots, 5$.

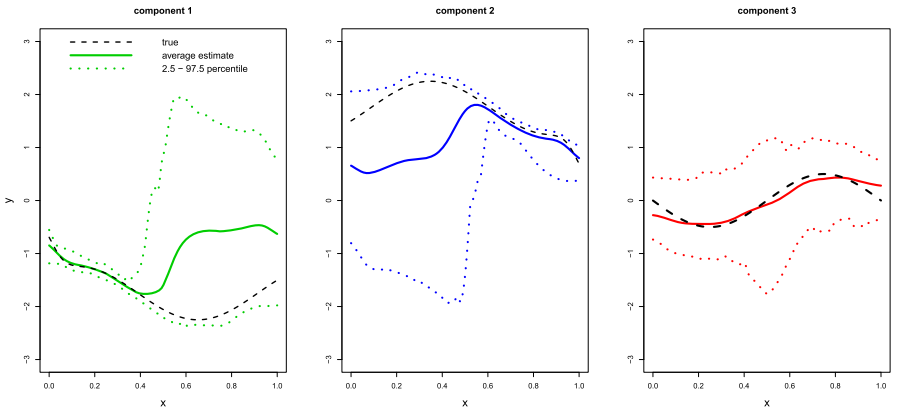
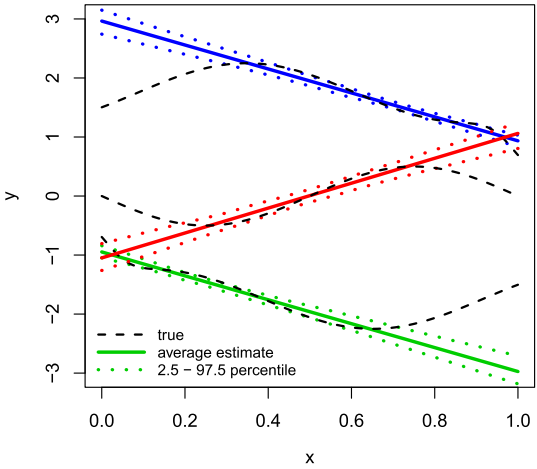


Fig. 11 Second simulation experiment: conditional means estimated by the SFMR model over 100 simulated datasets, for fixed $G = 3$

Fig. 12 Second simulation experiment: conditional means estimated by the FMR over 100 simulated datasets, for fixed $G = 3$



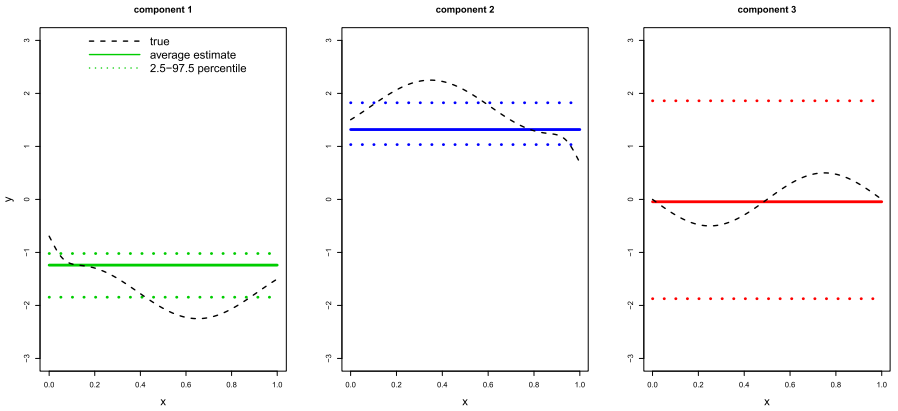


Fig. 13 Second simulation experiment: conditional means estimated by the SFMC model over 100 simulated datasets, for fixed $G = 3$

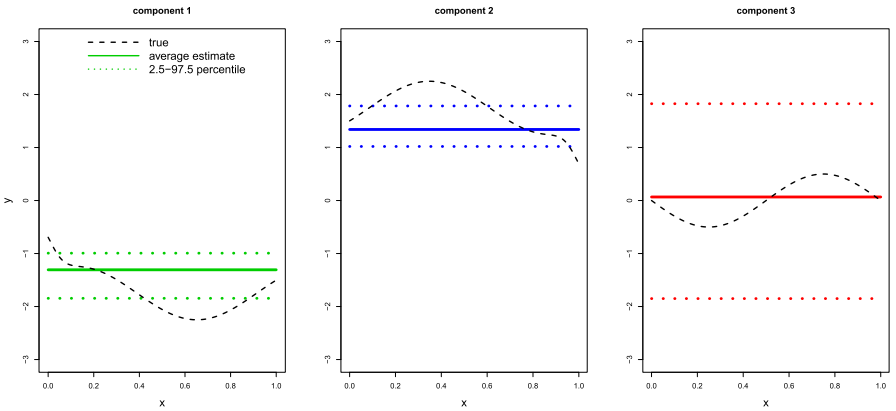


Fig. 14 Second simulation experiment: conditional means estimated by the FMC model over 100 simulated datasets, for fixed $G = 3$

Table 7 Second simulation experiment: mean (and standard deviation) of the RASE scores computed on the estimated conditional means over 100 simulated datasets

