Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE STATISTICHE

Ciclo XXXIII

Settore Concorsuale: 13/D1

Settore Scientifico Disciplinare: SECS-S/01 STATISTICA

FLEXIBLE BAYESIAN MODELLING OF CONCOMITANT COVARIATE EFFECTS IN MIXTURE MODELS

Presentata da: Marco Berrettini

Coordinatore Dottorato	**Supervisore**
Prof.ssa Monica Chiogna	**Prof. Giuliano Galimberti**

Esame finale anno 2021

Abstract

Mixture models provide a useful tool to account for unobserved heterogeneity, and are the basis of many model-based clustering methods. In order to gain additional flexibility, some model parameters can be expressed as functions of concomitant covariates. In particular, prior probabilities of latent group membership can be linked to concomitant covariates through a multinomial logistic regression model, where each of these so-called component weights is associated with a linear predictor involving one or more of these variables. In this Thesis, this approach is extended by replacing the linear predictors with additive ones, where the contributions of some/all concomitant covariates can be represented by smooth functions. An estimation procedure within the Bayesian paradigm is proposed. In particular, a data augmentation scheme based on difference random utility models is exploited, and smoothness of the covariate effects is controlled by suitable choices for the prior distributions of the spline coefficients. This methodology is then extended to include flexible covariates effects also on the component densities. The performance of the proposed methodologies is investigated via simulation experiments and applications to real data. The content of the Thesis is organized as follows. In Chapter 1, a literature review about mixture models and mixture models with covariate effects is provided. After a brief introduction on Bayesian additive models with P-splines, the general specification for the proposed method is presented in Chapter 2, together with the associated Bayesian inference procedure. This approach is adapted to the specific case of categorical and continuous manifest variables in Chapter 3 and Chapter 4, respectively. In Chapter 5, the proposed methodology is extended to include flexible covariate effects also in the component densities. Finally, conclusions and remarks on the Thesis are collected in Chapter 6.

Contents

Chapter 1

Introduction

1.1 Mixture models

The relevance of finite mixture models in data analysis is testified by the ever-increasing rate at which articles on theoretical and practical aspects of mixture models appear in the scientific literature. This is because they can be exploited to provide computationally convenient representations for modelling complex distributions of data on statistical phenomena. Fields in which mixture models have been successfully applied include agriculture, astronomy, bioinformatics, biology, economics, engineering, genetics, imaging, marketing, medicine, neuroscience, physics, psychiatry, psychology and social sciences. In these applications, finite mixture models provide a variety of tools, including cluster and latent class analyses, discriminant analysis, image analysis, and survival analysis, in addition to their more direct role of providing models for complex multimodal distributions. Finite mixture models provide a straightforward, but very flexible, extension of homogeneous statistical models. The price to pay for this flexibility is that inference for these models is challenging, because of the discrete latent structure that causes certain technical difficulties in estimation, and the need to decide on the unknown number of groups, states, or clusters. Extensive reviews of mixture models and their application are given in Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), Lindsay (1995), Böhning (1999), McLachlan and Peel (2004), Frühwirth-Schnatter (2006), Mengersen et al. (2011), and McNicholas (2016). In addition, mixture models are addressed in several books involving classification, machine learning, and other fields in multivariate analysis.

1.1.1 General definition

Many statistical models involve finite mixture distributions in some way. Consider a population made up of G subgroups mixed at random in proportion to their group size $\pi_1, \dots, \pi_G$. Assume that interest lies in a **Q**-dimensional random variable or vector $\mathbf{Y}$, whose distribution is heterogeneous across and homogeneous within the subgroups. Random vector $\mathbf{Y}$ takes values in a sample space $\mathcal{Y} \subset \mathcal{R}^Q$, which may be discrete or continuous. The distribution of $\mathbf{Y}$ is generally characterized by its probability density (or mass) function $f(\mathbf{y})$, where $f(\cdot)$ denotes a generic probability density function. Even if the vector $\mathbf{Y}$ is discrete, $f(\mathbf{y})$ can still be viewed as a density, according to a counting measure. Due to heterogeneity, $\mathbf{Y}$ has a different probability density function $f_g(\mathbf{y})$ in each group $g = 1, \dots, G$.

Random variable $\mathbf{Y}$ is said to arise from a finite mixture distribution if the probability density function $f(\mathbf{y})$ takes the following form:

$$f(\mathbf{y}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{y}), \quad \forall \mathbf{y} \in \mathcal{Y}. \tag{1.1}$$

A single density $f_g(\mathbf{y})$ is referred to as the component density, while G denotes the number of components. The parameters $\pi_1, \dots, \pi_G$ are called the (component) weights; the vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_G)$ is sometimes referred to as the (component) weight distribution (Frühwirth-Schnatter, 2006), taking value in the unit simplex $\mathcal{E}_G$ (which is a subspace of $\mathcal{R}^G$), defined by the following constraints:

$$\begin{aligned} &\pi_g \geq 0, \quad g = 1, \dots, G; \\ &\pi_1 + \dots + \pi_G = 1. \end{aligned} \tag{1.2}$$

In most applications one assumes that all component densities arise from the same parametric distribution family with density $f(\mathbf{y}|\boldsymbol{\theta}_g)$, which is taken to be known up to a vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G) \in \Theta^G$. In this case, the mixture can be written as

$$f(\mathbf{y}|\boldsymbol{\psi}) = \sum_{g=1}^{G} \pi_g f(\mathbf{y}|\boldsymbol{\theta}_g), \tag{1.3}$$

where $\boldsymbol{\psi} = (\boldsymbol{\theta}, \pi_1, \dots, \pi_{G-1})$ denotes the vector of unknown parameters taking values in the parameter space $\boldsymbol{\Psi} = \Theta^G \times \mathcal{E}_G$.

1.1.2 Identifiability

The issue of identifiability of a mixture distribution is essential for parameter estimation. A parametric family of densities, indexed by a parameter $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, defined over a sample space $\mathcal{Y}$, is said to be identifiable if any two values $\boldsymbol{\psi}$ and $\boldsymbol{\psi}^*$ in $\boldsymbol{\Psi}$ define the same probability law on $\mathcal{Y}$ if and only if $\boldsymbol{\psi}$ and $\boldsymbol{\psi}^*$ are identical; see, e.g., Rothenberg (1971). In terms of the corresponding probability densities $f(\mathbf{y}|\boldsymbol{\psi})$ and $f(\mathbf{y}|\boldsymbol{\psi}^*)$, this condition can be written as:

$$f(\mathbf{y}|\boldsymbol{\psi}) = f(\mathbf{y}|\boldsymbol{\psi}^*), \text{ for almost all } \mathbf{y} \in \mathcal{Y} \iff \boldsymbol{\psi} = \boldsymbol{\psi}^*. \tag{1.4}$$

Identifiability problems for finite mixture distributions are studied in Teicher (1963), Yakowitz and Spragins (1968) and Chandra (1977); a detailed discussion can be found in Frühwirth-Schnatter (2006, Section 1.3). One of the main issues with identifiabiliy of mixture models is related to the invariance of a mixture distribution to relabeling the components, as first noted by Redner and Walker (1984).

For the general finite mixture distribution with G components defined in Equation 1.3, there exist $G!$ equivalent ways of arranging the components. Therefore, usually identifiability in this context is defined by taking into account this potential issue. Let $f(\mathbf{y}|\boldsymbol{\psi}) = \sum_{g=1}^{G} \pi_g f(\mathbf{y}_g|\boldsymbol{\theta}_g)$ and $f(\mathbf{y}|\boldsymbol{\psi}^*) = \sum_{g=1}^{G^*} \pi_g^* f(\mathbf{y}_g|\boldsymbol{\theta}_g^*)$ be any two members of a parametric family of mixture densities. This class of finite mixtures is said to be identifiable for $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, if $f(\mathbf{y}|\boldsymbol{\psi}) = f(\mathbf{y}|\boldsymbol{\psi}^*)$ for almost all $\mathbf{y} \in \mathcal{Y}$, if and only if, $G = G^*$ and the component labels can be permuted so that $\pi_g = \pi_g^*$ and $f(\mathbf{y}_g|\boldsymbol{\theta}_g) = f(\mathbf{y}_g|\boldsymbol{\theta}_g^*)$, for $g = 1, \ldots, G$.

The lack of identifiability of $\boldsymbol{\psi}$ due to the interchanging of component labels can be overcome by the imposition of an appropriate constraint on $\boldsymbol{\psi}$ (e.g. ordering parameters). In the Bayesian context, this issue is referred to as the label-switching problem. Besides the theoretical issues related to identifiability, label switching can also be observed in practice during sampling.

1.1.3 Hierarchical representation

Any standard finite mixture model may be described as a hierarchical latent variable model, where the distribution of the observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ depends on a hidden G-dimensional label vector $\mathbf{D}_i = (D_{1i}, \ldots, D_{Gi})$, whose g-th element D_{gi} is defined to be one or zero, according to whether the component the i-th unit comes from is the g-th or not, for $i = 1, \ldots, n$.

On a first layer of the model, the joint sampling distribution of $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is specified conditional on the whole sequence of indicators $\mathbf{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_n)$:

$$f(\mathbf{y}|\mathbf{D}, \boldsymbol{\psi}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\mathbf{D}_i, \boldsymbol{\psi}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\psi}_{\mathbf{D}_i}). \tag{1.5}$$

In the standard finite mixture model it is assumed that the component indicators $\mathbf{D}_1, \ldots, \mathbf{D}_n$ are independent, and their joint distribution can be written as:

$$f(\mathbf{D}|\boldsymbol{\pi}) = \prod_{i=1}^{n} f(\mathbf{D}_i|\boldsymbol{\pi}), \tag{1.6}$$

with $\Pr(D_{gi} = 1|\boldsymbol{\pi}) = \pi_g$. Titterington (1990) proposed the name "hidden multinomial model" for the standard finite mixture model, because $\mathbf{D}_i$ is distributed according to a multinomial distribution consisting of one draw on G categories with probabilities $\pi_1, \ldots, \pi_G$; that is,

$$\Pr(\mathbf{D}_i = \mathbf{d}_i|\boldsymbol{\pi}) = \prod_{g=1}^{G} \pi_g^{d_{gi}}, \quad \sum_{g=1}^{G} d_{gi} = 1, \tag{1.7}$$

or $\mathbf{D}_i \sim MulNom_G(1; \boldsymbol{\pi})$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$. Equivalently, a categorical random variable C_i, $i = 1, \ldots, n$, may be defined, taking on the value c_i among $1, \ldots, G$ with probabilities $\pi_1, \ldots, \pi_G$, respectively. These two representations will be used interchangeably throughout the Thesis, since $D_{gi} = \mathbb{1}(c_i = g)$, where $\mathbb{1}(\cdot)$ denotes the indicator function.

1.1.4 Classification for known component parameters

Assume that n observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ are randomly drawn from a finite mixture of distributions with density $f(\mathbf{y}|\boldsymbol{\psi})$ indexed by a parameter $\boldsymbol{\psi} \in \boldsymbol{\Psi}$; these observations should be used to make inferences about the underlying group structure. Assume also that the resulting finite mixture model,

$$f(\mathbf{y}_i|\boldsymbol{\psi}) = \sum_{g=1}^{G} \pi_g f(\mathbf{y}_i|\boldsymbol{\theta}_g), \tag{1.8}$$

is known exactly, with precise values assigned to the number of components G, the component parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G$ and the weight distribution $\boldsymbol{\pi}$, and the only challenge is to classify the set of n observations $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ into each component. This classification problem is a common and old issue; see Cormack (1971), McLachlan and Basford (1988) and Everitt et al. (2001) for a review.

Classification of a single observation $\mathbf{y}_i$ aims at deriving the conditional probability $\Pr(D_{gi} = 1|\mathbf{y}_i; \boldsymbol{\psi})$ of the event $\{D_{gi} = 1\}$, having observed the event $\{\mathbf{Y}_i = \mathbf{y}_i\}$. Bayes' rule shows how to compute this probability for each component and observation from a discrete mixture distribution:

$$\Pr(D_{gi} = 1|\mathbf{Y}_i = \mathbf{y}_i; \boldsymbol{\psi}) = \frac{\Pr(D_{gi} = 1|\boldsymbol{\psi})\Pr(\mathbf{Y}_i = \mathbf{y}_i|D_{gi} = 1; \boldsymbol{\psi})}{\sum_{g=1}^{G} \Pr(D_{gi} = 1|\boldsymbol{\psi})\Pr(\mathbf{Y}_i = \mathbf{y}_i|D_{gi} = 1; \boldsymbol{\psi})}, \tag{1.9}$$

where $\Pr(D_{gi} = 1|\boldsymbol{\psi})$ is the prior probability that the i-th observation $\mathbf{y}_i$ comes from class g, and is equal to the class size π_g. For a discrete mixture, $\Pr(\mathbf{Y}_i = \mathbf{y}_i|D_{gi} = 1; \boldsymbol{\psi})$ is obtained from the component-specific probability density function $f(\mathbf{y}_i|\boldsymbol{\theta}_g)$. Thus, it is convenient to rewrite Bayes' rule, given in Equation (1.9), in the following way:

$$\Pr(D_{gi} = 1|\mathbf{Y}_i = \mathbf{y}_i; \boldsymbol{\psi}) = \frac{\pi_g f(\mathbf{y}_i|\boldsymbol{\theta}_g)}{f(\mathbf{y}_i|\boldsymbol{\psi})}, \tag{1.10}$$

as this result also holds when dealing with observations from continuous rather than discrete mixtures. The denominator in Equation (1.10) remains the same, whatever the value of g, and is equal to the sum of the numerators for $g = 1, \ldots, G$. For this reason, Bayes' rule is usually formulated up to proportionality:

$$\Pr(D_{gi} = 1|\mathbf{Y}_i = \mathbf{y}_i; \boldsymbol{\psi}) \propto \pi_g f(\mathbf{y}_i|\boldsymbol{\theta}_g). \tag{1.11}$$

The right-hand side is evaluated for each group, and the resulting values are normalized, to obtain a proper posterior distribution.

1.1.5 Maximum likelihood estimation

In this Section, it is assumed that the true number of distinct components G and the parametric distribution family of the component densities in the mixture distribution (1.8) are known, while the weight distribution $\boldsymbol{\pi}$, the component parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G$, and the indicator variables $\mathbf{D}$ are not. In this case, estimation of the parameters of the mixture distribution is not straightforward, since no method leads to an analytical solution. Thus, some computational procedure is required in practical estimation. The method of moments was the most widely applied estimation technique in the early days. With the availability of powerful computers and elaborated numerical algorithms, maximum likelihood (ML) method became the preferred one for parameter estimation in finite mixture models.

Let $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be n independent and randomly selected observations from the mixture distribution in Equation (1.8), and let $\boldsymbol{\psi} = (\boldsymbol{\theta}, \pi_1, \ldots, \pi_{G-1})$ denote all the unknown parameters in the mixture model that need to be estimated from the data. The mixture likelihood function $L(\boldsymbol{\psi}|\mathbf{y})$ is defined as the joint distribution of $\mathbf{y}_1, \ldots, \mathbf{y}_n$, considered as a function of $\boldsymbol{\psi}$:

$$L(\boldsymbol{\psi}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\psi}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f(\mathbf{y}_i|\boldsymbol{\theta}_g). \tag{1.12}$$

The ML estimator $\hat{\boldsymbol{\psi}}$ is obtained by maximizing the mixture likelihood $L(\boldsymbol{\psi}|\mathbf{y})$, defined in Equation (1.12), with respect to $\boldsymbol{\psi}$, i.e.

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} L(\boldsymbol{\psi}|\mathbf{y}). \tag{1.13}$$

A first issue arises for the important special case of mixtures of normal distributions, whose corresponding mixture likelihood is unbounded. This was first noted by Kiefer and Wolfowitz (1956), for the following mixture of two normal distributions:

$$Y \sim (1 - \pi_2)N(\mu, 1) + \pi_2 N(\mu, \sigma_2^2), \tag{1.14}$$

where π_2 is fixed, whereas the mean μ and the variance σ_2^2 are unknown. In this example, each observation in an arbitrary data set $\mathbf{y} = (y_1, \ldots, y_n)$, of arbitrary sample size n, gives rise to a singularity in the mixture likelihood function, since

$$\lim_{\sigma_2^2 \to 0} f(\mathbf{y}|\mu = y_i, \sigma_2^2) = \infty, \quad \text{for any } i = 1, \ldots, n. \tag{1.15}$$

The unboundedness of the mixture likelihood function is also relevant for the mixture of multivariate normal distributions, as first noted by Day (1969). Thus, the ML estimator as global maximizer of the mixture likelihood function does not always exist. Nevertheless, under certain boundedness conditions on the partial derivatives of $L(\boldsymbol{\psi}|\mathbf{y})$ with respect to the components of $\boldsymbol{\psi}$, Redner and Walker (1984, p. 211) prove that, in any sufficiently small neighborhood of the true parameter vector $\boldsymbol{\psi}^{\text{true}}$, for a sufficiently large sample size n, there exists a unique solution of the likelihood equation

$$\frac{\partial}{\partial \boldsymbol{\psi}} L(\boldsymbol{\psi}|\mathbf{y}) = 0, \tag{1.16}$$

which locally maximizes the log-likelihood function. Moreover, provided that certain regularity conditions hold (Casella and Berger, 2002), this ML estimator $\hat{\boldsymbol{\psi}}$ is consistent, efficient and asymptotically normal, i.e.

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^{\text{true}}) \to_d MVN(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\psi})), \tag{1.17}$$

where $\rightarrow_d$ denotes convergence in distribution and $\mathcal{I}(\boldsymbol{\psi}^{\text{true}})$ is the expected Fisher information matrix, defined as

$$\mathcal{I}^{-1}(\boldsymbol{\psi}) = -\int_{\mathcal{Y}} \left(\frac{\partial}{\partial \boldsymbol{\psi}} \log L(\boldsymbol{\psi}|\mathbf{y}) \right) \left(\frac{\partial}{\partial \boldsymbol{\psi}} \log L(\boldsymbol{\psi}|\mathbf{y}) \right)' L(\boldsymbol{\psi}|\mathbf{y}) \, \mathrm{d}\mathbf{y}. \tag{1.18}$$

Redner et al. (1981) and Atienza et al. (2007) provide a detailed discussion on the properties of ML estimators in the mixture models context.

ML estimation for finite mixture models was initially performed using direct method such as the Newton (Hasselblad, 1966) or the gradient method (Quandt, 1972). Nowadays, the Expectation-Maximization (EM) algorithm introduced by Dempster et al. (1977) is probably the most commonly applied method to find the ML estimator in finite mixture models; see Redner and Walker (1984) for a review. A disadvantage of this algorithm, compared to direct maximization of the likelihood function, is its much slower convergence rate. In order to overcome this issue, several authors use hybrid algorithms that combine the EM algorithm with Newton's method for mixture estimation; see, for example, Aitkin and Aitkin (1996). McLachlan and Peel (2004) give a thorough discussion of non-Bayesian parameter estimation for finite mixtures, with emphasis on ML estimation based on the EM algorithm. They also warn that the sample size n has to be very large before asymptotic theory of maximum likelihood applies, particularly for mixture models, and that, furthermore, the regularity conditions are often violated.