RASE_{μ_g}	$\mu_1(x)$	$\mu_2(x)$	$\mu_3(x)$
SFMRC	0.086 (0.035)	0.086 (0.038)	0.145 (0.034)
SFMR	1.073 (1.078)	0.995 (1.041)	0.472 (0.188)
FMRC	0.509 (0.042)	0.507 (0.048)	0.101 (0.472)
FMR	0.474 (0.123)	0.459 (0.069)	0.343 (0.164)
SFMC	0.662 (0.108)	0.617 (0.132)	1.248 (0.532)
FMC	0.627 (0.140)	0.604 (0.145)	1.181 (0.459)

Table 8 Second simulation experiment: average AICM, BICM, ARI and sARI (number of times each method ranks first) over 100 simulated datasets, for fixed $G = 3$

	AICM	(best)	BICM	(best)	ARI	(best)	sARI	(best)
SFMRC	252.8	(100)	602.4	(99)	0.906	(99)	0.845	(96)
SFMR	755.0	(0)	1261.8	(1)	0.326	(0)	0.260	(0)
FMRC	1641.7	(0)	1965.6	(0)	0.854	(1)	0.797	(4)
FMR	1463.8	(0)	2734.3	(0)	0.804	(0)	0.568	(0)
SFMC	1041.3	(0)	2523.5	(0)	0.525	(0)	0.513	(0)
FMC	1066.6	(0)	2516.7	(0)	0.520	(0)	0.512	(0)

Table 9 shows that, when the selection is based on AICM, the SFMRC is the only one which is able to consistently pick the correct number of nonempty components.

This leads to more favorable results for the SFMRC, if ARI and sARI computed with reference to the best models selected by each method are compared; see Table 10. On the contrary, results worsen when BICM is considered as model selection criterion. As shown in Table 11, the tendency of BICM to underestimate the actual number of nonempty components is confirmed, and in this second simulation experiment also SFMRC models suffer from this. As a consequence, the values of ARI and sARI

Table 9 Second simulation experiment: number of nonempty components selected for each method, according to AICM, over 100 simulated datasets

	$\tilde{G}=1$	$\tilde{G}=2$	$\tilde{G}=3$	$\tilde{G}=4$	$\tilde{G}=5$
SFMRC	–	–	100	–	–
SFMR	–	–	12	88	–
FMRC	14	62	17	7	–
FMR	22	–	47	27	4
SFMC	–	13	10	5	72
FMC	–	62	14	8	16

Table 10 Second simulation experiment: average AICM, ARI and sARI (number of times each method ranks first) over 100 simulated datasets, for optimal G according to AICM

	AICM	(best)	ARI	(best)	sARI	(best)
SFMRC	245.4	(100)	0.905	(99)	0.846	(97)
SFMR	588.7	(0)	0.292	(0)	0.257	(0)
FMRC	1278.6	(0)	0.599	(1)	0.586	(3)
FMR	1363.0	(0)	0.623	(0)	0.439	(0)
SFMC	844.6	(0)	0.516	(0)	0.505	(0)
FMC	968.3	(0)	0.574	(0)	0.576	(0)

Table 11 Second simulation experiment: number of nonempty components selected for each method, according to BICM, over 100 simulated datasets

	$\tilde{G}=1$	$\tilde{G}=2$	$\tilde{G}=3$	$\tilde{G}=4$	$\tilde{G}=5$
SFMRC	4	63	33	–	–
SFMR	100	–	–	–	–
FMRC	100	–	–	–	–
FMR	100	–	–	–	-
SFMC	1	99	–	–	–
FMC	3	96	1	–	–

computed for the best models selected using BICM are negatively impacted (data not shown).

5 Real data applications

5.1 Baseball salaries data

Watnik (1998) provides a dataset consisting of information about players for the 1992 Major League Baseball season. In particular, their salaries are considered as the response, along with measures of the 337 players' previous year's performances. Notice that this dataset is already well known in the literature on FRMC (see, for example, Khalili and Chen 2007; Chamroukhi and Huynh 2018). For simplicity, the analysis described in this section focuses on one of the quantitative covariates, the number of runs, taken as a measure of a player's contribution to the team. More specifically, the effect of this variable on player salaries is studied, by fitting the six different mixture

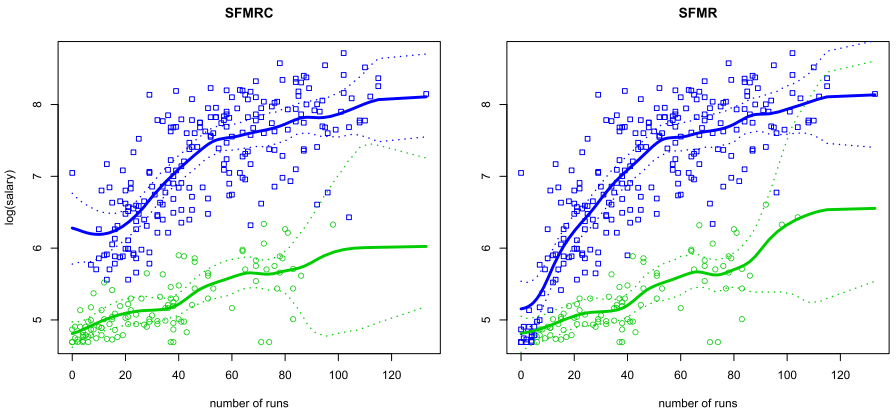


Fig. 15 Baseball salaries data: estimated posterior conditional means (and pointwise 95% posterior credible bands) obtained from the SFMRC (left panel) and the SFMR (right panel)

of regression models considered in Sect. 4 for a fixed number of components ranging from $G = 1$ to $G = 4$. As suggested by Watnik (1998), due to asymmetry, the response is preemptively transformed by taking the natural logarithm.