1.1.6 Bayesian parameter estimation

From a Bayesian perspective, all information contained in the data $\mathbf{y}$ about $\boldsymbol{\psi}$ is summarized into the posterior density, which is derived using Bayes' theorem, by combining the data-dependent mixture likelihood function $f(\mathbf{y}|\boldsymbol{\psi})$ in Equation (1.12) with a prior density $f(\boldsymbol{\psi})$:

$$f(\boldsymbol{\psi}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\psi}) f(\boldsymbol{\psi}), \tag{1.19}$$

where the posterior density may be known up to a normalizing constant given by $f(\mathbf{y})$. For Bayesian estimation, it is necessary to elicit the prior distribution $f(\boldsymbol{\psi})$. Specifying a prior distribution is often a subjective task, where the user selects a (usually, proper) density over the parameters to represent their knowledge - and uncertainty - about the phenomenon prior to observing data. For finite mixture models, such priors are usually obtained by adopting distributions that are conjugate to the complete-data likelihood function. First, it is often assumed that the parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G$ are independent of the weight distribution $\boldsymbol{\pi}$:

$$f(\boldsymbol{\psi}) = f(\boldsymbol{\pi}) f(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G). \tag{1.20}$$

For finite mixture models, a common choice for the prior to assign to the weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$ is the Dirichlet distribution $Dir(\delta_0, \ldots, \delta_0)$ with hyperparameters assumed to be the same, leading to an invariant and flat prior. The nature of the prior for the component parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G$ depends on the distribution family underlying the mixture distribution. To formulate a joint prior for $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G$, the parameters are assumed to be independent a priori, conditional on some hyperparameter $\boldsymbol{\phi}$:

$$f(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G | \boldsymbol{\phi}) = \prod_{g=1}^{G} f(\boldsymbol{\theta}_g | \boldsymbol{\phi}). \tag{1.21}$$

Results from a Bayesian analysis of finite mixture models using subjective prior information may be sensitive to particular choices of hyperparameters: to reduce this sensitivity, it is common practice to use hierarchical priors in the context of finite mixture modeling. Such priors treat the hyperparameter $\boldsymbol{\phi}$ as an unknown quantity with a prior of its own:

$$f(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, \boldsymbol{\phi}) = f(\boldsymbol{\phi}) \prod_{g=1}^{G} f(\boldsymbol{\theta}_g | \boldsymbol{\phi}). \tag{1.22}$$

In any case, the prior distribution has to be selected with some care.

1.1.7 Data augmentation and the MCMC algorithm

Following Dempster et al. (1977), any mixture model may be seen as an incomplete data problem by considering the component indicators $\mathbf{D}$ as missing data. The benefit of this data augmentation (Tanner and Wong, 1987) is that, conditional on $\mathbf{D}$, priors that are conjugate to the complete-data likelihood become available. The sampling distribution $f(\mathbf{y}, \mathbf{D} | \boldsymbol{\psi})$ of the complete data $(\mathbf{y}, \mathbf{D})$, regarded as a function of the unknown parameter $\boldsymbol{\psi}$, can be specified by exploiting the hierarchical representation of a finite mixture model, given in Section 1.1.3, as follows:

$$f(\mathbf{y}, \mathbf{D} | \boldsymbol{\psi}) = f(\mathbf{y} | \mathbf{D}, \boldsymbol{\psi}) f(\mathbf{D} | \boldsymbol{\psi}) = \prod_{i=1}^{n} f(\mathbf{y}_i | \mathbf{D}_i, \boldsymbol{\psi}) f(\mathbf{D}_i | \boldsymbol{\psi}). \tag{1.23}$$

Since $f(\mathbf{y}_i | D_{gi} = 1, \boldsymbol{\psi}) = f(\mathbf{y}_i | \boldsymbol{\theta}_g)$ and $\Pr(D_{gi} = 1 | \boldsymbol{\psi}) = \pi_g$, the complete-data likelihood function in (1.23) can be rewritten as

$$f(\mathbf{y}, \mathbf{D} | \boldsymbol{\psi}) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\pi_g f(\mathbf{y}_i | \boldsymbol{\theta}_g) \right]^{D_{gi}}. \tag{1.24}$$

Bayesian inference on a general mixture model through data augmentation explores the augmented parameter space of $(\mathbf{D}, \boldsymbol{\psi})$ by sampling from the complete-data posterior distribution $f(\mathbf{D}, \boldsymbol{\psi}|\mathbf{y})$, given by

$$f(\mathbf{D}, \boldsymbol{\psi}|\mathbf{y}) \propto f(\mathbf{y}, \mathbf{D}|\boldsymbol{\psi}) f(\boldsymbol{\psi}), \tag{1.25}$$

with the complete-data likelihood $f(\mathbf{y}, \mathbf{D}|\boldsymbol{\psi})$ defined as in Equation (1.24). Furthermore, conditional on knowing the parameter vector $\boldsymbol{\psi}$, the posterior distribution of the component indicators takes a very simple form, as for the classification problem studied in Section 1.1.4.

It is then quite straightforward to sample from the posterior in Equation (1.25) using Markov chain Monte Carlo (MCMC) methods, in particular Gibbs sampling (Geman and Geman, 1984), where $\boldsymbol{\psi}$ is sampled conditional on knowing $\mathbf{D}$, and $\mathbf{D}$ is sampled conditional on knowing $\boldsymbol{\psi}$. Pioneering papers realizing the importance of Gibbs sampling for Bayesian estimation in mixture models are Evans et al. (1992), Smith and Roberts (1993), Diebolt and Robert (1994) and Escobar and West (1995). Some authors, e.g. Celeux et al. (2000), use a Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953) to simulate $\boldsymbol{\psi}$ from the mixture posterior $f(\boldsymbol{\psi}|\mathbf{y})$ by iteratively proposing a new parameter from an arbitrary proposal density. In their seminal paper, Richardson and Green (1997) suggest applying the reversible jump MH algorithm, introduced by Green (1995), to select the number of components in a mixture model, whereas Stephens (2000) apply birth-and-death MCMC methods; see Green (2003) for a review about trans-dimensional methods.

The posterior in Equation (1.25), can be sampled through the following MCMC scheme, formulated for the general case, where each observation $\mathbf{y}_i$ may also be multivariate. The hierarchical prior in Equation (1.22) is imposed on the component parameters $\boldsymbol{\theta}$. The unconstrained algorithm starts with some classification $\mathbf{D}^{(0)}$ and by selecting a starting value for the hyperparameters $\boldsymbol{\phi}^{(0)}$. Then the following steps have to be repeated for $t = 1, \ldots, T_0, \ldots, T + T_0$:

1. parameter simulation conditional on the classification $\mathbf{D}^{(t-1)}$:

   (a) sample $\boldsymbol{\pi}^{(t)}$ from the Dirichlet distribution $Dir(\delta_1(\mathbf{D}^{(t-1)}), \ldots, \delta_G(\mathbf{D}^{(t-1)}))$, where $\delta_g(\mathbf{D}^{(t-1)})$ is given by

   $$\delta_g(\mathbf{D}^{(t-1)}) = \delta_0 + \sum_{i=1}^{n} D_{gi}^{(t-1)}; \tag{1.26}$$

   (b) sample the component parameters $\boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_G^{(t)}$ from the complete-data posterior $f(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G|\mathbf{D}^{(t-1)}, \mathbf{y}, \boldsymbol{\phi}^{(t-1)})$;

(c) sample the hyperparameter $\boldsymbol{\phi}^{(t)}$ from $f(\boldsymbol{\phi}|\boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_G^{(t)})$;

2. classification of each observation $\mathbf{y}_i$ conditional on knowing $\boldsymbol{\psi}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_G^{(t)}, \boldsymbol{\pi}^{(t)})$: sample $\mathbf{D}_i^{(t)}$ independently, for $i = 1, \ldots, n$, from the conditional posterior distribution $f(\mathbf{D}_i|\boldsymbol{\psi}^{(t)}, \mathbf{y}_i)$, which is given by

$$\Pr(D_{gi} = 1|\boldsymbol{\psi}^{(t)}, \mathbf{y}_i) \propto \pi_g^{(t)} f(\mathbf{y}_i|\boldsymbol{\theta}_g^{(t)}); \tag{1.27}$$

3. increase t by one, and return to step 1.

Finally, the first T_0 draws, corresponding to the so called burn-in phase, are discarded. Practical MCMC convergence diagnostics for finite mixture models are discussed in Robert et al. (1999). For further details and examples of MCMC algorithms for mixture models see also Frühwirth-Schnatter (2006).

A deterministic alternative to MCMC techniques is provided by methods from machine learning that approximate probability densities through optimization, known as variational approximations. Thanks to the reduced computational burden and the resulting speed, this approach has been successfully applied in the mixture models framework; see, for instance, Wang et al. (2006) and McGrory and Titterington (2007). Neverthless, differently from MCMC methods, variational inference does not provide guarantees of producing asymptotically exact samples from the target density, as it can only find a density close to the target. See Blei et al. (2017) for a recent review.

1.1.8 Model selection

While ratio test is usually one of the preferred likelihood-based methods for selecting parametric models, its application to model selection in finite mixture models creates some difficulty, as basic conditions of incompatibility between the spaces for the null and the alternative hypotheses are not met. For selecting the dimension (number of components) of the model, one may look at the general index proposed by Akaike (1974), known as Akaike information criterion (AIC), that accounts for model complexity. He suggests to pick the model $\mathcal{M}_G$ that minimizes

$$\mathrm{AIC}(G) = -2 \log f(\mathbf{y}|\hat{\boldsymbol{\psi}}_{(G)}, \mathcal{M}_G) + 2\ \dim(\boldsymbol{\psi}_{(G)}), \tag{1.28}$$

where $\hat{\boldsymbol{\psi}}_{(G)}$ is the maximum-likelihood estimator of the parameters in model $\mathcal{M}_G$ and $\dim(\boldsymbol{\psi}_{(G)})$ is the "global" dimension of the model, which acts as a correction term by introducing a penalty for high-dimensional models that provide little additional fit in comparison to simpler models. Without the term $\dim(\boldsymbol{\psi}_{(G)})$, one would choose the model that maximizes the likelihood function. The AIC is independent of the sample size, but it has been shown to be inconsistent as it favors models that overfit the data, even asymptotically; see Bozdogan (1987) and Leroux (1992).

An alternative to AIC is the Bayesian information criterion (BIC), proposed by Schwarz (1978). This index is defined by

$$\mathrm{BIC}(G) = -2\log f(\mathbf{y}|\hat{\boldsymbol{\psi}}_{(G)}, \mathcal{M}_G) + (\log n)\dim(\boldsymbol{\psi}_{(G)}). \tag{1.29}$$

Keribin (2000) proved that BIC asimptotically does not overestimate the "true" number of components, if the component density is correctly specified. If this condition does not hold, BIC tends to select a larger number of components, as shown by simulation studies reported in Biernacki and Govaert (1997) and Biernacki et al. (2000). Olteanu and Rynkiewicz (2011) proved that the consistency property of the BIC holds also for the class of mixture of experts models, which will be introduced in Section 1.2.

Several classification-based information criteria for choosing the number of components have been developed for finite mixture models in the clustering context. Some of these criteria involve entropy $\mathrm{EN}(G, \hat{\boldsymbol{\psi}}_{(G)})$, defined as follows:

$$\mathrm{EN}(G, \hat{\boldsymbol{\psi}}_{(G)}) = \sum_{g=1}^{G}\sum_{i=1}^{n} \Pr(D_{gi} = 1|\mathbf{y}_i, \hat{\boldsymbol{\psi}}_{(G)}) \log \Pr(D_{gi} = 1|\mathbf{y}_i, \hat{\boldsymbol{\psi}}_{(G)}). \tag{1.30}$$

In particular, the classification likelihood criterion (CLC) by Biernacki and Govaert (1997) penalizes the log-likelihood function by entropy $\mathrm{EN}(G, \hat{\boldsymbol{\psi}}_{(G)})$ rather than complexity:

$$\mathrm{CLC}(G) = -2\log f(\mathbf{y}|\hat{\boldsymbol{\psi}}_{(G)}, \mathcal{M}_G) + 2\ \mathrm{EN}(G, \hat{\boldsymbol{\psi}}_{(G)}). \tag{1.31}$$

The ICL-BIC criterion (Biernacki et al., 2000) considers both model complexity and entropy to penalize the log-likelihood function:

$$\text{ICL-BIC}(G) = \mathrm{BIC}(G) + 2\ \mathrm{EN}(G, \hat{\boldsymbol{\psi}}_{(G)}). \tag{1.32}$$

The ICL-BIC has been studied in a detailed comparison with the BIC provided by Baudry et al. (2015).

An alternative popular criterion for Bayesian model selection is the deviance information criterion (DIC) (Spiegelhalter et al., 2002), which is defined on the general principle of balancing goodness of fit and complexity, where the former is measured by the log-likelihood, whereas the latter is approximated by looking at the difference between the posterior mean deviance and the deviance of the posterior means. However, as discussed in

Celeux et al. (2006), the application of DIC to finite mixture models is not without problems, since the inclusion of discrete latent quantities in the data generating process calls for different definitions of the DIC.

Raftery et al. (2007) leverage the (approximate) posterior distribution of the log-likelihood to derive simulation-based anologous to AIC and BIC model selection criteria, called AICM and BICM. AICM's formula is very simple, since it involves only the likelihoods from the posterior simulation iterations:

$$\mathrm{AICM}(G) = -\frac{1}{(T-T_0)}\sum_{t=T_0+1}^{T} 2\log f(\mathbf{y}|\boldsymbol{\psi}^{(t)}, \mathcal{M}_G) + 2s_l^2(\mathcal{M}_G), \tag{1.33}$$

where $s_l^2(\mathcal{M}_G)$ is the variance of the log-likelihood of model $\mathcal{M}_G$ computed on the posterior sample. The following BICM expression, instead, requires some caution due to the presence of the sample size n:

$$\mathrm{BICM}(G) = -\frac{2}{(T-T_0)}\sum_{t=T_0+1}^{T} \log f(\mathbf{y}|\boldsymbol{\psi}^{(t)}, \mathcal{M}_G) + 2(\log n - 1)s_l^2(\mathcal{M}_G). \tag{1.34}$$

The issue with this criterion is that sample size n is not always well-defined, for example in hierarchical models, since the BICM refers to independent units and, therefore, it may not be able to handle correlated observations. Berger et al. (2003) and Pauler (1998)provide a discussion about this topic. Following Pauler (1998), Raftery et al. (2007) propose a modified definition of BICM, that requires the evaluation of the effective sample size involved in the estimation of each unknown parameter.

1.2 Mixture of experts models

Mixtures of experts (MoE) models provide a way to extend mixture models, allowing the model parameters to depend on concomitant covariate information. The terminology "mixture of experts" includes a wide class of mixture models. Indeed, although this nomenclature arises from the machine-learning literature (Jacobs et al., 1991), the class of mixture of experts models were already present in the statistical literature under the form of switching regression models (Quandt, 1972), concomitant variable latent-class models (Dayton and Macready, 1988), clusterwise regression models (DeSarbo and Cron, 1988) and mixed models (Wang et al., 1996).

The MoE framework makes flexible modelling easy, allowing a wide range of application. Among others, MoE models for rank data (Gormley et al., 2008), for network data (Gormley and Murphy, 2010b), for time series data (Frühwirth-Schnatter et al., 2012), for non-normal data (Chamroukhi, 2015; Nguyen and McLachlan, 2016) and for longitudinal data (Tang and Qu, 2016) have been developed. See Nguyen and Chamroukhi (2018) for a recent review on MoE models.

1.2.1 The family of mixture of experts models

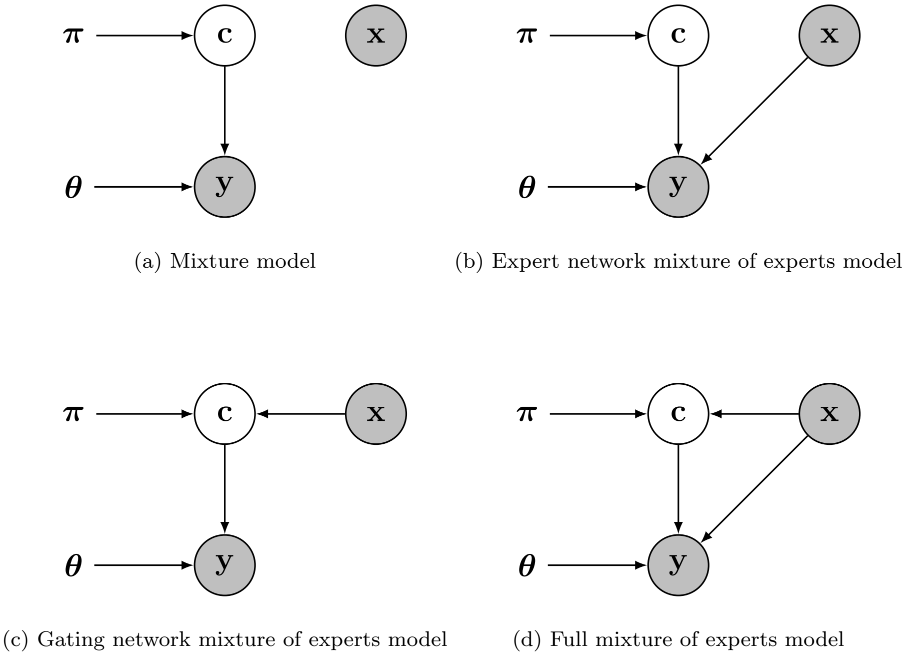
Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be an independent and identically distributed sample of outcomes from a population modelled by a G component finite mixture model. Depending on the application context, the outcome variable can be univariate or multivariate, discrete or continuous, or with a more complex structure. Each observation $i = 1, \ldots, n$ has J associated covariates, which are denoted $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_J})$. MoE models extend finite mixture models by allowing model parameters to be functions of the concomitant variables $\mathbf{x}_i$.

Any mixture model which incorporates covariates or concomitant variables falls within the mixture of experts framework. Figure 1.1 shows the graphical model representation (dependence graph) of the suite of four models in the MoE framework, freely adapted from Murphy and Murphy (2019). The representation in Figure 1.1 involves the latent cluster membership of each outcome variable, denoted by $\mathbf{c} = (c_1, \ldots, c_n)$, where $c_i = g$ if observation i belongs to cluster g, as introduced in Section 1.1.3. Following Murphy and Murphy (2019), the four different classes of models represented in Figure 1.1 can be interpreted as:

(a) mixture model: the outcome variable distribution depends on the latent cluster membership variable $\mathbf{c}$. The model is independent of the covariates; i.e. $f(\mathbf{y}_i, c_i) = \pi_{c_i} f(\mathbf{y}_i | \boldsymbol{\theta}_{c_i})$;

(b) expert network mixture of experts model: the outcome variable distribution depends on both the covariates $\mathbf{x}$ and the latent cluster membership variable $\mathbf{c}$; the distribution of the latent variable is independent of the covariates; i.e $f(\mathbf{y}_i, c_i | \mathbf{x}_i) = \pi_{c_i} f(\mathbf{y}_i | \boldsymbol{\theta}_{c_i}(\mathbf{x}_i))$. This class of models is also known in the literature as mixture of regression models (Frühwirth-Schnatter, 2006, Chapter 8).

(c) gating network mixture of experts model: the outcome variable distribution depends on the latent cluster membership variable $\mathbf{c}$ and the distribution of the latent variable depends on the covariates $\mathbf{x}$; i.e. $f(\mathbf{y}_i, c_i | \mathbf{x}_i) = \pi_{c_i}(\mathbf{x}_i) f(\mathbf{y}_i | \boldsymbol{\theta}_{c_i})$. For discrete outcome variables, these models are also referred to as concomitant-variables latent class model (Dayton and Macready, 1988).



(a) Mixture model

(b) Expert network mixture of experts model

(c) Gating network mixture of experts model

(d) Full mixture of experts model

Figure 1.1: Graphical model representation of the mixture of experts models. Depending on the presence (or absence) of edges between the covariates $\mathbf{x}$, the latent variable $\mathbf{c}$ and the response variable $\mathbf{y}$, four different models can be defined within the MoE framework; grey-colored circle represents observed quantities. Freely adapted from Murphy and Murphy (2019).

(d) Full mixture of experts model: the outcome variable distribution depends on both the covariates $\mathbf{x}$ and on the latent cluster membership variable $\mathbf{c}$. Additionally, the distribution of the latent variable depends on the covariates $\mathbf{x}$; i.e $f(\mathbf{y}_i, c_i|\mathbf{x}_i) = \pi_{c_i}(\mathbf{x}_i) f(\mathbf{y}_i|\boldsymbol{\theta}_{c_i}(\mathbf{x}_i))$. In this case, it is crucial to carefully design the dependence graph, as the general model could suffer from non-identifiability.

MoE models can be considered as members of the class of conditional mixture models (Bishop, 2006), since, for a given set of covariates $\mathbf{x}_i$, the distribution of $\mathbf{y}_i$ is a finite mixture model. Jacobs et al. (1991) consider the component densities $f(\mathbf{y}_i|\boldsymbol{\theta}_g(\mathbf{x}_i))$ as the experts, which model different parts of the input space, and the component weights $\pi_g(\mathbf{x}_i)$ as the gating networks, hence the mixture of experts terminology. The way the different models within the MoE framework depend on the covariates is typically application specific. In particular, Jacobs et al. (1991) model the component weights using a multinomial logit regression model. Arbitrarily selecting a "reference" class – for example, the G-th –, one can assume that the log-odds of the latent class (prior) membership π_g, with respect to that class G, are linear functions of the covariates:

$$\log \frac{\pi_g(\mathbf{x}_i)}{\pi_G(\mathbf{x}_i)} = \mathbf{x}_i' \boldsymbol{\gamma}_g, \quad g = 1, \ldots, G-1, \tag{1.35}$$

where $\boldsymbol{\gamma}_g = (\gamma_{g0}, \gamma_{g1}, \ldots, \gamma_{gJ})$ denotes the vector of coefficients corresponding to the g-th latent class. To make the model identifiable, $\boldsymbol{\gamma}_G$ is set equal to the null vector. Following some simple algebra, the Equation (1.35) can be rewritten as:

$$\pi_g(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i' \boldsymbol{\gamma}_g)}{1 + \sum_{g=1}^{G-1} \exp(\mathbf{x}_i' \boldsymbol{\gamma}_g)}, \quad g = 1, \ldots, G-1. \tag{1.36}$$

It is worth noting that MoE models do not necessarily require the use of a multinomial logit model to express the component weights as functions of the covariates. Geweke and Keane (2007) employ a model similar to an MoE model, where the component weights have a multinomial probit structure. See also Xu et al. (1995) and Nguyen et al. (2019) for alternative solutions based on Gaussian gating functions. The form of the distribution $f(\mathbf{y}_i|\boldsymbol{\theta}_g(\mathbf{x}_i))$ depends on the type of outcome data under study.

1.2.2 Statistical inference

Jacobs et al. (1991) and Jordan and Jacobs (1994)derive maximum likelihood (ML) estimates for MoE models via the expectation-maximization (EM) algorithm. The EM algorithm for fitting MoE models is straightforward in principle, but the M step is usually more difficult in practice than the M step for standard mixture models. This is usually due to a complex component density and/or component weights model, or a large parameter set. A modified version of the EM algorithm, the expectation-and-conditional maximization (ECM) algorithm proposed by Meng and Rubin (1993) is therefore often employed, where the likelihood can be fruitfully factorized and the parameter estimates depend on each other. In the ECM algorithm, the M step consists of a series of conditional maximization steps. Because in the context of MoE models no closed form expression is available for the ML estimator of the parameters $\boldsymbol{\gamma}_g$, the conditional M step requires the use of a numerical optimization technique, or, as in Gormley and Murphy (2008) and Nguyen and McLachlan (2016), a minorization-maximization (MM) algorithm (Hunter and Lange, 2004). The class of MM algorithms, in which a minorizing function is iteratively maximized and updated, is thoroughly covered in Lange (2016). Alternatively, one may consider estimating parameters $\boldsymbol{\gamma}_g$ only at convergence, as in Vermunt (2010).

Estimation of MoE models can also be achieved within the Bayesian paradigm, either via a variational approach (Bishop and Svensén, 2012) or using a Markov chain Monte Carlo (MCMC) algorithm. The latter approach is used, for example, in Peng et al. (1996), Gormley and Murphy (2010a) and Frühwirth-Schnatter et al. (2012). Both the Gibbs sampler (Geman and Geman, 1984) and the Metropolis-Hastings (Metropolis et al., 1953) algorithm are typically required. Again, the specific MCMC algorithm, and the form of the prior distributions, depend on the nature of the MoE model under study and on the type of the response.

As is standard in Bayesian estimation of mixture models (Diebolt and Robert, 1994), fitting MoE models is greatly simplified by augmenting the observed data with the latent group indicator variable $\mathbf{c} = (c_1, \ldots, c_n)$, or, equivalently, component indicators $\mathbf{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_n)$, with $D_{gi} = \mathbb{1}(c_i = g)$, for $g = 1, \ldots, G$. Nevertheless, performing inference on MoE models can be difficult in practice. A conditional multivariate normal prior is an intuitive prior for the model parameters $\boldsymbol{\gamma}_g$, $g = 1, \ldots, G-1$, in the component weights, but it is non-conjugate. Hence, the full conditional distribution is not available in closed form and a Metropolis-Hastings (MH) step is required, at least, to sample the component weight parameters. This solution brings issues such as choosing suitable proposal distributions and tuning parameters. Gormley and Murphy (2010b) detail an approach for deriving proposal distributions with attractive properties, within the context of a MoE model for network data.

Alternatively, Frühwirth-Schnatter et al. (2012) exploit data augmentation of the multinomial logit regression (MNLR) model, based on the difference random utility model (dRUM) representation, in the context of MoE models. As previously shown by Frühwirth-Schnatter and Frühwirth (2010), the MNLR model has the following representation as a binary logit model, conditional on knowing $\lambda_{li} = \exp(\mathbf{x}_i'\gamma_l)$ for all $l \neq g$:

$$\begin{aligned} z_{gi} &= \mathbf{x}_i'\boldsymbol{\gamma}_g - \log\left(\sum_{l\neq g} \lambda_{li}\right) + \epsilon_{gi}, \\ D_{gi} &= \mathbb{1}_{(}z_{gi} > 0), \end{aligned} \tag{1.37}$$

where z_{gi} is a latent variable, ϵ_{gi} are i.i.d. errors following a logistic distribution and D_{gi} is the binary outcome variable introduced in Section 1.1.3. Given $\lambda_{1i}, \ldots, \lambda_{G-1,i}$ and $\mathbf{D}_i$, the latent variables $(z_{1i}, \ldots, z_{G-1,i})$ follow exponential distributions and can be easily sampled in a data augmented implementation. Following Scott (2011), natural proposal distributions are available to implement an MH step to sample $\boldsymbol{\gamma}_g|(\boldsymbol{\gamma}_{-g}, \mathbf{D}, \mathbf{z}_g)$ conditional on $\mathbf{z}_g = (z_{g1}, \ldots, z_{gn})$, for $g = 1, \ldots, G-1$.

To avoid any MH step, Frühwirth-Schnatter et al. (2012) apply auxiliary mixture sampling as introduced by Frühwirth-Schnatter and Frühwirth (2010) to approximate the logistic distribution of each ϵ_{gi} by a finite scale mixture of H normal distributions with zero means and parameters (s_h^2, w_h). The same authors obtain a 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 indicator r_{gi} is introduced as yet another latent variable. Conditional on the latent variables $\mathbf{z}_g$ and the indicators $\mathbf{r}_g = (r_{g1}, \ldots, r_{gn})$, the binary logit regression model reduces to a linear Gaussian regression model. Hence, the posterior $\boldsymbol{\gamma}_g|(\boldsymbol{\gamma}_{-g}, \mathbf{D}, \mathbf{z}_g, \mathbf{r}_g)$ is Gaussian and a Gibbs step is available to sample $\boldsymbol{\gamma}_g$, conditional on $\mathbf{z}_g$ and $\mathbf{r}_g$, for $g = 1, \ldots, G-1$. Finally, each component indicators r_{gi} is sampled from a discrete distribution conditional on z_{gi} and $\boldsymbol{\gamma}$.

1.2.3 An MCMC algorithm based on data augmentation

Based on the representation of Section 1.2.2, the following MCMC scheme is presented by Frühwirth-Schnatter et al. (2012):

1. for $g = 1, \ldots, G-1$, sample the $(J+1)$-dimensional regression coefficients $\boldsymbol{\gamma}_g$ conditional on $\mathbf{z}_g$ and $\mathbf{r}_g$. Using a normal prior $\boldsymbol{\gamma}_g \sim MVN(\mathbf{0}, v\mathbf{I}_{J+1})$, with $\mathbf{I}$ denoting the identity matrix and v set to a

high value (for example equal to 100), leads to a diffuse a prior. The conditional posterior is then given by a multivariate normal density with precision matrix $\mathbf{P}_{\gamma_g}$ and mean $\mathbf{m}_{\gamma_g}$:

$$\mathbf{P}_{\gamma_g} = \mathbf{X}'\mathbf{W}_g^{-1}\mathbf{X} + v^{-1}\mathbf{I}_{J+1}, \quad \mathbf{m}_{\gamma_g} = \mathbf{P}_{\gamma_g}^{-1}\mathbf{X}'\mathbf{W}_g^{-1}\left(\mathbf{z}_g + \log\sum_{l\neq g}\boldsymbol{\lambda}_l\right), \tag{1.38}$$

where $\mathbf{W}_g$ is a $n \times n$ diagonal matrix with nonzero elements equal to the randomly drawn variances $(\omega_{1g} = s^2_{r_{g1}}, \dots, \omega_{ng} = s^2_{r_{gn}})$ for the g-th group;

2. sample all (partial) differences of utilities $z_{1i}, \dots, z_{G-1,i}$ simultaneously for each i from

$$z_{gi} = \log\left(\frac{\lambda_{gi}}{\log\sum_{l\neq g}\lambda_{li}}U_{gi} + D_{gi}\right) - \log\left(1 - U_{gi} + \frac{\lambda_{gi}}{\log\sum_{l\neq g}\lambda_{li}}D_{gi}\right), \tag{1.39}$$

with $U_{gi} \sim Unif(0,1)$;

3. for $i = 1, \dots, n$ and $g = 1, \dots, G-1$, sample the component indicators r_{gi} conditional on z_{gi} from

$$\Pr(r_{gi} = h|z_{gi}, \boldsymbol{\gamma}_g) \propto \frac{w_h}{s_h}\exp\left[-\frac{1}{2}\left(\frac{z_{gi} - \mathbf{x}'\boldsymbol{\gamma}_g + \log\sum_{l\neq g}\lambda_{li}}{s_h}\right)^2\right]. \tag{1.40}$$

4. sample the component parameters $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G$ given the allocations $\mathbf{D}_i$, ..., $\mathbf{D}_n$;

5. classify each individual i according to Bayes' rule, by drawing $\mathbf{D}_i, \dots, \mathbf{D}_n$ from their full conditional

$$\Pr(D_{gi} = 1|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\gamma}, \boldsymbol{\theta}) \propto \frac{\lambda_{gi}}{\sum_{g=1}^G \lambda_{gi}} f(\mathbf{y}_i|\boldsymbol{\theta}_g). \tag{1.41}$$

1.3 Summary of the remaining chapters

In this Thesis, a more flexible specification of the covariate effects is considered for the whole suite of mixture of experts models. Starting from the gating network mixture of experts models, this aim is achieved by replacing linear predictors with additive ones in the multinomial logistic regression model (1.35) which defines the effects of the concomitant covariates on the component weights $\pi_1, \dots, \pi_{G-1}$. To achieve a parsimonious representation of the smooth functions included in the predictors, the Bayesian P-splines approach suggested by Lang and Brezger (2004) is used. Following Frühwirth-Schnatter et al. (2012), data augmentation is exploited to represent the flexible multinomial logistic regression model as a (partial) difference random utility model (dRUM), and their MCMC algorithm (reported in Section 1.2.3) is adapted to perform parameter estimation.

The remainder of the Thesis is organized as follows. After a brief introduction on Bayesian Generalized Additive Models with P-splines, the general specification for a flexible gating network mixture of experts model is presented in Chapter 2, together with the associated Bayesian inference procedure. In Chapter 3 and Chapter 4, this methodology is adapted to the specific cases of categorical and continuous manifest variables, respectively, and some results on simulation experiments and applications on real data are provided. In Chapter 5, the methodology is extended to include flexible covariate effects also on the parameters of the component densities. In the applications, this full mixtures of experts model is compared, among others, to a mixture of additive models with P-splines. Finally, in Chapter 6, some conclusions and remarks are discussed.

Chapter 2

Flexible modelling of the mixture weights

2.1 Generalized additive models

Generalized additive models (Hastie and Tibshirani, 1990) can be a useful modelling tool for data analysis. A generalized additive model is a generalized linear model with a linear predictor involving a sum of smooth functions $s(\cdot)$ of covariates:

$$h(\mathrm{E}(\mathbf{Y}|\mathbf{X})) = \eta = s_1(X_1) + s_2(X_2) + \dots . \tag{2.1}$$

Here $h(\cdot)$ is the link function which put into relationship the predictor η with the expected value of the response $\mathbf{Y}$, conditional on the covariates. Model 2.1 allows for rather flexible specification of the dependence between the response and the observed covariates. This flexibility comes at the cost of two theoretical issues: it is necessary both to specify an analytical form for the smooth functions and to choose how smooth they should be. Several proposals are available for modeling and estimating the smooth functions $s(\cdot)$; see, e.g., Hastie et al. (2009) and Wood (2017) for an overview.

Consider a set of independent and identically distributed observations $\{y_i\}$, $i = 1, \dots, n$. Each observation i has an associated vector of covariates $\mathbf{x}_i = (1,\, x_{i1},\, \dots,\, x_{iJ^*-1},\, x_{iJ^*},\, x_{iJ^*+1},\, \dots,\, x_{iJ})$, of which the last $J - J^*$ are metrical, with $J^* \in \{0, 1, \dots, J\}$. Given covariates and unknown parameters, the predictor predictor is defined as:

$$\eta_i = \gamma_0 + x_{i1}\gamma_1 + \dots + x_{iJ^*}\gamma_{J^*} + s_{J^*+1}(x_{i,J^*+1}) + \dots + s_J(x_{iJ}), \tag{2.2}$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_J^*)$ denotes the vector of coefficients corresponding to the parametric part of the predictor, and cubic splines are chosen to approximate the smooth functions $s_{J^*+1}(\cdot), \dots, s_J(\cdot)$.

The choice of cubic splines is motivated by some appealing theoretical properties. Consider a set of points $\{x_i, y_i\}, i = 1, \dots, n$, where $x_i < x_i + 1$ for each i. The cubic spline interpolating these points is a piecewise function made up of sections of cubic polynomial, one for each interval $[x_i, x_i + 1]$, which are joined together so that the whole function is continuous up to second derivative. The points at which the sections are joined (and the two end points) are known as the knots of the spline. In each section within two consecutive knots, the funcion has constant coefficients which in turn can vary from section to section. Cubic splines are the smoothest possible interpolant through any set of data, in the sense that among all the possible functions $s(\cdot)$ that

- are continuous in $[x_1, x_n]$,
- have absolutely continuous first derivative,
- interpolate $\{x_i, y_i\}$, $i = 1, \dots, n$,

the cubic spline is the one that presents the lowest integrated squared second derivative

$$\int_{x_1}^{x_n} s''(x)^2 \mathrm{d}x, \tag{2.3}$$

which can be interpreted as a measure of the wiggliness of the curve; see Green and Silverman (1993) for a proof.

This property is exploited by Reinsch (1967) to also prove that, among all the previously defined functions $s(\cdot)$, for a given λ, the one minimizing

$$\sum_{i=1}^{n} (y_i - s(x_i))^2 + \lambda \int_{x_1}^{x_n} s''(x)^2 \mathrm{d}x \tag{2.4}$$

is a cubic spline. In other words, smoothing splines naturally arise as the solution to the minimization problem defined in Equation (2.4), where λ is a tuning parameter used to control the weight associated to the penalty term (2.3) that account for the wiggliness of the curve.

Cubic splines can be conveniently represented as a linear combination of known basis functions and unknown coefficients, which means the model in Equation (2.2) is still linear in the parameters. A wide variety of basis functions is present in the literature, see Wood (2017, Chapter 5) for a review.

2.1.1 P-splines

The use of B-spline basis is appealing because these basis functions are strictly local – each basis function is non-zero only over the intervals between $r + 2$ adjacent knots, where r is the degree of the basis ($r = 2$ for the

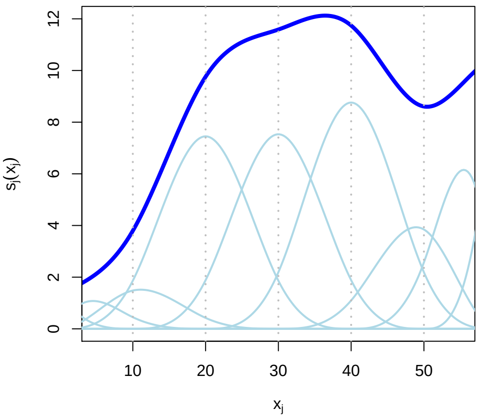


Figure 2.1: Illustration of a smooth curve by rank $m = 9$ B-spline bases.

cubic spline). To define an m parameter cubic B-spline basis, one needs to define $m + r$ equally spaced knots, $k_1 < k_2 < \cdots < k_{m+r-1} < k_{m+r}$, and the interval the spline is to be evaluated over lies within $[k_r, k_m]$ (so that the first and last r knot locations are essentially arbitrary). Such spline can then be written as a linear combination of m B-spline basis functions $B_{j\rho}(x_j)$ with coefficients $\beta_{j\rho}$, that is,

$$s(x_{ij}) = \sum_{\rho=1}^{m} B_{j\rho}(x_{ij})\beta_{j\rho}, \quad i = 1, \ldots, n, \tag{2.5}$$

where, following e.g. De Boor (1972), each B-spline basis function $B_{j\rho}(x_{ij})$ is most conveniently defined recursively for each unit i, as

$$\begin{aligned} B_{j\rho}(x_{ij}) &= B_{j\rho}^{(r)}(x_{ij}) = \tfrac{x_{ij}-k_\rho}{k_{\rho+r}-k_\rho} B_{j\rho}^{(r-1)}(x_{ij}) + \tfrac{k_{\rho+r+1}-x_{ij}}{k_{\rho+r+1}-k_{\rho+1}}, \\ B_{j\rho}^{(0)}(x_{ij}) &= \begin{cases} 1, & \text{if } k_\rho \le x_{ij} < k_{\rho+1}, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{2.6}$$

Figure 2.1 shows an example of a cubic B-spline with 5 knots, whose locations are indicated by the vertical dotted lines in grey. The light blue curves show B-spline basis functions multiplied by the associated coefficients: each is non-zero over 4 intervals. Their sum gives the spline itself, represented by the thicker blue curve above.

B-splines were developed as a stable basis for large scale spline interpolation, see De Boor et al. (1978) for further details. However, the real statistical interest in B-splines focuses on their penalized version, commonly known as P-splines, introduced by Eilers and Marx (1996). The authors suggest to work with a moderately large number of equally spaced knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on differences of adjacent B-spline coefficients to guarantee sufficient smoothness of the fitted curves. Estimation can be carried out by direct maximization of the penalized likelihood (Marx and Eilers, 1998) or via backfitting (Hastie and Tibshirani, 1990).

2.1.2 Bayesian approach

In a Bayesian approach, unknown parameters $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{j,m})$, $j = J^* + 1, \ldots, J$, as well as $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{J^*})$, can be considered as random variables, with appropriate prior distributions. Lang and Brezger (2004) define priors for the regression parameters $\boldsymbol{\beta}_j$ by replacing the difference penalties proposed by Eilers and Marx (1996) with their stochastic analogues. In particular, first differences correspond to a first-order random walk:

$$\beta_{j\rho} = \beta_{j,\rho-1} + u_{j\rho}, \quad u_{j\rho} \sim N(0, \tau_j^2). \tag{2.7}$$

The amount of smoothness is controlled by the variance parameters τ_j^2, which corresponds to the reciprocal of the smoothing parameters in the frequentist approach; this parameter protects against possibile overfitting if a large number of knots is chosen. The priors in Equation (2.7) can be equivalently written in the form of global smoothness priors:

$$\boldsymbol{\beta}_j | \tau_j^2 \propto \exp\left(-\frac{1}{2\tau_j^2} \boldsymbol{\beta}_j' \mathbf{K}_j \boldsymbol{\beta}_j \right), \quad \mathbf{K}_j = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \\ , & & & & \end{bmatrix}. \tag{2.8}$$

with penalty matrix $\mathbf{K}_j$ given by $\mathbf{K}_j = \boldsymbol{\Delta}_1' \boldsymbol{\Delta}_1$, where $\boldsymbol{\Delta}_1$ is the first order difference matrix. Penalty matrix $\mathbf{K}_j$ is rank deficient, as $\text{rank}(\mathbf{K}_j) = m - 1$, therefore, the prior in Equation (2.8) is improper. It is worth mentioning that these kind of models are usually referred to, in the literature, as intrinsic Gaussian Markov random fields (Rue and Held, 2005).