In light of the tendency of BICM to underestimate the number of nonempty components highlighted in the simulation experiments, in this application the optimal value for G is selected according to AICM. The number of nonempty components $\tilde{G}$ resulted to be equal to 2 for the two semiparametric mixture of regression models (SFMRC and SFMR) as well as the two mixture of Gaussians (SFMC and FMC), and equal to 1 for the two parametric mixture of regression models (FMRC and FMR). Among these six models, the SFMRC presents the smallest AICM (663.4), followed by the SFMR (733.7), the SFMC (781.1) and the FMC (817.3), while the remaining best parametric mixture of regression models, having $G = 1$, collapse to the same model, with the highest AICM (888.1).

As Fig. 15 shows, the main difference between the semiparametric mixture of regression models seems to be related to the allocation of players with a low number of runs. The SFMRC keeps the two clusters well separated, by assigning all of these units to the lower one, whereas the SFMR creates some overlap, such that the functions describing the conditional means, $\hat{\mu}_1(x)$ and $\hat{\mu}_2(x)$, almost intersect one another. Figure 15 shows, in both cases, the presence of a nonlinear effect of the number of runs on the log-salary for the upper cluster, while the bands does not exclude a linear effect for the lower cluster.

Fixing the number of components $G = 2$, the FMR allocates the players similarly to the SFMRC. In particular, only 9 units out of 337 (ARI = 0.894) differ in cluster allocation between the two models. Focusing on the parametric regression approaches, the main difference between the allocations seems to be related to few among the lowest paid players having a number of runs ranging between 30 and 90, which are assigned to the upper component by the FMRC. This probably induces variability in the estimated mean functions of the latter, which present wider bands, if compared to the ones estimated by FMRC; see Fig. 16.

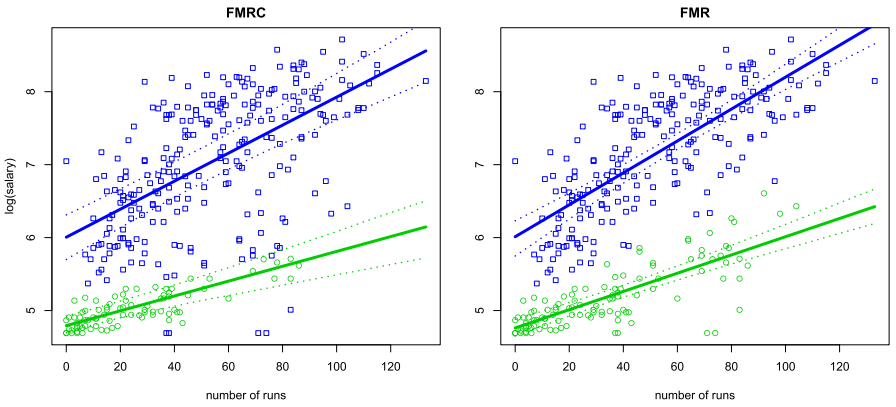


Fig. 16 Baseball salaries data: estimated posterior conditional means (and pointwise 95% posterior credible bands) obtained from the FMRC (left panel) and the FMR (right panel), for fixed $G = 2$

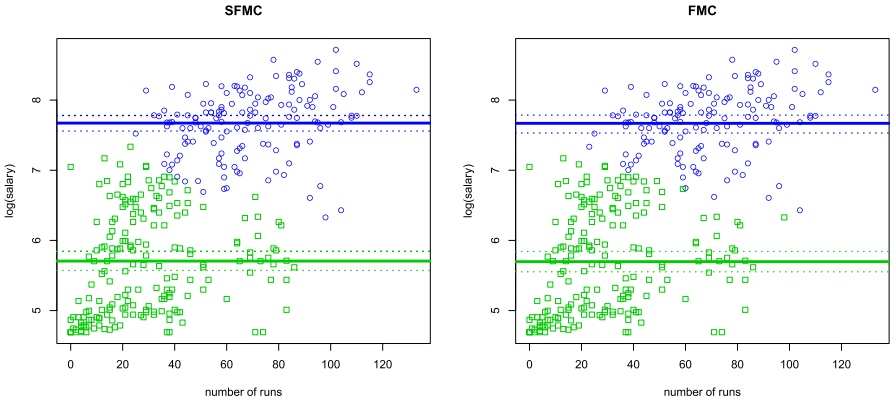


Fig. 17 Baseball salaries data: estimated posterior conditional means (and pointwise 95% posterior credible bands) obtained from the SFMC (left panel) and the FMC (right panel), for fixed $G = 2$

The two approaches with constant conditional means produce a sensibly different partition with respect to the other methods, and similar to each other, detecting a cluster of highly paid players (around 500.000 dollars or more); see Fig. 17. Here, the boundary (not drawn) separating the two groups is not perfectly horizontal due to the covariate effect on the log-odds of the weights; see Fig. 18. In particular, all four mixture models with concomitants agree about the presence of a decreasing trend in the effect of the number of runs on the log-odds of the mixture weight $\eta_1(x)$, but the semiparametric methods estimate nonlinear functions that cannot be approximated properly by a straight line. The ability to pick this underlying effect is also the main reason for the differences observed between the performances of the two semiparametric regression approaches.