For full Bayesian inference, each unknown variance parameter τ_j^2 is also considered as random, and estimated together with the corresponding j-th vector of unknown coefficients $\boldsymbol{\beta}_j$, for $j = J^* + 1, \ldots, J$. Therefore, hyperpriors are assigned to variances $\tau^2_{J^*+1}, \ldots, \tau^2_J$ in a further layer of the hierarchy, choosing them to be dispersed inverse gamma priors $\tau^2_j \sim IG(a_j, b_j)$, $j = J^* + 1, \ldots, J$. Common choices for the hyperparameters are $a_j = b_j = 10^{-3}$, or $a_j = 1$ and $b_j = 5 \times 10^{-3}$ or lower, leading to almost diffuse priors for τ^2_j.

2.2 Semiparametric gating network mixture of experts models

2.2.1 Model specification

Consider a sample of independent and identically distributed observations $\{\mathbf{y}_i\}$, $i = 1, \ldots, n$, from an heterogeneous population, and suppose that the distribution of $\mathbf{y}$ is described by a G-components finite mixture model with weights $\pi_1, \ldots, \pi_G$, such that $0 < \pi_g < 1$ and $\sum_{g=1}^{G} \pi_g = 1.$. The distribution in each component $g = 1, \ldots, G$, is described by the probability (density) function $f(\mathbf{y}_i|\boldsymbol{\theta}_g)$ with parameters $\boldsymbol{\theta}_g$. Observation i has vector of associated covariates $\mathbf{x}_i = (1,\, x_{i1},\, \ldots,\, x_{iJ^*-1},\, x_{iJ^*},\, x_{iJ^*+1},\, \ldots,\, x_{iJ})$, of which the last $J - J^*$ are metrical, with $J^* \in \{0, 1, \ldots, J\}$.

The gating network mixtures of experts model introduced in Section 1.2.1 extends the finite mixture model by allowing the distribution of the latent variable to depend on the concomitant variables:

$$f(\mathbf{y}_i|\mathbf{x}_i) = \sum_{g=1}^{G} \pi_g(\mathbf{x}_i) f(\mathbf{y}_i|\boldsymbol{\theta}_g). \tag{2.9}$$

Rather than modelling the component weights by a multinomial logit regression model, the linear predictors in Equation (1.35) are extended by exploiting the additive structure introduced in Section 2.1. More precisely, the assumption of linearity of the covariate effects on the log-odds of the mixture weights is relaxed by defining an additive predictor for each component g as follows:

$$\log \frac{\pi_g(\mathbf{x}_i)}{\pi_G(\mathbf{x}_i)} = \eta_g(\mathbf{x}_i) = \eta_{gi} = \sum_{j=0}^{J^*} x_{ij}\gamma_{gj} + \sum_{j=J^*+1}^{J} s_{gj}(x_{ij}), \tag{2.10}$$

where $\boldsymbol{\gamma}_g = (\gamma_{g0}, \gamma_{g1} \ldots, \gamma_{gJ^*})$ denotes the vector of coefficients for the usual parametric part of the predictor, and $s_{g,J^*+1}(\cdot), \ldots, s_{gJ}(\cdot)$ are unknown smooth functions of the metrical covariates for the g-th class. As in Section 2.1.1, splines are used to express these smooth functions:

$$s_{gj}(x_{ij}) = \sum_{\rho=1}^{m} B_{j\rho}(x_{ij})\beta_{gj\rho}, \tag{2.11}$$

where $B_{j\rho}(\cdot)$ is a B-spline basis function for a cubic spline, for $\rho = 1, \dots, m$, and $\boldsymbol{\beta}_{gj} = (\beta_{gj1}, \dots, \beta_{gjm})$ denotes the associated vector of coefficients for the g-th component, $j = J^* + 1, \dots, J$. To ensure identifiability of the additive predictors, each function $s_{gj}(x_j)$ is constrained to have zero mean, that is,

$$\frac{1}{\text{range}(x_j)} \int s_{gj}(x_j)\,\mathrm{d}x_j = 0, \quad j = J^* + 1, \dots, J. \tag{2.12}$$

These constraints can be incorporated into estimation by centering functions $s_{gj}(x_j)$ about their means: in practice, the average of the smooth functions will be incorporate into the overall intercept γ_0. See Section 2.2.3 for further details about Bayesian constrained sampling.

By defining the $n \times m$ design matrices $\mathbf{B}_j$ – whose generic ρ-th element in i-th row is given by $B_{j\rho}(x_{ij})$ – the predictor (2.10) can be rewritten in matrix notation as

$$\boldsymbol{\eta}_g = \eta_g(\mathbf{x}) = \mathbf{X}\boldsymbol{\gamma}_g + \sum_{j=J^*+1}^{J} \mathbf{B}_j \boldsymbol{\beta}_{gj}, \tag{2.13}$$

where $\mathbf{X}$ is the design matrix of fixed effects and $\boldsymbol{\beta}_{gj} = (\beta_{gj1}, \dots, \beta_{gjm})$. Following some simple algebra, the component weights can be expressed, for each unit i, as

$$\pi_g(\mathbf{x}_i) = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\gamma}_g + \sum_{j=J^*+1}^{J} \mathbf{B}_{ij}'\boldsymbol{\beta}_{gj}\right)}{1 + \sum_{g=1}^{G-1} \exp\left(\mathbf{x}_i'\boldsymbol{\gamma}_g + \sum_{j=J^*+1}^{J} \mathbf{B}_{ij}'\boldsymbol{\beta}_{gj}\right)}, \quad g = 1, \dots, G, \tag{2.14}$$

where $\mathbf{B}_{ij}$ is a vector containing the elements of the i-th row of $\mathbf{B}_j$. Essentially, by resorting to spline functions, it is possible to extend the model in Equation (1.36) to its flexible version, preserving linearity of the predictor $\boldsymbol{\eta}_g$ with respect to the unknown parameters $\boldsymbol{\gamma}_g$ and $\boldsymbol{\beta}_g$.

2.2.2 Bayesian Inference

To represent in a convenient way the previously described multinomial logistic semiparametric regression model (2.10) in the Bayesian context, the data augmentation scheme in Equation (1.37) is followed. By doing so, the multinomial model reduces to a binary model involving the component indicator $\mathbf{D}_i = (D_{1i}, \ldots, D_{Gi})$, introduced in Section 1.1.3, and (partial) differences of random utilies z_{gi}:

$$\begin{aligned} z_{gi} &= \eta_{gi} - \log\left(\sum_{l \neq g} \lambda_{li}\right) + \epsilon_{gi}, \\ D_{gi} &= \mathbb{1}(z_{gi} > 0). \end{aligned} \tag{2.15}$$

With respect to the partial dRUM representation in Equation (1.37), here z_{gi} is specified in a more flexible way, with additive predictor η_{gi} defined involving a sum of smooth function – as in Equation (2.10) – and $\lambda_{gi} = \exp(\eta_{gi})$. Gaussian priors are assumed for the fixed effects parameters: $\boldsymbol{\gamma}_g \sim MVN(\mathbf{0}, v\mathbf{I}_{J^*+1})$, with variance hyperparameter v set so a sufficiently large value (e.g. equal to 100), in case a non-informative prior is needed. To penalize the m B-spline parameters related to the nonlinear part of the predictor, m is set high (e.g. equal to 23), and the priors for these coefficients are defined following Lang and Brezger (2004):

$$\beta_{gj\rho} = \beta_{gj,\rho-1} + u_{gj\rho}, \quad u_{gj\rho} \sim N(0, \tau^2_{gj}), \tag{2.16}$$

or, equivalently,

$$\boldsymbol{\beta}_{gj} | \tau^2_{gj} \propto \exp\left(-\frac{1}{2\tau^2_{gj}} \boldsymbol{\beta}'_{gj} \mathbf{K}_j \boldsymbol{\beta}_{gj}\right), \tag{2.17}$$

with penalty matrix $\mathbf{K}_j$ defined as in Equation (2.8), and $\tau^2_{gj} \sim IG(a_j = 1, b_j = 5 \times 10^{-3})$. The logistic distribution of the i.i.d. errors ϵ_{gi} is approximated by a finite mixture of normals, following Frühwirth-Schnatter and Frühwirth (2010), as

$$f(\epsilon_{gi}) \approx \sum_{h=1}^{H} w_h f_N(0, s^2_h), \tag{2.18}$$

with $f_N(\cdot)$ denoting the normal density function. The parameters of the finite mixture approximation in Equation (2.18), (w_h, s^2_h), for $h = 1, \ldots, H$ are obtained by minimizing the Kullback-Leibler divergence (Kullback and Leibler, 1951) and can be found in Table 2 of Frühwirth-Schnatter and Frühwirth (2010), along with an evaluation of goodness of these approximations. The number of components H in the approximating mixture model (2.18) is fixed equal to 6 whenever possibile, although good approximations can be obtained for as small as $H = 3$. In a second step of data augmentation, the component indicator $r_{gi} \sim MulNom_H(1; w_1, \ldots, w_H)$ is introduced as yet another latent variable. Regarding the parameters of the component densities, conjugacy between priors and the likelihood function can be exploited in order to sample from the posterior distribution using Gibbs steps, depending on the specific nature of the component densities $f(\mathbf{y}|\boldsymbol{\theta}_g)$.

2.2.3 The MCMC Algorithm

Based on the representation in Section 2.2.2, a new MCMC algorithm is implemented for fixed G, by integrating the scheme proposed by Frühwirth-Schnatter et al. (2012) – detailed in Section 1.2.3 – with the Bayesian P-spline approach proposed by Lang and Brezger (2004) (Section 2.1.1). The steps to be followed are:

1. sample the regression coefficients $\boldsymbol{\beta}_g$ conditional on $\mathbf{z}_g$ and $\mathbf{r}_g$, $g = 1, \ldots, G-1$. Using the prior in Equation (2.17), the full conditional of $\boldsymbol{\beta}_{gj}$ is given by a multivariate normal density. Straightforward calculations (Brezger and Lang, 2006) show that the precision matrix $\mathbf{P}_{gj}$ and the mean $\mathbf{m}_{gj}$ of $\boldsymbol{\beta}_{gj}|\cdot$ are given by

$$\begin{aligned} \mathbf{P}_{gj} &= \mathbf{B}_j' \mathbf{W}_g^{-1} \mathbf{B}_j + \frac{1}{\tau_{gj}^2} \mathbf{K}_j, \\ \mathbf{m}_{gj} &= \mathbf{P}_{gj}^{-1} \mathbf{B}_j' \mathbf{W}_g^{-1} \left(\mathbf{z}_g - \tilde{\boldsymbol{\eta}}_{g,-j} + \log \sum_{l \neq g} \boldsymbol{\lambda}_l \right), \end{aligned} \tag{2.19}$$

   where $\tilde{\boldsymbol{\eta}}_{g,-j}$ is the part of the predictor associated with all, but the j-th, effects in the model, and $\mathbf{W}_g$ 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;

2. center each smooth function $s_{gj}(x_j)$. Imposing the constraint in Equation (2.12) is equivalent to sampling $\boldsymbol{\beta}_{gj}|(\mathbf{1}_n'\mathbf{B}_j\boldsymbol{\beta}_{gj} = \mathbf{0})$, for each $j = J^* + 1, \ldots, J$, where $\mathbf{1}$ denotes a vector of length n with all elements equal to 1. Following Algorithm 2.6 in Rue and Held (2005), this can be done by trasforming each vector of coefficients $\boldsymbol{\beta}_{gj}$ as follows:

$$\tilde{\boldsymbol{\beta}}_{gj} = \boldsymbol{\beta}_{gj} - \mathbf{P}_{gj}^{-1}\mathbf{B}_j'\mathbf{1}_n \left(\mathbf{1}_n'\mathbf{B}_j\mathbf{P}_{gj}^{-1}\mathbf{B}_j'\mathbf{1}_n\right)^{-1} \mathbf{1}_n'\mathbf{B}_j\boldsymbol{\beta}_{gj}; \tag{2.20}$$

3. sample the fixed effects parameters $\boldsymbol{\gamma}_g$ from a multivariate normal den-

sity with precision matrix $\mathbf{P}_{\gamma_g}$ and mean vector $\mathbf{m}_{\gamma_g}$:

$$\begin{aligned}
\mathbf{P}_{\gamma_g} &= \mathbf{X}'\mathbf{W}_g^{-1}\mathbf{X} + \frac{1}{v}\mathbf{I}_{J^*+1}, \\
\mathbf{m}_{\gamma_g} &= \mathbf{P}_{\gamma_g}^{-1}\mathbf{X}'\mathbf{W}_g^{-1}\left(\mathbf{z}_g - \tilde{\boldsymbol{\eta}}_{\boldsymbol{g},-\gamma} + \log\sum_{l\neq g}\boldsymbol{\lambda}_l\right).
\end{aligned} \tag{2.21}$$

Here $\tilde{\boldsymbol{\eta}}_{g,-\gamma}$ represents the nonlinear part of the predictor for the g-th component;

4. sample the variance parameters τ^2_{gj} conditional on $\tilde{\boldsymbol{\beta}}_{gj}$:

$$\tau^2_{gj}|\tilde{\boldsymbol{\beta}}_{gj} \sim IG\left(a_{gj} + \frac{\mathrm{rank}(\mathbf{K}_j)}{2}, b_{gj} + \frac{1}{2}\tilde{\boldsymbol{\beta}}'_{gj}\mathbf{K}_j\tilde{\boldsymbol{\beta}}_{gj}\right); \tag{2.22}$$

5. for each unit $i = 1, \dots, n$, sample all (partial) differences of utilities $z_{1i}, \dots, z_{G-1,i}$ simultaneously from:

$$z_{gi} = \log\left(\frac{\lambda_{gi}}{\log\sum_{l\neq g}\lambda_{li}}U_{gi} + D_{gi}\right) - \log\left(1 - U_{gi} + \frac{\lambda_{gi}}{\log\sum_{l\neq g}\lambda_{li}}D_{gi}\right), \tag{2.23}$$

with $U_{gi} \sim Unif(0, 1)$;

6. sample the component indicators r_{gi} conditional on z_{gi} from:

$$\Pr(r_{gi} = h|z_{gi}, \tilde{\boldsymbol{\beta}}_g, \boldsymbol{\gamma}_g) \propto \frac{w_h}{s_h}\exp\left[-\frac{1}{2}\left(\frac{z_{gi} - \eta_{gi} + \log\sum_{l\neq g}\lambda_{li}}{s_h}\right)^2\right]; \tag{2.24}$$

7. sample the component parameters $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G$ given the component indicators $\mathbf{D}_i, \dots, \mathbf{D}_n$. This step depends on the specific features of the outcome variables and the choice for the component density functions $f(\mathbf{y}|\boldsymbol{\theta}_g)$, $g = 1, \dots, G$. See the next chapters for further details;

8. classify each unit i according to Bayes' rule: draw $\mathbf{D}_i, \dots, \mathbf{D}_n$ from the following discrete probability distribution which combines the likelihood and the prior:

$$\Pr(D_{gi} = 1|\mathbf{y}_i, \mathbf{x}_i, \tilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \propto \frac{\lambda_{gi}}{\sum_{g=1}^{G}\lambda_{gi}}f(\mathbf{y}_i|\boldsymbol{\theta}_g). \tag{2.25}$$

This MCMC algorithm and all the variants presented in this Thesis have been implemented in R code (R Core Team, 2020) and are available on GitHub at the following link: github.com/MarcoBerrettini/sMoE. All the computations in this Thesis are carried out using R.

2.2.4 A note on the computational cost of the algorithm

Obviously, the improved flexibility has a cost in terms of computational complexity. Assume that all the J concomitant covariates are metrical (i.e. $J^* = 0$) and that one is interested in estimating all of their corresponding effects through the algorithm proposed in Section 2.2.3. In this case, inverting the posterior precision matrix $\mathbf{P}_{gj}$ in Equation (2.19) is the operation which mostly inflates the computational cost of the algorithm. Indeed, the complexity of computing the inverse of an $m \times m$ matrix through the basic Gauss-Jordan elimination method (Althoen and Mclaughlin, 1987) is $\mathcal{O}(m^3)$, although it can be reduced via more efficient procedures, e.g. the Coppersmith-Winograd algorithm (Coppersmith and Winograd, 1987). Moreover, for any fixed g, this operation is repeated J times, once for each covariate. Note, in fact, that, as each row of any B-spline basis matrix $\mathbf{B}_j$, $j = 1, \ldots, J$, sums to 1, and given that the penalty matrix $\mathbf{K}_j$ does not have full rank, it is necessary to separately draw $\boldsymbol{\beta}_{g1}, \ldots, \boldsymbol{\beta}_{gJ}$, since a posterior precision matrix for the whole vector of coefficients $\boldsymbol{\beta}_g$ would not be invertible and would require to resort to alternative solutions, such as the Moore-Penrose inverse (Moore, 1920). When $J > 1$, this implies the presence of the additional centering step in Equation (2.20), for each $\boldsymbol{\beta}_{gj}$, $j = 1, \ldots, J$, together with a further step for estimating the intercept. As anticipated in Section 2.2.1, this procedure guarantees identifiability for the additive predictors.

For comparison purposes, assume now that the same J covariate effects are estimated via the parametric approach by Frühwirth-Schnatter et al. (2012), thus setting $J^* = J$ and reducing the MCMC algorithm proposed in Section 2.2.3 to the one reported in Section 1.2.3. Here, for any given g, the posterior variance matrix of all the regression coefficients, that are included in vector $\boldsymbol{\gamma}_g = (\gamma_0, \gamma_1, \ldots, \gamma_J)$, is matrix $P_{\boldsymbol{\gamma}_g}$ in Equation (1.38). Its inverse can be easily computed all at once, with complexity $\mathcal{O}((J + 1)^3)$, if the standard Gauss-Jordan algorithm is exploited. Notice that J is usually (way) lower than m, causing the parametric approach to be sensibly faster with respect to the semiparametric one. This difference, in terms of time required by each algorithm to complete the pre-specified number of iterations, is reported together with the results of some simulation studies in the next chapters.

2.2.5 Posterior inference and model selection

Once the MCMC algorithm has completed the prefixed number T of iterations, posterior inference is carried out by estimating each parameter's posterior mean over the last $T - T_0$ draws of the chains, with T_0 defining the burn-in phase. Posterior quantities can be computed for the smooth functions by considering them as linear combinations of spline bases and the corresponding regression coefficients' estimates. The uncertainty associated to the smooth functions is quantified via their pointwise percentiles (usually 2.5 - 97.5 or 5 - 95), for each function over the last $T - T_0$ posterior draws.

Observations can be allocated into the G components using the maximum-a-posteriori (MAP) rule. In particular, each unit $i = 1, \dots, n$ is assigned to the component $\hat{c}_i$ such that

$$\hat{c}_i = \arg\max_g \left(\sum_{t=T_0}^{T} D_{1i}^{(t)}, \dots, \sum_{t=T_0}^{T} D_{Gi}^{(t)} \right). \tag{2.26}$$

where $\mathbf{D}_i^{(t)} = (D_{1i}^{(t)}, \dots, D_{Gi}^{(t)})$ represents the allocation vector for unit i at iteration t. Sometimes, using the MAP rule, one or more components could have no units assigned to them: thus, it might be worth distinguishing between the number of components G and the number of non-empty components, denoted as

$$\tilde{G} = \sum_{g=1}^{G} \mathbb{1} \left(\sum_{i=1}^{n} \mathbb{1}(\hat{c}_i = g) > 0 \right). \tag{2.27}$$

Choosing the number of components in a mixture model is an important problem, which originated many efforts in the statistical literature. In most approaches, selecting G is related to the number of free parameters, which is not clear for the proposed model, due to the presence of regulariziation induced by the prior distribution on the regression coefficients. As reported in Section 1.1.8, one simple solution in the Bayesian framework is given by the AICM (Raftery et al., 2007), whose formula depends only on the log-likelihoods from the posterior simulation:

$$\text{AICM} = 2(\bar{l} - s_l^2), \tag{2.28}$$

where $\bar{l}$ and s_l^2 are the sample mean and variance of the sequence of log-likelihoods $f(\mathbf{y}_i|\boldsymbol{\theta}_{\mathbf{D}_i}^{(t)})$, for each iteration $t = T_0, \dots, T$, after the burn-in. AICM has already been applied successfully in the mixture modelling context, for example by Erosheva et al. (2007), Gormley and Murphy (2010b), Gormley and Murphy (2011), and Mollica and Tardella (2017).

2.2.6 Label switching

As for any finite mixture model, label switching may occur during MCMC sampling; see Frühwirth-Schnatter (2006, Section 3.5) for a review. To identify a mixture of experts model, Frühwirth-Schnatter et al. (2012) suggest to focus on a subset of a group-specific parameter and apply k-means clustering (with G clusters) to the posterior draws. MCMC draws belonging to the same group are assigned to the same cluster by k-means clustering, and the resulting classification sequences $\zeta_t = (S_1^{(t)}, \dots, S_G^{(t)})$ – where each $S_g^{(t)}$, g=1,…,G, is a classification index taking values in {1,…,G} – show how to re-arrange the group-specific parameters for each iteration $t = 1, \dots, T$, even if label switching occurred during sampling. In particular, if the mixture is not overfitting the number G of groups, ζ_t is a permutation of $\{1, \dots, G\}$, and a unique labeling is achieved by reordering the draws in the following way:

- relabel the hidden allocations $\mathbf{D}_1, \dots, \mathbf{D}_n$ through the inverse ζ_t^{-1}: re-arrange $\mathbf{D}_1, \dots, \mathbf{D}_G$ by $\mathbf{D}_{\zeta_t^{-1}(1)}, \dots, \mathbf{D}_{\zeta_t^{-1}(G)}$, respectively;
- relabel the group-specific parameters through $\zeta_t^{-1}(1)$, …, $\zeta_t^{-1}(G)$: re-arrange $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G$ by $\boldsymbol{\theta}_{\zeta_t^{-1}(1)}, \dots, \boldsymbol{\theta}_{\zeta_t^{-1}(G)}$;
- relabel the regression coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$), corresponding to the linear and nonlinear part of the predictor in the multinomial logistic regression model: substitute $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_G$ by $\boldsymbol{\gamma}_{\zeta_t^{-1}(1)} - \boldsymbol{\gamma}_{\zeta_t^{-1}(G)}, \boldsymbol{\gamma}_{\zeta_t^{-1}(2)} - \boldsymbol{\gamma}_{\zeta_t^{-1}(G)}, \dots, \boldsymbol{\gamma}_{\zeta_t^{-1}(G)} - \boldsymbol{\gamma}_{\zeta_t^{-1}(G)} = \mathbf{0}$ and, thanks to the fact that the additive predictors are still linear in the parameters, substitute $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_G$ by $\boldsymbol{\beta}_{\zeta_t^{-1}(1)} - \boldsymbol{\beta}_{\zeta_t^{-1}(G)}, \boldsymbol{\beta}_{\zeta_t^{-1}(2)} - \boldsymbol{\beta}_{\zeta_t^{-1}(G)}, \dots, \boldsymbol{\beta}_{\zeta_t^{-1}(G)} - \boldsymbol{\beta}_{\zeta_t^{-1}(G)} = \mathbf{0}$. Subtracting $\boldsymbol{\gamma}_{\zeta_t^{-1}(G)}$ and $\boldsymbol{\beta}_{\zeta_t^{-1}(G)}$, respectively, from all draws ensures that the regression coefficients of the baseline are all equal to 0 in the identified model.

Chapter 3

Categorical manifest variables

3.1 Model specification and Bayesian inference

For the discrete case, each manifest variables $\mathbf{Y}_q$, $q = 1, \dots, Q$, has P_q possible outcomes, and takes value $y_{iqp} = 1$ if the i-th unit presents the p-th category for the q-th variable, $y_{iqp} = 0$ otherwise. For each group $g = 1, \dots, G$, and for each manifest variable a multinomial distribution is assumed, with probabilities $\boldsymbol{\xi}_{gq} = (\xi_{gq1}, \dots, \xi_{gqP_q})$. Furthermore, conditioning on $\mathbf{x}_i$ and assuming conditional independence between the Q responses, $\mathbf{y}_i$ has the following distribution

$$f(\mathbf{y}_i|\mathbf{x}_i) = \sum_{g=1}^{G} \pi(\mathbf{x}_i) \prod_{q=1}^{Q} \prod_{p=1}^{P_q} (\xi_{gqp})^{y_{iqp}}. \tag{3.1}$$

Model in Equation (3.1) may be referred to as latent class models with concomitant covariates (Dayton and Macready, 1988) or latent class regression models (Linzer et al., 2011).

Component weights $\pi_1(\mathbf{x}_i), \dots, \pi_{G-1}(\mathbf{x}_i)$ are defined as in Equation (2.14) and a Dirichlet distribution is assigned to each vector of conditional probabilities $\boldsymbol{\xi}_{gq}$, with hyperparameters $\iota_1, \dots, \iota_{P_q}$ all set equal to 1 for a flat prior. Thus, conditional on the latent component indicator $\mathbf{D}_g$, $g = 1, \dots, G$, the conditional probabilities can be sampled during step 7 of the MCMC algorithm introduced in Section 2.2.3 from the following full conditional:

$$\boldsymbol{\xi}_{gq}|\mathbf{D}, \mathbf{y} \sim Dir\left(\iota_1 + \sum_{i=1}^{n}(D_{gi} \cdot y_{iq1}), \dots, \iota_{P_q} + \sum_{i=1}^{n}(D_{gi} \cdot y_{iqP_q})\right). \tag{3.2}$$

Finally, the conditional posterior of the indicators can be made explicit for $g = 1, \dots, G$ as:

$$\Pr(D_{gi} = 1 | \mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}_g, \boldsymbol{\gamma}_g, \boldsymbol{\xi}_g) \propto \frac{\lambda_{gi}}{\sum_{g=1}^{G} \lambda_{gi}} \prod_{q=1}^{Q} \prod_{p=1}^{P_q} (\xi_{gqp})^{y_{iqp}} \,. \tag{3.3}$$

3.2 Simulation study

The performance of the proposed semiparametric approach based on Bayesian P-splines is investigated in a simulated environment. Two experiments are carried out, differing for the true number of components and the distribution of the manifest variables. The quality of the estimates for the covariates' effects is evaluated through a comparison between the true effects and the estimated posterior marginal effects, obtained using both the proposed semiparametric method and restricting the additive predictor to be a linear function of the covariates. To assess the performance of the estimators $\hat{s}_{gj}(\cdot)$ of the unknown regression functions $s_{gj}(\cdot)$, the square root of the average squared errors (hereafter, RASE) is also considered, defined as:

$$\text{RASE}_{s_{gj}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\hat{s}_{gj}(x_{ji}^{*}) - s_{gj}(x_{ji}^{*}) \right)^2}, \quad g = 1, \dots, G-1; \, j = 1, \dots, J, \tag{3.4}$$

where $\{x_{ji}^{*}\}$, $i = 1, \dots, n$, are grid points taken evenly in the range of each covariate.

Regarding the clustering performance, a comparison between the resulting and the true allocations is made in terms of both Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) and soft ARI (Flynt et al., 2019). While ARI measures similarity between two data clusterings (hard partitions), soft ARI (sARI) can incorporate the posterior allocation distributions (soft partitions). Both indexes take values ranging from a maximum of +1, if the two partitions are identical (apart from labelling), to a negative minimum if the similarity between the two partitions is lower than the expected similarity of all pairwise comparisons between clusterings specified by a random model. For this second part of the analysis, Bayesian latent class (BLC) models, which ignore the effects of concomitant information on the components' weights, are also considered. For each method and each value of G, 4000 MCMC draws are simulated after a burn-in of 1000 draws. The optimal choice for G is based on the AICM.

Table 3.1: Cluster-specific conditional probabilities.

ξ_{gqp}	g=1	g=2
$\boldsymbol{\xi}_{g1}$	(0.7, 0.1, 0.2)	(0.2, 0.7, 0.1)
$\boldsymbol{\xi}_{g2}$	(0.2, 0.8)	(0.7, 0.3)
$\boldsymbol{\xi}_{g3}$	(0.3, 0.6, 0.1)	(0.1, 0.3, 0.6)
$\boldsymbol{\xi}_{g4}$	(0.1, 0.1, 0.5, 0.3)	(0.5, 0.3, 0.1, 0.1)
$\boldsymbol{\xi}_{g5}$	(0.1, 0.1, 0.8)	(0.1, 0.8, 0.1)

Table 3.2: Mean (and standard deviation) of the RASE scores, computed on the estimated log-odds of the mixture weights over 100 simulated datasets.

RASE$_{s_{gj}}$	$s_{11}(x^*)$	$s_{12}(x^*)$
Semiparametric	0.255 (0.057)	0.314 (0.088)
Parametric	0.860 (0.007)	1.346 (0.002)

3.2.1 First simulation experiment: $G = 2$

A batch of 100 independent datasets is generated, with $n = 1000$, from a 2-component mixture distribution for $Q = 5$ categorical manifest variables, whose component-specific distributions are reported in Table 3.1. The log-odds of component weights is assumed to depend on 2 uniformly distributed covariates x_1 and x_2 in $[0, 1]$, as

$$\eta_1(x_1, x_2) = 2(\sin(3\pi x_1)e^{(-x_1)} + (3x_2 - 1.5)^2) - 0.5. \tag{3.5}$$

Figures 3.1 and 3.2 show the marginal effects, $s_{11}(x_1)$ and $s_{12}(x_2)$, along with the average posterior estimated effects across the 100 simulations obtained using both the semiparametric and the parametric approach, respectively. It appears that the underlying trend can be recovered through the proposed semiparametric method. Conversely, the parametric competitor clearly cannot properly approximate the nonlinear non-monotonic trend. These results are summarized in terms of RASE in Table 3.2. Goodness of fit (versus complexity) and quality of the allocations, for fixed $G = 2$, are analyzed in Table 3.3. It can be noticed the semiparametric approach outperforms its competitors in tems of AICM, ARI and sARI.

For each dataset, the three algorithms are run setting the number of components equal to 2, 3 and 4. Table 3.4 shows the number of non-empty components for the optimal models selected according to AICM for each

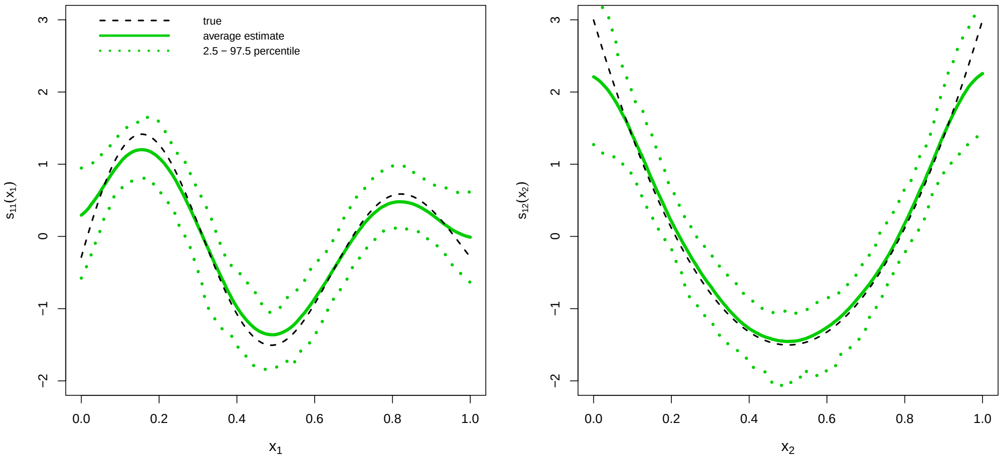


Figure 3.1: Average posterior effects (green solid line) and 2.5 – 97.5 pointwise percentiles (green dotted lines) of the concomitant covariates on the log-odds of mixture weights, estimated with the semiparametric approach over 100 simulated datasets. The true effect is represented by a black dashed line.

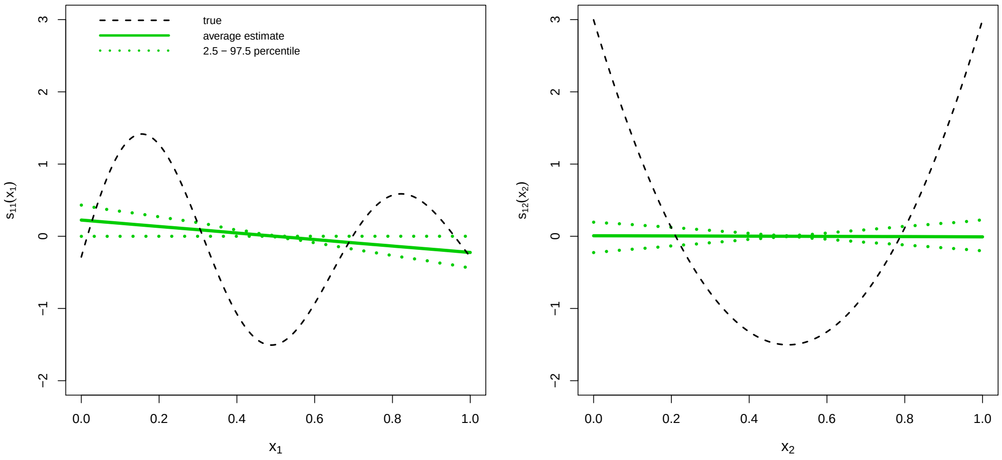


Figure 3.2: Average posterior effects (green solid line) and 2.5 – 97.5 pointwise percentiles (green dotted lines) of the concomitant covariates on the log-odds of mixture weights, estimated with the parametric approach over the 100 simulated datasets. The true effect is represented by a black dashed line.

Table 3.3: Average AICM, ARI and sARI (number of times each model ranks first) values for 100 simulated datasets, for fixed $G = 2$.

	AICM	(best)	ARI	(best)	sARI	(best)
Semiparametric MoE	8466.5	(96)	0.834	(99)	0.760	(100)
Parametric MoE	8533.7	(4)	0.794	(0)	0.704	(0)
BLC model	9063.4	(0)	0.794	(0)	0.704	(0)

Table 3.4: Optimal number of non-empty components selected for each method, according to AICM.

Method	$\tilde{G}$=2	$\tilde{G}$=3	$\tilde{G}$=4
Semiparametric MoE	100	-	-
Parametric MoE	26	69	5
BLC model	99	1	-

method for all the simulated datasets. It is evident that, in this scenario, restricting the additive predictor to be linear leads to a wrong choice of the number of non-empty components $\tilde{G}$ most of the times. If the statistics in Table 3.3 are compared with the ones computed with reference to the best models selected, according to Table 3.4, the results do not change much for the BLC model. There is a negligible difference for the semiparametric MoE models due to the fact that, in one sample, the best model selected according to AICM formally presents 3 components, although 1 of them is empty. Obviously, for the parametric MoE models there is an improvement in terms of average AICM (8510.0), while both the quantities that measures the accuracy of the allocations get sensibly worse (average ARI = 0.516, average sARI = 0.460).

3.2.2 Second simulation experiment: $G = 6$

A batch of 100 independent datasets is generated, with $n = 1000$, from a 6-component mixture distribution for $Q = 12$ categorical manifest variables, whose component-specific distributions are in Table 3.5. Similarly to the simulation study in Section 3.2.1, the component weights are assumed to depend on 2 uniformly distributed covariates x_1 and x_2 in $[0, 1]$. In particular, the log-odds of the weights are defined as follows:

$$
\begin{aligned}
\eta_1(x_1,x_2) &= 0.7(\sin(3\pi x_1)e^{-x_1} + (3x_2-1.5)^2) - 0.5;\\
\eta_2(x_1,x_2) &= 0.5e^{-x_1^2} - 0.8;\\
\eta_3(x_1,x_2) &= 0.5\sin(6x_1-1) + e^{-16(3x_1-0.5)^2} + \frac{e^{-30(x_2-0.3)}}{1+e^{-30(x_2-0.3)}};\\
\eta_4(x_1,x_2) &= 0.6\Bigg(3.4827\tilde{x}_1 - 4.7422\tilde{x}_1^2 + 3.3035\tilde{x}_1^3 - 1.2605\tilde{x}_1^4 + 0.251\tilde{x}_1^5\\
&\quad - 0.0204\tilde{x}_1^6 + \frac{e^{-20(x_2-0.4)}}{1+e^{-20(x_2-0.4)}}\Bigg), \quad \text{with } \tilde{x}_1 = 2.5x_1 + 0.5;\\
\eta_5(x_1,x_2) &= 0.5\left(\frac{e^{-10x_1}}{1+e^{-10x_1}} + \frac{e^{-50(x_2-0.3)}}{1+e^{-50(x_2-0.3)}}\right).
\end{aligned}
\tag{3.6}
$$

Despite the different setting, most of the conclusions that can be drawn for this simulation study are in line with ones in Section 3.2.1. Figures 3.3 to 3.7 show the marginal effects $s_{g1}(x_1)$ (first row) and $s_{g2}(x_2)$ (second row), for $g = 1,\ldots,G-1$, along with the average posterior estimated effects across the 100 simulations obtained on the simulated dataset using both the semiparametric and the parametric approach. The latter provides decent approximations, since most of the underlying trends are (almost) linear or monotonic. Moreover, oversmoothing can be noticed in the estimates obtained by the semiparametric MoE model, although on average results seem to be better than those obtained with the parametric MoE model, and the bands fully contain most of the effects. RASE scores reported in Table 3.6 confirm these remarks. Table 3.7 shows that the advantages in terms of goodness of fit do not correspond to such differences in terms of quality of the allocations, when the true number of components is fixed $G = 6$. Nevertheless, the semiparametric MoE appears to prevail most of times, especially in terms of sARI. For each dataset, the three algorithms are run setting the number of components G ranging from 2 to 8. All the competing methods detect the right number of non-empty groups, apart from 7 times: 5 for the parametric MoE model, twice for the BLC model. Thus, if the previous comparison based on clustering performance is repeated for each competing model, the conclusions do not change sensibly. For this second simulation study, a comparison is made also in terms of execution time of the algorithms estimating the three competing models.

Figure 3.8 shows through a box plot how the different computational complexity between the three algorithms affect the time employed by each

Table 3.5: Cluster-specific conditional probabilities.

ξ_{gqp}	$\boldsymbol{\xi}_{g1} = \boldsymbol{\xi}_{g2} = \boldsymbol{\xi}_{g3}$	$\boldsymbol{\xi}_{g4} = \boldsymbol{\xi}_{g5} = \boldsymbol{\xi}_{g6}$	$\boldsymbol{\xi}_{g7} = \boldsymbol{\xi}_{g8} = \boldsymbol{\xi}_{g9}$	$\boldsymbol{\xi}_{g10} = \boldsymbol{\xi}_{g11} = \boldsymbol{\xi}_{g12}$
g=1	(0.7, 0.1, 0.2)	(0.7, 0.1, 0.2)	(0.7, 0.1, 0.2)	(0.7, 0.1, 0.2)
g=2	(0.7, 0.2, 0.1)	(0.7, 0.2, 0.1)	(0.2, 0.1, 0.7)	(0.2, 0.1, 0.7)
g=3	(0.1, 0.2, 0.7)	(0.1, 0.2, 0.7)	(0.2, 0.7, 0.1)	(0.2, 0.7, 0.1)
g=4	(0.7, 0.1, 0.2)	(0.2, 0.1, 0.7)	(0.1, 0.2, 0.7)	(0.1, 0.7, 0.2)
g=5	(0.1, 0.7, 0.2)	(0.1, 0.7, 0.2)	(0.1, 0.7, 0.2)	(0.1, 0.7, 0.2)
g=6	(0.1, 0.7, 0.2)	(0.1, 0.2, 0.7)	(0.2, 0.1, 0.7)	(0.7, 0.1, 0.2)

Table 3.6: Mean (and standard deviation) of the RASE scores, computed on the estimated log-odds of the mixture weights over 100 simulated datasets.