The partition induced by the SFMRC identifies a cluster, the lower one (in green), which might be broadly interpreted as the cluster of "underrated" (or "underpaid", with respect to the others) baseball players. In fact, while it is obvious players with better

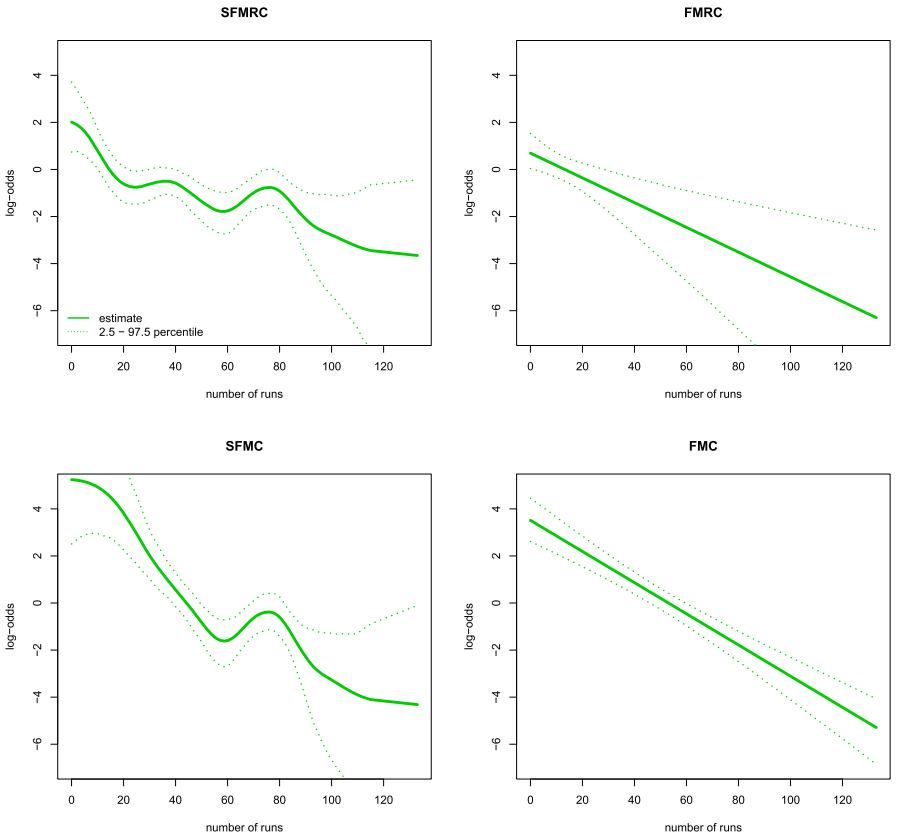


Fig. 18 Baseball salaries data: estimated posterior effects on the log-odds (and pointwise 95% posterior credible bands) obtained from the SFMRC (top left panel), the FMRC (top right panel), the SFMC (bottom left panel), the FMC (bottom right panel) for fixed $G = 2$

performances get paid more, as corroborated by the increasing trends of both means, there seems to be a group of players whose salary is substantially lower than that of players achieving similar performances (in terms of number of runs), belonging to the upper group (in blue). Indeed, the two estimated mean functions $\hat{\mu}_1(x)$ and $\hat{\mu}_2(x)$ in Fig. 15 appear almost parallel. A partial explanation of this result can be found in some additional pieces of information available in the dataset. In particular, there is a variable indicating the "free agency eligibility" of each player, i.e. if that player could have gone to a team of his choice in 1992. At the time -Watnik (1998) explains- only players with a certain amount of experience were eligible for free agency (134 out of 337) and, thus, able to market themselves to the highest bidder. On the contrary, if a player not "free agency eligible" wanted to play, he had to accept what his team was willing to pay him, or go with his team to an appointed "arbitrator", who would choose between the player's suggested salary and the team's one. However, "arbitration eligibility", which is included in the dataset as a variable as well, was for players (65 out of 337, in the dataset) who had some experience in the League, although not enough to be eligible

Table 12 Baseball salaries data: comparison between the resulting allocations of the SFMRC and (free agency or arbitration) eligibility

Cluster	Free agency or arbitration		
	Not eligible	Eligible	
Lower (green)	109	6	115
Upper (blue)	29	193	222
	138	199	337

Table 13 Nitrogen oxide data: AICM produced by the six models for optimal G

	AICM	$\tilde{G}$
SFMRC	−53.6	2
SFMR	−65.3	2
FMRC	−52.1	2
FMR	−51.2	2
SFMC	112.0	3
FMC	110.9	3

for free agency. For interpretation purpose, the two above described categories, "free agency eligible" and "arbitration eligible" players are merged, and Table 12 compares the partition induced by SFMRC with the one obtained by distinguishing between (free agency or arbitration) eligible and noneligible players. The resulting ARI (0.626) is the highest observed among the six models. Indeed, it can be noticed that almost all the eligible players (193 out of 199) belong to the upper (blue) cluster, together with 29 players who apparently had been able to obtain an "adequate" salary without probing the market.