RASE$_{s_{gj}}$	Semiparametric	Parametric
$s_{11}(x^*)$	0.251 (0.070)	0.378 (0.077)
$s_{12}(x^*)$	0.242 (0.089)	0.505 (0.035)
$s_{21}(x^*)$	0.159 (0.077)	0.208 (0.139)
$s_{22}(x^*)$	0.122 (0.058)	0.106 (0.079)
$s_{31}(x^*)$	0.326 (0.094)	0.426 (0.121)
$s_{32}(x^*)$	0.210 (0.095)	0.289 (0.201)
$s_{41}(x^*)$	0.143 (0.064)	0.213 (0.132)
$s_{42}(x^*)$	0.164 (0.074)	0.163 (0.077)
$s_{51}(x^*)$	0.139 (0.062)	0.212 (0.122)
$s_{52}(x^*)$	0.167 (0.065)	0.187 (0.064)

Table 3.7: Average AICM, ARI and sARI (number of times each model ranks first) over 100 simulated datasets, for fixed $G = 6$.

	AICM	(best)	ARI	(best)	sARI	(best)
Semiparametric MoE	20229.3	(77)	0.777	(62)	0.687	(93)
Parametric MoE	21128.1	(23)	0.771	(24)	0.681	(7)
BLCA	22360.3	(0)	0.771	(14)	0.679	(0)

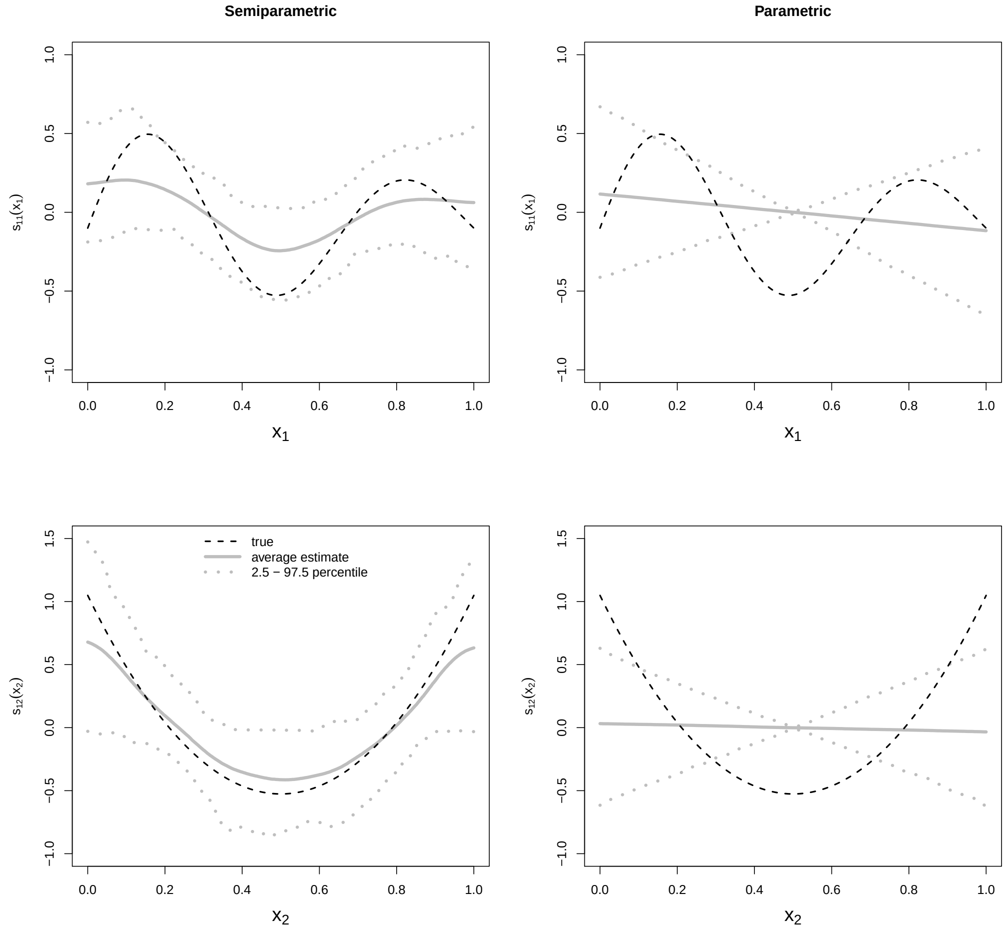


Figure 3.3: Average posterior effects (grey solid line) and 2.5 – 97.5 pointwise percentiles (grey dotted lines) of the concomitant covariates on the log-odds of mixture weights for group 1, estimated with the semiparametric approach over 100 simulated datasets. The true effect is represented by a black dashed line.

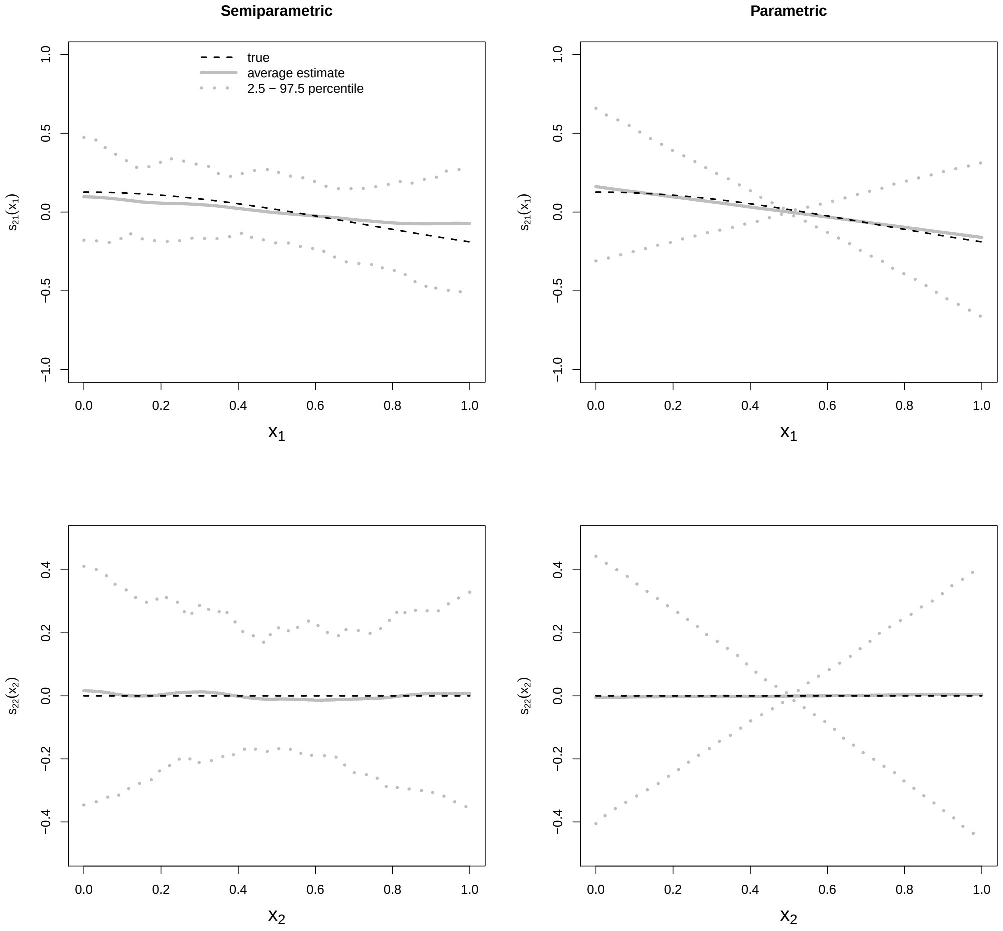


Figure 3.4: Average posterior effects (grey solid line) and 2.5 – 97.5 pointwise percentiles (grey dotted lines) of the concomitant covariates on the log-odds of mixture weights for group 2, estimated with the semiparametric approach over 100 simulated datasets. The true effect is represented by a black dashed line.

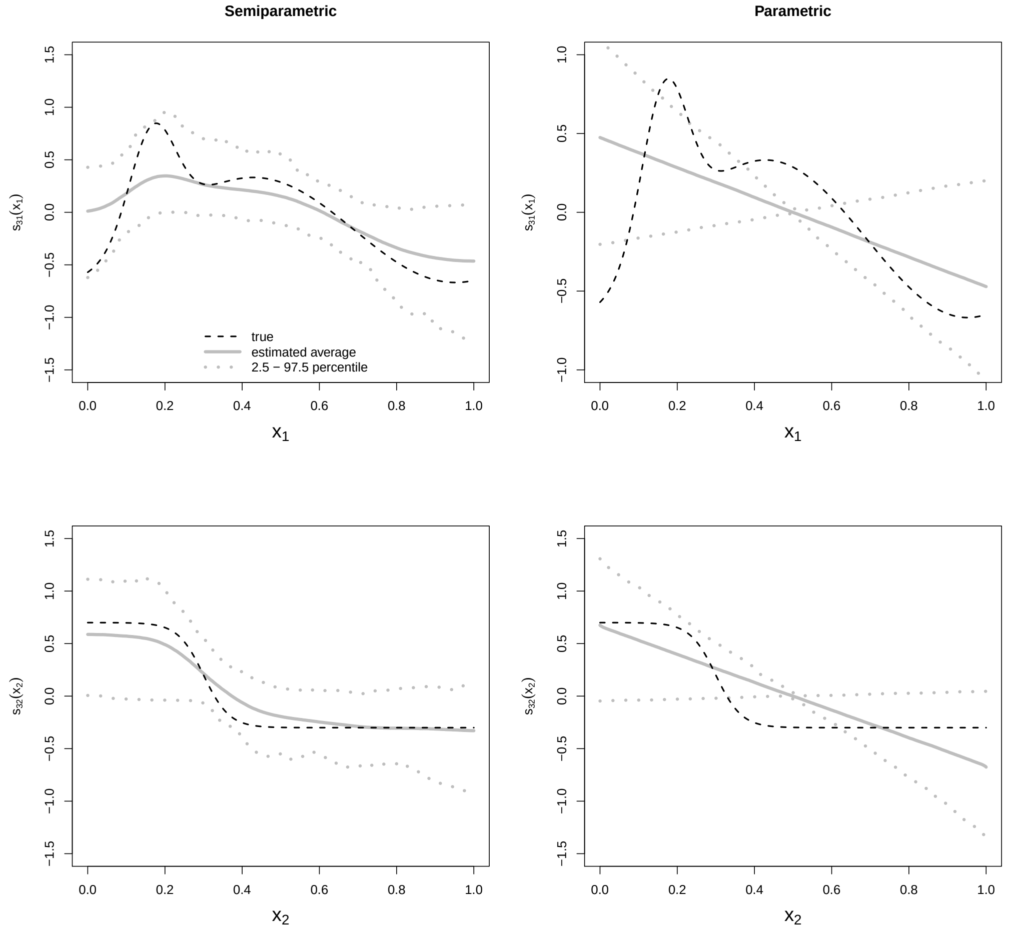


Figure 3.5: Average posterior effects (grey solid line) and 2.5 – 97.5 pointwise percentiles (grey dotted lines) of the concomitant covariates on the log-odds of mixture weights for group 3, estimated with the semiparametric approach over 100 simulated datasets. The true effect is represented by a black dashed line.

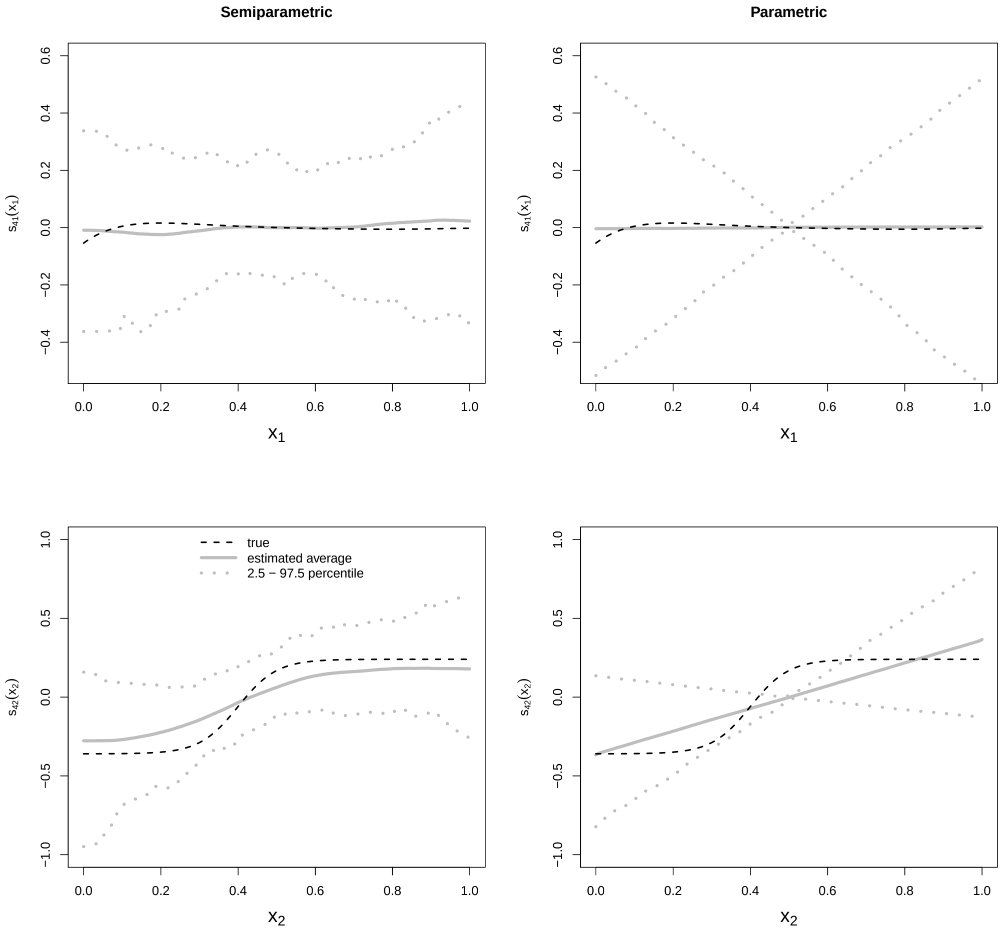


Figure 3.6: Average posterior effects (grey solid line) and 2.5 – 97.5 pointwise percentiles (grey dotted lines) of the concomitant covariates on the log-odds of mixture weights for group 4, estimated with the semiparametric approach over 100 simulated datasets. The true effect is represented by a black dashed line.

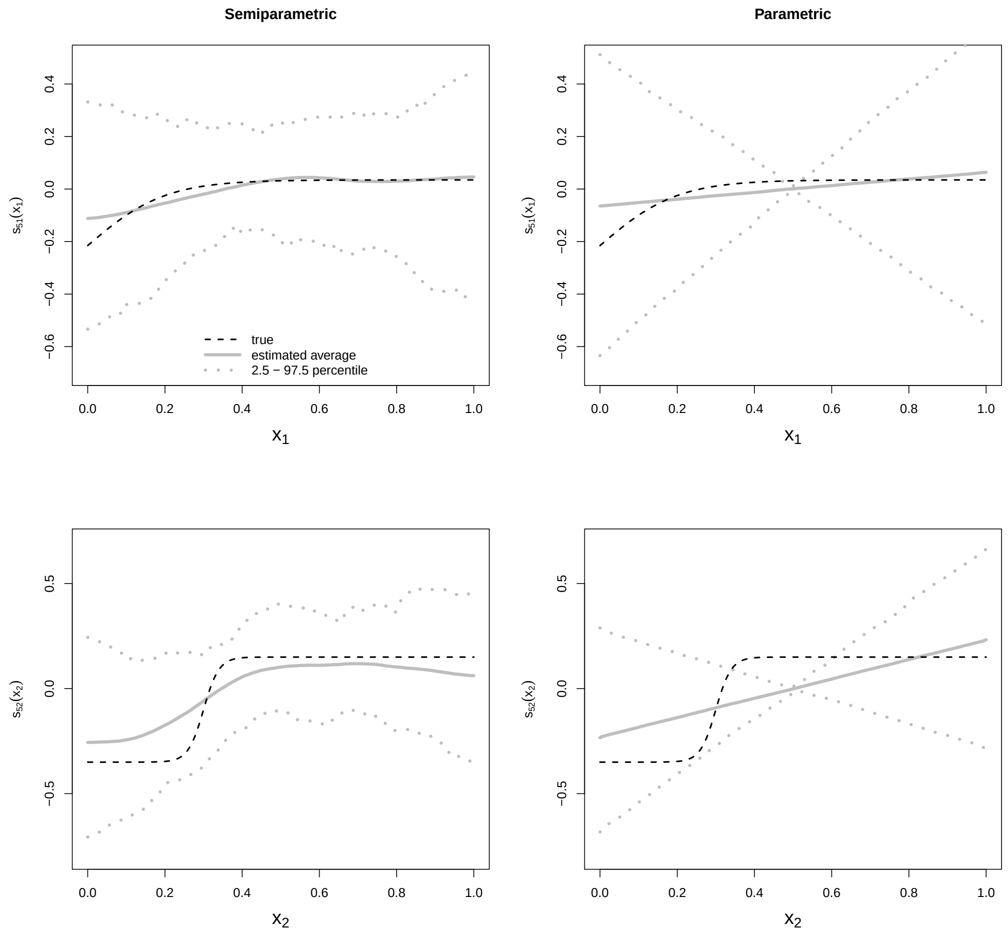


Figure 3.7: Average posterior effects (grey solid line) and 2.5 – 97.5 pointwise percentiles (grey dotted lines) of the concomitant covariates on the log-odds of mixture weights for group 5, estimated with the semiparametric approach over 100 simulated datasets. The true effect is represented by a black dashed line.

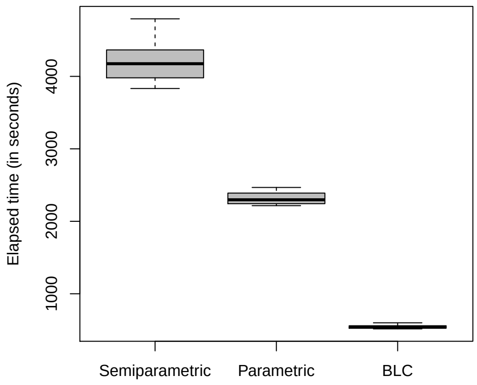


Figure 3.8: Time employed by each algorithm to complete 5000 iterations (initialization and posterior inference included), for each of the 100 replications, with fixed G=6.

of them, for each of the 100 samples, with the current setting. The reported times refer to analyses performed using an IBM x3750 M4 server with 4 Intel Xeon E5-4620 processors with 8 cores and 128GB RAM. As expected, the introduction of concomitant covariates result in an increase in the execution times. Furthermore, the impact of the increase in complexity due to the use of Bayesian P-spline is also evident.

3.3 Application: Brexit data

The proposed empirical analysis is based on data about the Parliamentary votes on Brexit, sometimes referred to as "meaningful votes", that are the parliamentary votes under the terms of Section 13 of the United Kingdom's European Union (Withdrawal) Act 2018 (Parliament, 2018). This Act requires the government of the United Kingdom (UK) to bring forward an amendable parliamentary motion at the end of the Article 50 negotiations between the government and the European Union (EU) to ratify the Brexit withdrawal agreement.

The sample considered consists of $n = 638$ members of the UK Parliament (MPs) who voted 16 divisions during the period 25/03/2019 – 01/04/2019. In the United Kingdom, a member of Parliament (MP) is an individual elected to serve in the House of Commons, the lower house of the Parliament of the

United Kingdom. The Commons is an elected body consisting of 650 members representing constituencies (electoral areas). Among the MPs not taken into account in this study there are 4 Speakers (one speaker and 3 deputies) who neither took part nor voted, 7 members of the political party Sinn Féin which followed a policy of abstentionism (refusing to attend Parliament or vote on bills) and one MP who passed away on February 17th, 2019 and was replaced after April 4th, 2019.