Rather than using the additional information on eligibility to validate the clustering results, without including it into the models (and thus, treating it as a potential source of unobserved heterogeneity), this binary variable could be explicitly included into the models as an additional (binary) covariate. This analysis is focused only on the two semiparametric mixture of regression models (SFMRC and SFMR). Not surprisingly, the inclusion of this additional covariate leads the AICM to reduce the optimal number of components to $G = 1$ for both of them. This is consistent with the fact that the two conditional means as estimated by the two-component SFMRC (without the binary covariate) are almost parallel, and can be reasonably approximated by a single curve plus a vertical shift depending on the value taken by the binary covariate. In addition, this simplification in the model resulted in an AICM value lower than the one associated with the two-component SFMRC. This second part of the application should allow the Reader to appreciate the efficacy of SFMRC models in detecting possible sources of unobserved heterogeneity. Appendix A contains details about how to extend both the specification of the model given in Sect. 2 and the MCMC algorithm provided in Sect. 3.1 in order to include a binary covariate.

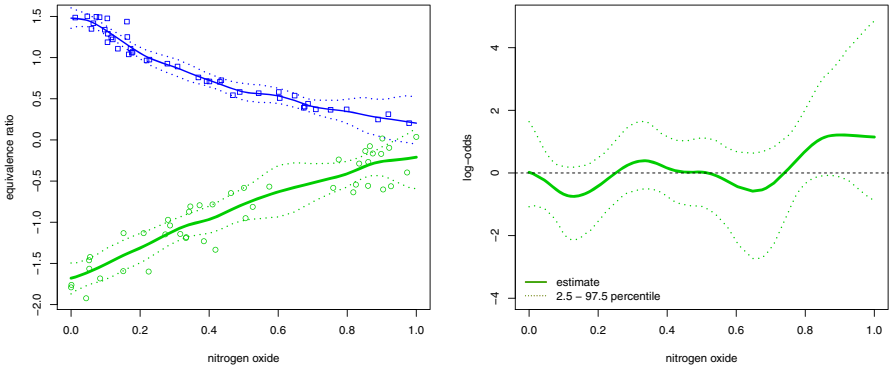


Fig. 19 Nitrogen oxide data: estimated posterior conditional means (left panel) and log-odds of mixture weights (right panel), with pointwise 95% posterior credible bands for the two components

5.2 Nitrogen oxide data

First introduced by Brinkman (1981), this dataset includes 88 observations about the concentration of nitric oxide in engine exhaust and the equivalence ratio, which represents a measure of the richness of the air-ethanol mix, for burning ethanol in a single-cylinder automobile test engine. Mixtures of regression models have been already fit to these data in Xiang and Yao (2018), where the equivalence ratio has been considered as the dependent variable y, and the concentration of nitric oxide is taken as concomitant covariate x. In this Paper, a similar analysis is performed using the proposed flexible Bayesian approach. Although the two-component structure seems quite clear in the scatterplot, six algorithms with different levels of flexibility have been run for 10000 iterations each (burn-in: 5000 draws) with G ranging from $G = 1$ to $G = 4$. According to AICM, all the regression models considered find two nonempty components, while the remaining methods find three. Figure 19 shows the covariate effects as estimated by the proposed semiparametric mixture of regressions with concomitant covariates. More specifically, in the left panel the estimated conditional means are reported, while in the right one the estimated log-odds $\eta_1(x)$ can be observed. Both plots are in line with the results by Xiang and Yao (2018) and indicate the lack of need for a flexible specification of any covariate effect in this case. In fact, the bands do not exclude linearity, although some mild nonlinearity seems to be present in the conditional means. This shows the efficacy of the penalisation induced by the selected priors in adjusting the proposed model to the required level of complexity. Coherently, AICM points at SFMR as the model to be preferred; see Table 13.

6 Conclusions

In this Paper, a general specification of mixture of regression models is proposed, allowing both component weights and conditional means to be nonlinear functions of a covariate. This general approach resort to spline functions for approximating the smooth effect of the concomitant variables. Parameter estimation is based on a

Bayesian approach through MCMC machinery. In principle, the Reader might question whether a full parametric approach, e.g. by considering a monomial set of bases to represent the map between component probabilities and covariates, could prove to be flexible enough to catch nonlinearity. Unfortunately, this parametric representation would require some arbitrary choices, such as the maximum degree for monomial bases, or the definition of an automatic selection criterion. The approach advocated for in this Paper bypasses this issue by controlling flexibility through the variance parameters of the spline coefficients, following Lang and Brezger (2004).

Using simulation experiments, the proposed method has proved to be a useful tool for recovering the underlying covariate effects -especially if indeed not linear- and, consequently, for estimating models with a better goodness of fit and leading to a more accurate allocation. The potential of the proposal has been illustrated also through applications to real data.

Although the results shown seem encouraging, the proposed model is characterised by some limitations and there are some issues that might deserve further investigation.

Firstly, there are limitations related to the assumption for both the manifest variable and the concomitant covariate to be univariate. The adaptation to the multivariate case would require particular attention to deal with the presence of component-specific covariance matrices and multiple regressors. Regarding the latter, each predictor could be expressed as a sum of smooth functions, as in the additive paradigm introduced by Hastie and Tibshirani (1990); hence, Bayesian P-splines could be used to approximate such nonlinear functions. However, conditions for identifiability should be revised, and, in particular, further constraints should be introduced to guarantee identifiability of the predictors.

In addition, the simulation experiments have highlighted the tendency of BICM to underestimate the actual number of components. As previously mentioned, this seems coherent with other results already reported in the literature (see, for example, Redivo et al. 2020). Since BICM has been introduced as an approximation of the (logarithm of) the marginal likelihood, it would be interesting to consider other estimators of this quantity. For example, Frühwirth-Schnatter (2019) has recently proposed two bridge sampling estimators of the marginal likelihood for FMRC models, and investigating their performance in the context of the semiparametric models proposed in this Paper could be the subject of future investigation.

Furthermore, adopting a probit representation (Geweke and Keane 2007) instead of the dRUM approach could provide another potential avenue to explore, in order to understand the benefit of the proposed modelling framework, especially when the ease of interpretation given by the logit formulation is not of relevance.

Alternative approaches to penalise the spline coefficients associated with the bases could be considered. For example, one could consider applying shrinkage to the variances τ_g^2 and δ_g^2 in a similar way to what Bitto and Frühwirth-Schnatter (2019) and Cadonna et al. (2020) suggest in the context of time varying parameter models.

As far as the computational implementation is concerned, as a consequence of the use of mixture of Gaussians to approximate the Logistic distribution, no Metropolis-Hastings steps are required in the proposed MCMC algorithm. Although this can be considered an advantage, the increase in the computational burden due to the introduction of an additional latent variable should not be ignored. Furthermore, the

implemented MCMC algorithm relies on the specification of a fixed value for the number of components. If this quantity is unknown, it is necessary to estimate it by running the algorithm many times with different inputs, which might be time consuming, especially when the "true" value is large. One solution could be incorporating the choice of G within the algorithm itself. For instance, a reversible jump MCMC algorithm could be exploited (Richardson and Green 1997), by designing appropriate dimension-changing moves. Alternatively, the issue of choosing the optimal value for G could be circumvented by focusing the attention on the posterior distribution of the number of nonempty components, through the combination of a large value for G with appropriate prior distributions, as suggested in Malsiner-Walli et al. (2016). This latter strategy seems more coherent with the peculiar behaviour observed in the simulation studies, where the proposed MCMC algorithm occasionally converges to a solution that is characterised by empty components.

Funding Open access funding provided by Alma Mater Studiorum - Universitá di Bologna within the CRUI-CARE Agreement.

A Including an additional binary covariate

Let $d_i \in \{0, 1\}$ be a binary covariate associated to unit $i = 1, \ldots, n$. Then, Equations (2) and (4) can be respectively rewritten as follows:

$$\eta_g(x_i, d_i) = \sum_{\rho=1}^{m} B_\rho(x_i)\gamma_{g\rho} + d_i\xi_g \,; \tag{21}$$

$$\mu_g(x_i, d_i) = \sum_{\rho=1}^{m} B_\rho(x_i)\beta_{g\rho} + d_i\zeta_g \,; \tag{22}$$

with ξ_g and ζ_g unknown regression coefficients. Assuming a diffuse Gaussian prior $N(0, \psi_g^2)$ on such parameters, with variance ψ_g^2 set sufficiently high (e.g., 100), leads to the following modifications to be applied to points 1 and 5 of the Gibbs sampler introduced in Sect. 3.1:

1a. For $g = 1, \ldots, G-1$, sample the regression coefficients' vector $\boldsymbol{\gamma}_g$ conditional on $\mathbf{z}_g$ and $\mathbf{r}_g$ from a multivariate Gaussian density with precision matrix $\mathbf{P}_g$ and

mean $\mathbf{m}_g$ given by

$$\begin{aligned}\mathbf{P}_g &= \mathbf{B}'\mathbf{W}_g^{-1}\mathbf{B} + \frac{1}{\delta_g^2}\mathbf{K},\\ \mathbf{m}_g &= \mathbf{P}_g^{-1}\mathbf{B}'\mathbf{W}_g^{-1}\left(\mathbf{z}_g - \mathbf{d}\xi_g + \log\boldsymbol{\lambda}_{-g}(\mathbf{x},\mathbf{d})\right),\end{aligned} \tag{23}$$

with $\boldsymbol{\lambda}_{-g}(\cdot)$ defined as in Equation (8), and $\mathbf{d} = (d_1, \ldots, d_n)$;