At the end of March 2019, the government had not won any of the meaningful votes. This led to a series of non-binding "indicative votes" on potential options for Brexit, and also to a delay of the departure date. The amendment tabled by Conservative MP Sir Oliver Letwin on March 25th, to take power to control business in the Commons away from the government on March 27th, to allow MPs to put forward business motions relating to Brexit, passed 329 – 302. Instead, the one tabled by Labour MP Dame Margaret Beckett was defeated 311 – 314. This amendment would have required Parliament to vote favourably for a "no deal" Brexit, or request an extension to Article 50 if the government was without a deal within seven days of leaving the European Union. The amended main motion (Letwin but not Beckett) passed 327 – 300. As a result of the Letwin amendment's success, indicative votes on Parliament's preferred Brexit options were held on March 27th. Eight propositions were voted upon, of which all eight failed:

- (B) No deal – Conservative MP Mr John Baron's option to immediately leave the EU without any deal (For: 160, Against: 400);

- (D) Common market 2.0 – Conservative MP Mr Nicholas Boles's proposal to join the Single Market and a customs union (For: 189 – against: 283);

- (E) EFTA and EEA – Conservative MP Mr George Eustice's proposal to remain in the Single Market outside of a customs union (For: 64 – Against: 377);

- (J) Customs union – Conservative MP Mr Kenneth Clarke's proposal for a permanent customs union (For: 265 – Against: 271);

- (K) Labour's plan – Labour's alternative position proposed by MP Mr Jeremy Corbyn, including a comprehensive customs union with the EU, close alignment with the Single Market, dynamic alignment on rights and protections, commitments on participation in EU agencies and funding programmes, and clear agreements on the detail of future security arrangements (For: 237 – Against: 307);

- (L) Revocation to avoid no deal – Scottish National Party MP Ms Joanna Cherry's proposal's to revoke Article 50 (For: 184 – Against: 293);
- (M) Confirmatory public vote – Labour MP Dame Margaret Beckett's proposal for a public vote on any withdrawal bill (For: 268 – Against: 295);
- (O) Maneged no deal – Conservative MP Mr Marcus Fysh's proposal to immediately leave the EU seeking a tariff-free trade agreement (For: 139 – Against: 422).

It is worth noting that only the first and the last propositions express clear pro-leave positions, while the other 6 aim at mitigating the effects of Brexit, or even stopping it.

As Parliament had agreed to an extension of Article 50 to June 30th, the possibility of a third meaningful vote was raised and took place on March 29th, 2019. Mrs Theresa May promised she would resign as Prime Minister if the Withdrawal Agreement passed. In the end, Mrs Theresa May's deal was voted down again (For: 286 – Against: 344), albeit by a smaller margin than in the previous two votes that took place on January, 15th, and March, 12th, respectively. Further indicative votes were held on April 1st on four propositions chosen by the Speaker, all of which failed:

- (C) Customs union by Conservative MP Mr Kenneth Clarke (For: 273 – Against: 276);
- (D) Common market 2.0 by Conservative MP Mr Nicholas Boles (For: 261 – Against: 282);
- (E) Confirmatory public vote by Labour MPs Mr Peter Kyle and Mr Phil Wilson (For: 280 – Against: 292);
- (G) Parliamentary supremacy by Scottish National Party MP Ms Joanna Cherry (For: 191 – Against: 292).

Notice that all the four proposals were modified version of those put to the vote on March 27th, although only the proposers of the confirmatory public vote and the title of Ms Cherry's one changed (previously, "revocation to avoid no deal"). The proposal of a third round of indicative votes, to be held on April, 8th, was then rejected. Due to huge opposition to the fourth withdrawal agreement, on May, 24th, Mrs Theresa May announced she would resign as Conservative Party leader and Prime Minister on June, 7th.

The main purpose of the analysis described in this paper is to identify groups of MPs whose opinions about Brexit, in terms of votes for the aforementioned divisions, are similar. Furthermore, the influence of some concomitant information related to the MPs themselves, or the constituencies they represent, on group membership is considered. In particular, $J = 3$ concomitant covariates are included in the analysis:

- age of the MP;
- share of Leave votes at the Brexit referendum in parliamentary constituencies;
- "safeness" of the seat of each MP.

Regarding the second covariate, it is worth noting the Brexit referendum vote was not counted by constituencies except in Northern Ireland. Some local councils (districts) republished local results by electoral ward or constituency. Some constituencies are coterminous with (overlap) their local government district. For the others, Hanretty (2017) estimated through a demographic model the Leave and Remain vote.

About the third covariate, Apostolova et al. (2017) analyzed and made available the results of the 2017 UK general election and, in particular, the number of votes taken by each party for each of the 650 constituencies. In this dataset, 12 main parties are considered, while all the others are gathered, unless one of these won the seat: the votes taken by the winning party are counted separately and placed in another category, for a total of $K = 13$ categories. To quantify how much an MP, or the party they represent, was appreciated in the constituency they were elected into, a measure of the degree of heterogeneity of votes among parties in that constituency is considered.

In particular, by denoting with $\alpha_{\kappa i}$ the share of votes taken by party $\kappa\,(\kappa = 1, \dots, K)$, the entropy of the votes in the i-th constituency can be computed as:

$$\mathrm{EN}(\boldsymbol{\alpha}_i) = -\sum_{\kappa=1}^{K} \alpha_{\kappa i} \log(\alpha_{\kappa i}). \tag{3.7}$$

In this case, the entropy $\mathrm{EN}(\boldsymbol{\alpha}_i)$ quantifies the uncertainty in predicting the number of votes taken by a party that is drawn at random in a given constituency. It ranges from a minimum of 0, which corresponds to a situation of no heterogeneity (i.e. all votes are taken by a single party), to a maximum of $\log(K)$, indicating that there is equidistribution of votes between the parties. Thus, by considering $\exp(\mathrm{EN}(\boldsymbol{\alpha}_i))$ a quantity that can be interpreted as the effective number of competing political parties (or candidates) in a given constituency is obtained.

3.3.1 Brexit results

The results of the 16 divisions mentioned in Section 3.3 are imported in `R` through the package "`hansard`". To identify groups of MPs with similar opinion about Brexit, values of G ranging from 1 to 15 are considered for the semiparametric MoE model described in Section 3.1. For each value of G, 4000 MCMC draws are considered after a burn-in of 1000 draws. The optimal number of components suggested by the AICM is 11, although Figure 3.9 shows that the AICM curve is quite flat between 9 and 14. Clusters' composition in terms of political party membership is shown in Table 3.8. The results relative to the most meaningful clusters detected in the analysis are described in the main text; additional results can be found in Appendix A. For interpretation reasons, it is worth mentioning that the log-odds $\eta_g(\mathbf{x})$, $g = 1, \ldots, 10$, are expressed using Cluster 11 as the reference. This cluster is the most numerous, with 217 MPs, and is characterized by an extreme pro-Leave (even pro-no-deal) position, as shown in Figure 3.10. Looking at the composition, 211 out of these 217 MPs are conservatives, including Mr John Baron, proposer of the no-deal, Mr Marcus Fysh, proposer of the managed no-deal, and Mr Boris Johnson, who subsequentely became prime minister, on July 24th, 2019.

Cluster 8 is the most heterogeneous in terms of political party membership. In fact, 47 MPs out of 54 belonging to the Scottish National Party (including Ms Joanna Cherry, who proposed the revocation of Article 50 to avoid no deal), Plaid Cymru and the Liberal Democrats are in this group, together with 7 more MPs belonging to 3 other different parties. Since there is no area where the amount of votes for both the two leading parties (Conservatives and Labourists) is negligible, the constituencies represented in this group are usually characterized by the presence of at least 3 political parties, i.e. the one that actually won plus the two leading parties. In other words, the probability for an MP to belong to this cluster is higher if the effective number of competing candidates is greater than 3. This threshold (nonlinear) effect is represented in the third plot of Figure 3.11, while it looks slightly smoother for Cluster 4 (Figure A.1 in the Appendix), mostly made up by the whole Democratic Unionist Party. A mild nonlinearity appears also in the second plot of Figure 3.11, representing the effect of the opinion about Brexit in the constituency. In particular, the probabiliy for an MP to be assigned to Cluster 8 decreases as the fraction of Leave votes in his constituency increases. In fact, most of the constituencies represented in this

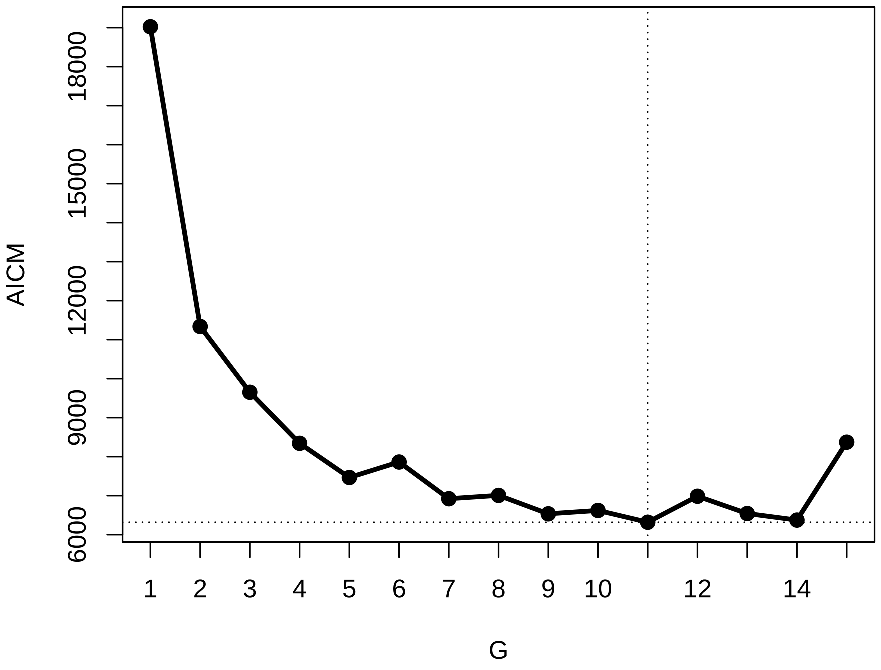


Figure 3.9: AICM values corresponding to different number of components G for the semiparametric gating network mixture of experts model

Table 3.8: Posterior allocation and political party membership of the $n = 638$ MPs. Adjusted Rand Index: 0.458. (C = Conservatives, DUP = Democratic Unionist Party, GP = Green Party, Ind = Independents, Lab = Labourists, LD = Liberal Democrats, PC = Plaid Cymru, SNP = Scottish National Party).

	Cluster										
Party	1	2	3	4	5	6	7	8	9	10	11
C	0	21	16	5	21	0	33	3	3	0	211
DUP	0	0	0	10	0	0	0	0	0	0	0
GP	0	0	0	0	0	0	0	0	1	0	0
Ind	0	0	0	0	0	1	0	2	12	0	1
Lab	19	1	4	0	0	108	1	2	6	101	5
LD	0	1	0	0	0	1	0	9	1	0	0
PC	0	0	0	0	0	0	0	4	0	0	0
SNP	0	0	0	0	0	0	1	34	0	0	0

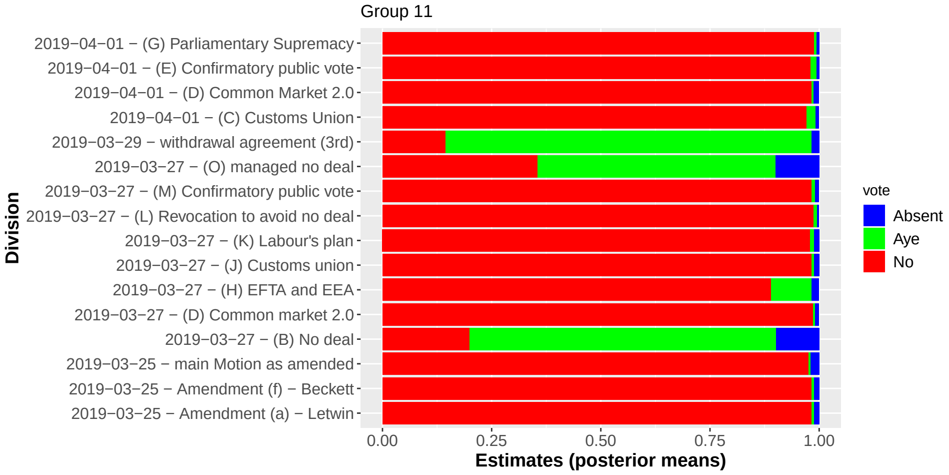


Figure 3.10: Vote estimates (posterior means) for Cluster 11.

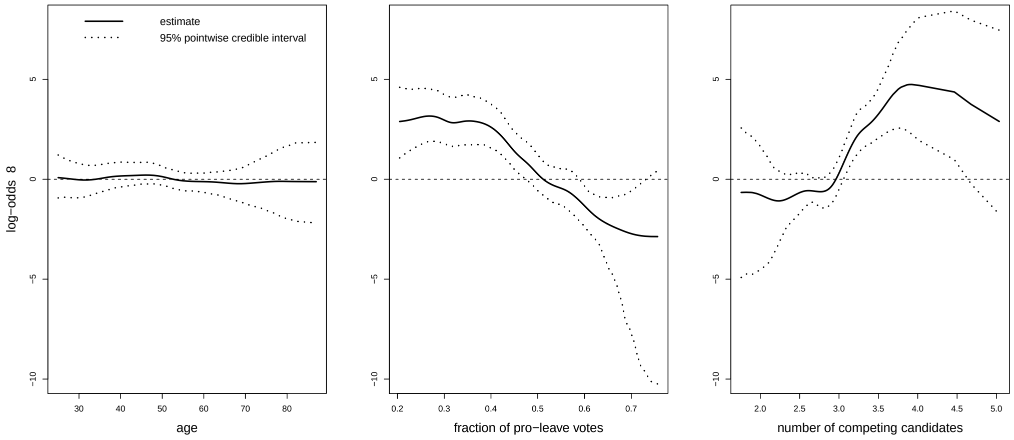


Figure 3.11: Estimated smooth effects for Cluster 8.

group are known to be pro-Remain, and so are the aforementioned parties, as confirmed by the posterior means of votes shown in Figure 3.12.

Cluster 9 is the last remaining group which is not mostly composed by MPs belonging to the leading parties. In particular it includes a large portion of the Independents group, made up by MPs who quitted the party they were elected with. Despite this, they seem to reflect the pro-Remain position of the constituencies where they were elected from (Figures A.3 and A.4 in the Appendix).

Also Cluster 6 and 10 are characterized by a pro-Remain position towards Brexit (see Figure 3.14 and Figure 3.16). Both are mostly made up by Labourists and represent constituencies where the competition is usually concentrated around the two leading parties. Hence, in this case the effective number of competing candidates has the opposite nonlinear effect on the probability of belonging to these groups (see the third plot in Figure 3.13 and Figure 3.15), if compared to the one observed for Cluster 9. The main difference between these two clusters is due to the opinion towards Brexit of the respectively represented constituencies. In particular, as it can be seen in the second plot of Figure 3.13 and Figure 3.15, the fraction of leave votes has a mild nonlinear effect on the probability of belonging to Cluster 10 (similar to the one observed in Figure 3.11), while it has no effect for Cluster 6. This seems to be reflected by a more extreme position of the MPs of Cluster 10, especially towards the revocation to avoid no deal (Motions L, G). The aforementioned Dame Margaret Beckett, first proposer of the confirmatory public vote, as well as of the homonymous amendment, belongs to this group, while

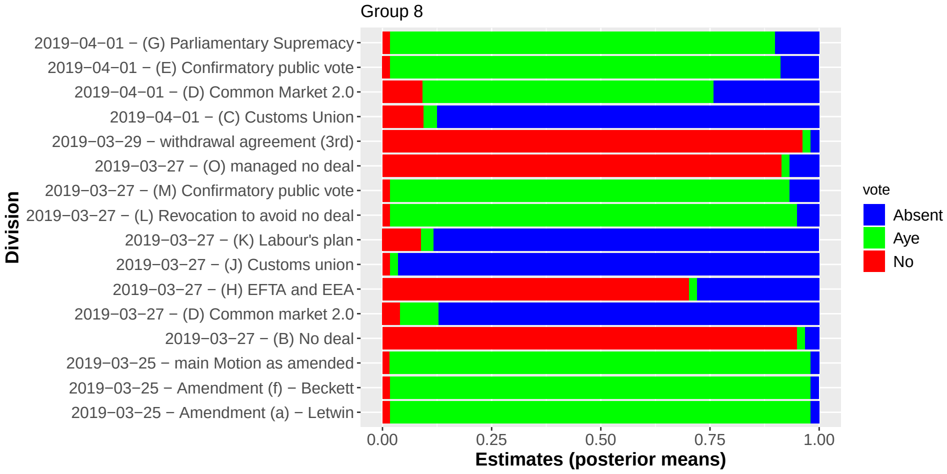


Figure 3.12: Vote estimates (posterior means) for Cluster 8.

the second co-proposers of the confirmatory public vote, Mr Peter Kyle and Mr Phil Wilson are divided between Clusters 10 and 6, respectively. Cluster 6 is also characterized by the presence of Mr Jeremy Corbyn, proposer of the Labour's alternative plan for Brexit, and leader of the Labour Party until the 2019 United Kingdom general election.

There is a third Labour group, Cluster 1, which represents the pro-Leave minority of the party. More precisely, Figure 3.17 shows that the MPs belonging to this cluster distinguish themselves from party members assigned to different clusters because they were elected in the few pro-Leave Labour constituencies. This seems to affect their opinion in terms of divisions, which sensibly departs from the party line. However, this cluster does not look homogeneous, but rather divided into two fractions. In particular Figure 3.18 shows 5 divisions where the "no" fraction is close to 50%.

Cluster 7 includes all of the MPs of the Cabinet (of the time), including Mrs Theresa May. A high rate of abstensionism is present in this group, apart from the Letwin-Beckett amendment and the third meaningful vote (Figure A.6 in the Appendix). According to Figure A.5 in the Appendix, covariates seem to have no effect here. Cluster 2 is (mostly) conservative as well, but the opinion of its MPs looks more heterogenous with respect to Clusters 11 and 7. In particular, a clear anti no-deal position emerges in Figure A.8 in the Appendix. This might be due to the characteristics of the MPs themselves and the constituencies they represent. In fact, although Conservatives, most of these MPs come from constituencies that expressed a pro-Remain position and this mild tendency is reflected in the second plot

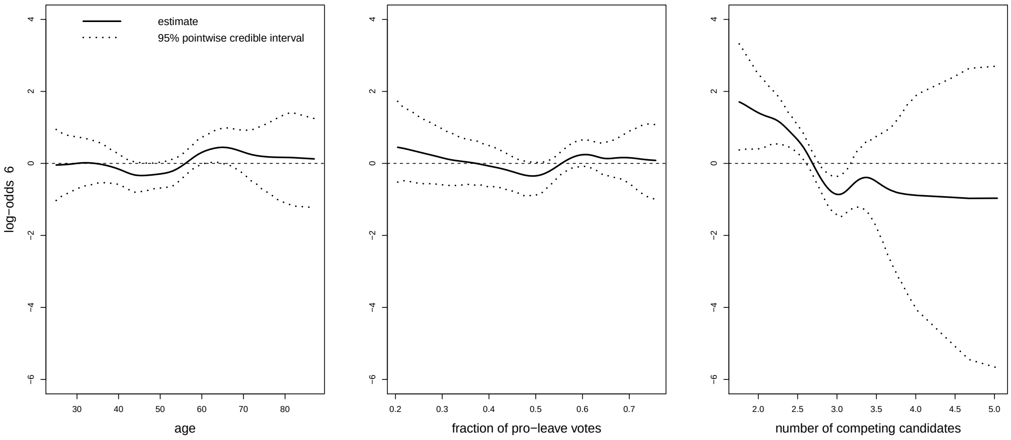


Figure 3.13: Estimated smooth effects for Cluster 6.