1b. Sample the $G-1$ regression coefficients ξ_g associated to the binary covariate, conditional on $\mathbf{z}_g$ and $\mathbf{r}_g$, from a Gaussian density with precision p_g and mean m_g given by

$$\begin{aligned}p_g &= \mathbf{d}'\mathbf{W}_g^{-1}\mathbf{d} + \frac{1}{v^2},\\ m_g &= \frac{1}{p_g}\mathbf{d}'\mathbf{W}_g^{-1}\left(\mathbf{z}_g - \mathbf{B}\boldsymbol{\gamma}_g + \log\boldsymbol{\lambda}_{-g}(\mathbf{x},\mathbf{d})\right);\end{aligned} \tag{24}$$

5a. For $g = 1, \ldots, G$, sample the regression coefficient $\boldsymbol{\beta}_g$, $g = 1, \ldots, G$ from a multivariate Gaussian density with covariance matrix $\mathbf{V}_g$ and mean $\boldsymbol{\nu}_g$

$$\mathbf{V}_g = \left(\frac{1}{\sigma_g^2}\mathbf{B}^{(g)'}\mathbf{B}^{(g)} + \frac{1}{\tau_{gj}^2}\mathbf{K}_j\right)^{-1}, \quad \boldsymbol{\nu}_g = \mathbf{V}_g\mathbf{B}^{(g)'}\left(\mathbf{y}^{(g)} - \mathbf{d}^{(g)}\zeta_g\right), \tag{25}$$

where, consistently with the notation provided in (14), $\mathbf{d}^{(g)}$ indicate a subvector of $\mathbf{d}$ with elements corresponding to the units allocated to the g-th group.

5b. Sample the G regression coefficients ζ_g, $g = 1, \ldots, G$ from a Gaussian density with covariance matrix v_g and mean ν_g

$$v_g = \left(\frac{1}{\sigma_g^2}\mathbf{d}^{(g)'}\mathbf{d}^{(g)} + \frac{1}{\psi_g^2}\right)^{-1}, \quad \nu_g = v_g\mathbf{d}^{(g)'}\left(\mathbf{y}^{(g)} - \mathbf{B}^{(g)}\boldsymbol{\beta}_g\right). \tag{26}$$

References

Berrettini M, Galimberti G, Ranciati S, Murphy TB (2021) Flexible Bayesian modelling of concomitant covariate effects in mixture models. arXiv preprint arXiv:2105.12852

Bitto A, Frühwirth-Schnatter S (2019) Achieving shrinkage in a time-varying parameter model framework. J Econom 210(1):75–97

Brezger A, Lang S (2006) Generalized structured additive regression based on Bayesian P-splines. Comput Stat Data Anal 50(4):967–991

Brinkman ND (1981) Ethanol fuel-single-cylinder engine study of efficiency and exhaust emissions. SAE Trans 90:1410–1424

Cadonna A, Frühwirth-Schnatter S, Knaus P (2020) Triple the gamma: a unifying shrinkage prior for variance and variable selection in sparse state space and TVP models. Econometrics 8(2):20

Celeux G, Frühwirth-Schnatter S, Robert CP (2019) Model selection for mixture models-perspectives and strategies. In: Handbook of mixture analysis. CRC Press, pp 271–307

Chamroukhi F, Huynh BT (2018) Regularized maximum-likelihood estimation of mixture-of-experts for regression and clustering. In: 2018 International joint conference on neural networks (IJCNN). IEEE, pp 1–8
Ding C (2006) Using regression mixture analysis in educational research. Pract Assess Res Eval 11(11):1–11
Dyer WJ, Pleck J, McBride B (2012) Using mixture regression to identify varying effects: a demonstration with paternal incarceration. J Marriage Fam 74(5):1129–1148
Erosheva EA, Fienberg SE, Joutard C (2007) Describing disability through individual-level mixture models for multivariate binary data. Ann Appl Stat 1(2):502–537
Flynt A, Dean N, Nugent R (2019) sARI: a soft agreement measure for class partitions incorporating assignment probabilities. Adv Data Anal Classif 13(1):303–323
Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer, Berlin
Frühwirth-Schnatter S (2019) Keeping the balance-bridge sampling for marginal likelihood estimation in finite mixture, mixture of experts and markov mixture models. Braz J Probab Stat 33(4):706–733
Frühwirth-Schnatter S, Frühwirth R (2010) Data augmentation and MCMC for binary and multinomial logit models. In: Statistical modelling and regression structures. Springer, pp 111–132
Frühwirth-Schnatter S, Pamminger C, Weber A, Winter-Ebmer R (2012) Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. J Appl Economet 27(7):1116–1137
Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis. Second edition. Chapman and Hall/CRC
Gershenfeld N (1997) Nonlinear inference and cluster-weighted modeling. Ann N Y Acad Sci 808(1):18–24
Geweke J, Keane M (2007) Smoothly mixing regressions. J Econom 138(1):252–290
Goldfeld S, Quandt R (1973) The estimation of structural shifts by switching regressions. In: Annals of economic and social measurement, volume 2, number 4, pp 475–485
Gormley IC, Frühwirth-Schnatter S (2019) Mixture of experts models. In: Handbook of mixture analysis. CRC Press, pp 271–307
Gormley IC, Murphy TB (2008) A mixture of experts model for rank data with applications in election studies. Ann Appl Stat 2(4):1452–1477
Gormley IC, Murphy TB (2010) A mixture of experts latent position cluster model for social network data. Stat Methodol 7(3):385–405
Gormley IC, Murphy TB (2011) Mixture of experts modelling with social science applications. In: Mixtures: estimation and applications. Wiley Online Library, pp 101–121
Green PJ, Richardson S (2002) Hidden Markov models and disease mapping. J Am Stat Assoc 97(460):1055–1070
Hastie TJ, Tibshirani RJ (1990) Generalized additive models, vol 43. CRC Press
Huang M, Li R, Wang S (2013) Nonparametric mixture of regression models. J Am Stat Assoc 108(503):929–941
Huang M, Yao W (2012) Mixture of regression models with varying mixing proportions: a semiparametric approach. J Am Stat Assoc 107(498):711–724
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218
Ingrassia S, Minotti SC, Vittadini G (2012) Local statistical modeling via a cluster-weighted approach with elliptical distributions. J Classif 29(3):363–401
Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Comput 3(1):79–87
Khalili A, Chen J (2007) Variable selection in finite mixture of regression models. J Am Stat Assoc 102(479):1025–1038
Lang S, Brezger A (2004) Bayesian P-splines. J Comput Graph Stat 13(1):183–212
Lê Cao K-A, Meugnier E, McLachlan GJ (2010) Integrative mixture of experts to combine clinical factors and gene markers. Bioinformatics 26(9):1192–1198
Li F, Villani M, Kohn R (2010) Flexible modeling of conditional distributions using smooth mixtures of asymmetric student t densities. J Stat Plan Inference 140(12):3638–3654
Li F, Villani M, Kohn R (2011) Modeling conditional densities using finite smooth mixtures. In: Mixtures: estimation and applications. Wiley Online Library, pp 123–144
Lu Z (2006) A regularized minimum cross-entropy algorithm on mixtures of experts for time series prediction and curve detection. Pattern Recognit Lett 27(9):947–955
Malsiner-Walli G, Frühwirth-Schnatter S, Grün B (2016) Model-based clustering based on sparse finite Gaussian mixtures. Stat Comput 26:303–324

McDonald SE, Shin S, Corona R, Maternick A, Graham-Bermann SA, Ascione FR, Williams JH (2016) Children exposed to intimate partner violence: Identifying differential effects of family environment on children's trauma and psychopathology symptoms through regression mixture models. Child Abuse Neglect 58:1–11

Mollica C, Tardella L (2017) Bayesian Plackett–Luce mixture models for partially ranked data. Psychometrika 82(2):442–458

Murphy K, Murphy TB (2020) Gaussian parsimonious clustering models with covariates and a noise component. Adv Data Anal Classif 14:293–325

Murphy K, Viroli C, Gormley IC (2020) Infinite mixtures of infinite factor analysers. Bayesian Anal 15(3):937–963

R Core Team (2020) R: A Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Raftery A, Newton M, Satagopan J, Krivitsky P (2007) Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In: Bayesian statistics 8. Oxford University Press, pp 371–416

Ranciati S, Viroli C, Wit EC (2017) Mixture model with multiple allocations for clustering spatially correlated observations in the analysis of chip-seq data. Biom J 59(6):1301–1316

Redivo E, Nguyen HD, Gupta M (2020) Bayesian clustering of skewed and multimodal data using geometric skewed normal distributions. Comput Stat Data Anal 152:107040

Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). J R Stat Soc: Ser B (statistical methodology) 59(4):731–792

Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. Chapman and Hall/CRC

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). J R Stat Soc: Ser B (statistical methodology) 64(4):583–639

Tashman A, Frey RJ (2009) Modeling risk in arbitrage strategies using finite mixtures. Quant Finance 9(5):495–503

Turner TR (2000) Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. J R Stat Soc: Ser C (Appl Stat) 49(3):371–384

Van Horn ML, Jaki T, Masyn K, Howe G, Feaster DJ, Lamont AE, George MR, Kim M (2015) Evaluating differential effects using regression interactions and regression mixture models. Educ Psychol Measur 75(4):677–714

Villani M, Kohn R, Giordani P (2009) Regression density estimation using smooth adaptive gaussian mixtures. J Econom 153(2):155–173

Wang P, Puterman ML, Cockburn I, Le N (1996) Mixed Poisson regression models with covariate dependent rates. Biometrics 52(2):381–400

Watnik MR (1998) Pay for play: are baseball salaries based on performance? J Stat Educ 6(2)

Wedel M, DeSarbo WS (1993) A latent class binomial logit methodology for the analysis of paired comparison choice data. Decis Sci 24(6):1157–1170

Weigend AS, Shi S (2000) Predicting daily probability distributions of s&p500 returns. J Forecast 19(4):375–392

Xiang S, Yao W (2018) Semiparametric mixtures of nonparametric regressions. Ann Inst Stat Math 70(1):131–154

Xiang S, Yao W, Yang G et al (2019) An overview of semiparametric extensions of finite mixture models. Stat Sci 34(3):391–404

Xu L, Jordan M, Hinton GE (1994) An alternative model for mixtures of experts. In: Advances in neural information processing systems, vol 7. MIT press, pp 633–640

Young DS, Hunter DR (2010) Mixtures of regressions with predictor-dependent mixing proportions. Comput Stat Data Anal 54(10):2253–2266

Zhang Y, Zheng Q (2018) Semiparametric mixture of additive regression models. Commun Stat-Theory Methods 47(3):681–697