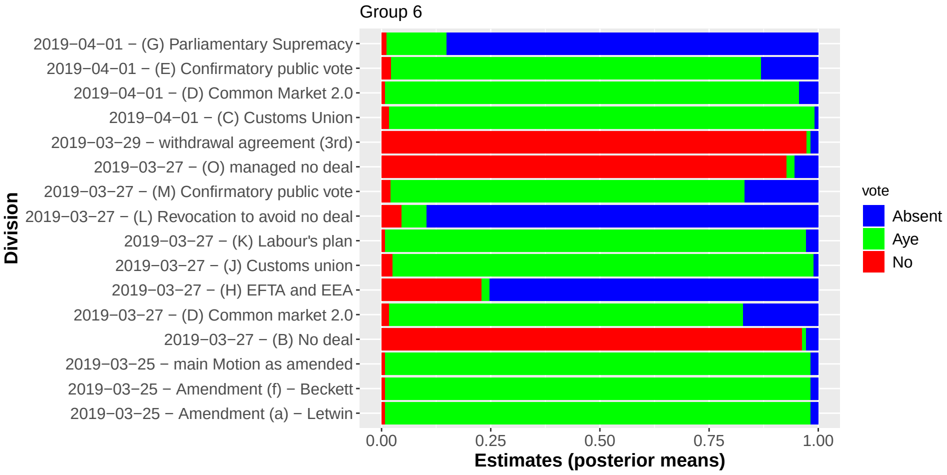


Figure 3.14: Vote estimates (posterior means) for Cluster 6.

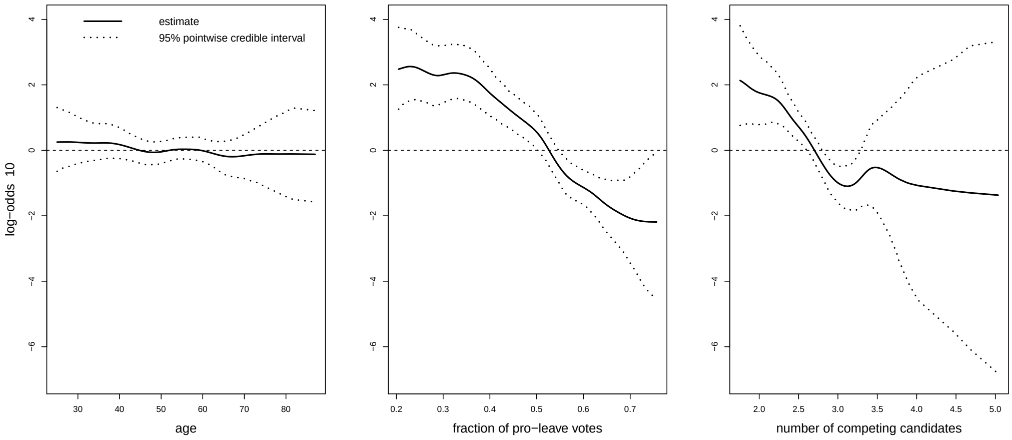


Figure 3.15: Estimated smooth effects for Cluster 10.

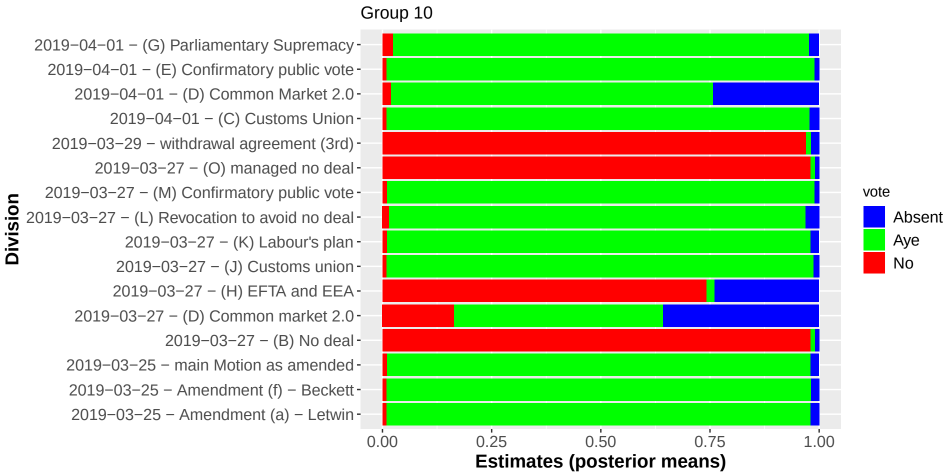


Figure 3.16: Vote estimates (posterior means) for Cluster 10.

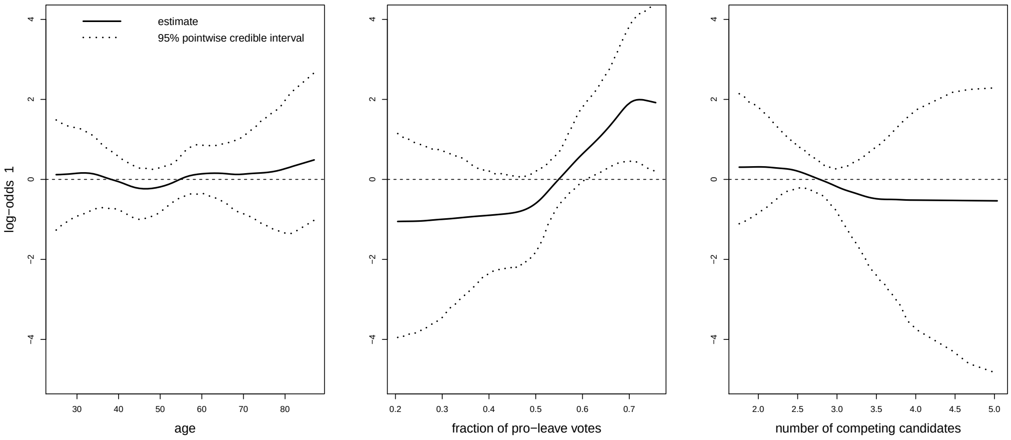


Figure 3.17: Estimated smooth effects for Cluster 1.

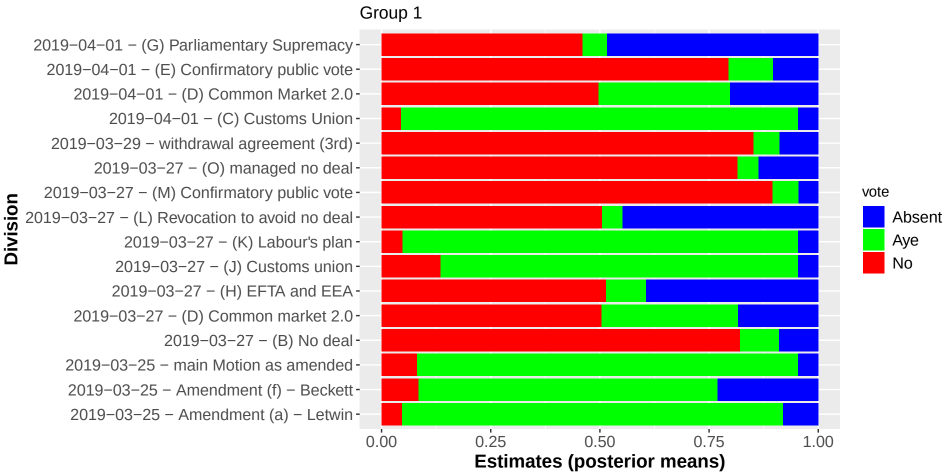


Figure 3.18: Vote estimates (posterior means) for Cluster 1.

of Figure A.7 in the Appendix. Both Mr Nicholas Boles and Mr Kenneth Clarke, proposers of the 2 couples of divisions for a customs union with the EU (“Common market 2.0” and, indeed, “Customs union”), belong to this group.

The remaining conservative clusters, Cluster 3 and Cluster 5, mainly distinguish themselves for their moderate positions expressed through the divisions (Figures A.10 and A.12) rather than their concomitant information (Figures A.9 and A.11). In particular, the latter is characterized by the presence of Mr George Eustice, whose proposal to remain in the Single Market outside of a customs union resulted in the division titled “(E) EFTA and EEA”, which has been favorably voted only by this cluster and Cluster 2.

Chapter 4

Continuous manifest variables

4.1 Model specification

Consider a sample comprised of continuous outcome observations $\{\mathbf{y}_i\}$, $i = 1, \ldots, n$, from a population that is clustered into a G-component finite mixture model. Let $\mathbf{c}$ be a vector of latent variables such that, for each unit i, $c_i = g$ if i belongs to cluster g. Furthermore, suppose that $\mathbf{c}$ has a discrete distribution with $\Pr(c_i = g|x_i) = \pi_g(\mathbf{x}_i)$, for $g = 1, 2, \ldots, G$, where $\mathbf{x}_i$ is a $(J+1)$-dimensional vector of covariates, including $J - J^*$ metrical variables $x_{J^*+1}, \ldots, x_J$, with $J^* \in \{0, 1, \ldots, J\}$, and weights $\pi_g(\mathbf{x}_i)$ are defined as in Equation (2.14), for $i = 1, \ldots, n$. In the univariate case, conditioning on $\mathbf{c}_i$ and $\mathbf{x}_i$, it is assumed that y_i follows a normal distribution with mean μ_{c_i} and variance $\sigma^2_{c_i}$. Hence, the conditional density of y_i given $\mathbf{x}_i$ can be written as a mixture of normals:

$$f(y_i|\mathbf{x}_i) = \sum_{g=1}^{G} \pi_g(\mathbf{x}_i) f_N\left(y_i|\mu_g, \sigma^2_g\right), \tag{4.1}$$

with $f_N(\cdot)$ being the density of a univariate normal distribution.

If $\mathbf{y}$ is Q-dimensional, then conditioning on $\mathbf{x}_i$, each $\mathbf{y}_i$ follows a finite mixture of multivariate normals:

$$f(\mathbf{y}_i|\mathbf{x}_i) = \sum_{g=1}^{G} \pi_g(\mathbf{x}_i) f_{MVN_Q}\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right). \tag{4.2}$$

where $f_{MVN_Q}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a Q-variate normal distribution, with vector of means $\boldsymbol{\mu}_g$ and positive definite covariance matrix $\boldsymbol{\Sigma}_g$, $g = 1, \ldots, G$.

4.2 Bayesian inference

Assume that manifest variable $\mathbf{Y}$ is Q-variate, with $Q > 1$. As in Bensmail et al. (1997), a conjugate prior can be used:

$$f(\boldsymbol{\mu}_{\mathbf{1}}, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G) = \prod_{g=1}^{G} f(\boldsymbol{\mu}_g|\boldsymbol{\Sigma}_g) f(\boldsymbol{\Sigma}_g). \tag{4.3}$$

An Inverse Wishart prior is assigned to the covariance matrix $\boldsymbol{\Sigma}_g$, $g = 1, \ldots, G$, and it is assumed that, conditional on $\boldsymbol{\Sigma}_g$, the g-th mean vector $\boldsymbol{\mu}_g$ is Q-variate normally distributed with variance depending on $\boldsymbol{\Sigma}_g$:

$$\boldsymbol{\Sigma}_g \sim IW(a_\Sigma, \mathbf{B}_\Sigma), \quad \boldsymbol{\mu}_g \sim MVN_Q\left(\boldsymbol{\mu}_0, \frac{1}{\upsilon}\boldsymbol{\Sigma}_g\right), \quad g = 1 \ldots, G, \tag{4.4}$$

where a_Σ, $\mathbf{B}_\Sigma$, $\boldsymbol{\mu}_0$, υ are known. In particular, Bensmail et al. (1997) suggest using the following data-dependent hyperparameters: $a_\Sigma = 2.5, \mathbf{B}_\Sigma = 0.5\mathbf{S}_y, \boldsymbol{\mu}_0 = \bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_Q), \upsilon = 1$, with $\bar{\mathbf{y}}$ and $\mathbf{S}_y$ denoting the sample mean vector and the sample covariance matrix, respectively.

Following Frühwirth-Schnatter (2006, Section 6.3), the component densities' parameters $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ can be sampled conditional on the component indicator $\mathbf{D}_g$, $g = 1, \ldots, G$, in step 7 of the MCMC algorithm described in Section 2.2.3, from the following full conditionals:

$$\begin{aligned}
\boldsymbol{\mu}_g|\boldsymbol{\Sigma}_g, \mathbf{D}_g \sim & MVN_Q\left(\frac{1}{\upsilon + \sum_{i=1}^{G} D_{gi}}\left(\upsilon\boldsymbol{\mu}_0 + \mathbf{y}'\mathbf{D}_g\right), \frac{1}{\upsilon + \sum_{i=1}^{n} D_{gi}}\boldsymbol{\Sigma}_g\right), \\
\boldsymbol{\Sigma}_g|\boldsymbol{\mu}_g, \mathbf{D}_g \sim & IW(a_\Sigma + 0.5\left(\sum_{i=1}^{n} D_{gi} + 1\right), \\
& \qquad \mathbf{B}_\Sigma + 0.5\upsilon(\boldsymbol{\mu}_g - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_g - \boldsymbol{\mu}_0)' + \left(\mathbf{y}'\mathbf{D}_g - \boldsymbol{\mu}_g\right)\left(\mathbf{y}'\mathbf{D}_g - \boldsymbol{\mu}_g\right)').
\end{aligned} \tag{4.5}$$

Finally, the posterior in the last step can be made explicit as:

$$\Pr(D_{gi} = 1|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}_g, \boldsymbol{\gamma}_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \propto \frac{\lambda_{gi}}{\sum_{g=1}^{G} \lambda_{gi}} f_{MVN_Q}\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right), \quad g = 1, \ldots, G. \tag{4.6}$$

The resulting MCMC algorithm can be easily adapted to the univariate case ($Q = 1$), in Equation (4.1), by replacing the prior distributions – and, consequently, the full conditionals – with their univariate analogues: the (univariate) normal distribution, for the mean parameter μ_g, and the Inverse Gamma distribution, for the variance parameter σ_g^2, $g = 1, \ldots, G$; see Frühwirth-Schnatter (2006, Section 6.1) or Marin et al. (2005).

4.3 Soccer player positions data

Pettersen et al. (2014) present a dataset of body-sensor traces and corresponding videos from three professional soccer games captured in late 2013 at the Alfheim Stadium in Tromsø, Norway. Tromsø - Stromsogodset is selected for this study, since it is the only one which is valid for the national competition. This game was played on November 3rd, 2013, and it ended with no scores. Player data, including field position, are sampled at 20 Hz using the ZXY Sport Tracking system.

The aim of this analysis is to apply the proposed method study how a player's position is affected by a teammate's one and possibily identify a finite number of different phases of the game. Obviously, this relationship depends on many factors, such as the two player's role and which area of the field they are supposed to cover: for example, the position of a striker should to be more influenced by another striker's position, rather then a defender's one. For this reason, this study focuses on couples of players playing close to each other. Due to privacy reasons, each player is identified only by a random numeric tag instead of his name or the number he wears, and attempts of re-identificantions are not allowed. Thus, after plotting each player's location on the field during the whole game, the study starts by concentrating on the player covering the right full-back position, identified with tag 9, and assuming that his longitude and latitude (y_1 and y_2, respectively) can reasonably be approximated by a bivariate normal distribution. Then, the two-dimensional location of the centre-back playing closer to him, Player 13, are taken as concomitant covariates (x_1, x_2) in the following gating network mixture of experts model:

$$f(\mathbf{y}_i|\mathbf{x}_i) = \sum_{g=1}^{G} \pi_g(\mathbf{x}_i) f_{MVN_2}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g),$$

where G is unknown and the effect of the covariates on the weights $\pi_g(\mathbf{x})$, $g = 1, \ldots, G$ is specified according to Equation 2.14. The results are presented in Section 4.3.1. The analysis is then repeated by focusing on the opposite side of the field, studying the position of the left full-back, Player 8, depending on the left centre-back, identified with tag 2, to understand if any simmetry is present between two sides of the backfield. The results for this second part of the study are given in Section 4.3.2.

To carry out the analysis, some assumptions are made. In particular, the observations are assumed to be independent across time: to make this assumption more realistic, the data are thinned out to 501 observations over more than 90 minutes of play, leading to a distance of approximately 10

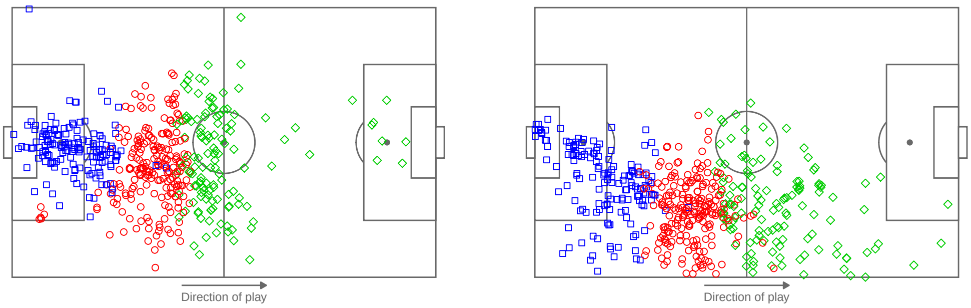


Figure 4.1: Locations of Player 13 (left plot) and Player 9 (right plot). Different colors and dot symbols correspond to different clusters.

seconds between each pair of consecutive observations. Since between the first and the second half of the game the direction of play changes, preparing this dataset requires a 180° rotation of the locations observed during the second half. The two dimensions of the location of the centre-backs, x_1 and x_2, representing the long and short side of the field, respectively, are assumed to have an additive effect on the log-odds of the component weights. For both analyses, the algorithm is run for fixed G ranging from 1 to 6. The results produced by the respective best models, in terms of AICM, are selected.

4.3.1 Right-back results

Focusing the right side of the field, the best model according to AICM has $G = 3$ components. Figure 4.1 shows the locations of the two players during the game, allocated according to the 3-component mixture of experts model. The clusters does not seem well separated. Indeed, without considering the position of Player 13, the best (according do AICM) finite mixture of normals with constant component weights suggests the presence of a single component. These clusters may be interpreted as phases of the game: in particular, the blue dots identify the defensive phase, the green triangles the offensive one, while the red square indicate an intermediate phase. For each component, posterior estimates of the mean and variance parameters $(\hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)$ are reported in Table 4.1. For illustrative purpose, the correspond-

Table 4.1: Estimated parameters of the component densities (and standard deviances).

	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_{12}$
Cluster 1	38.81	16.71	36.21	56.18	-3.12
	(0.55)	(0.59)	(4.90)	(3.86)	(6.22)
Cluster 2	17.42	23.98	73.85	79.61	-32.92
	(0.85)	(0.77)	(9.23)	(9.59)	(7.16)
Cluster 3	59.84	18.10	147.56	102.83	-40.14
	(1.24)	(0.95)	(18.55)	(13.4)	(12.18)

ing MCMC draws for Cluster 3 (green) are shown in Figures B.1 and B.2 in the Appendix B. The resulting estimated component densities $f_N(\hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)$, $g = 1, 2, 3$, are depicted in Figure 4.2. It can be noticed that the overlapping ellipses representing these bivariate normal distributions tend to exceed the limits of the field, warning that an asymmetric (or even truncated) alternative could be a more appropriate form to be assumed for these component densities.

The intermediated phase, originally associated to the first component (in red), is taken as the reference to define the log-odds of mixture weights as

$$\eta_g(\mathbf{x}_i) = \log \frac{\pi_g(\mathbf{x}_i)}{\pi_1(\mathbf{x}_i)}, \quad g = 2, 3.$$

The splines' coefficients are transformed according to Section 2.2.5, and the estimated effect of the location of Player 13 on the probability of both the defensive and the offensive phase of Player 9 are reported in Figure 4.3. The clusters differ mainly with respect to the long side (x_1) of the field, while the location on the short side seems to be less impactful. Lower values of the longitude for Player 13 seem to lead to a higher probability that Player 9 is in the defensive phase, implying him covering the backfield too. This probability drops as x_1 grows, increasing the odds of the offensive phase, characterized by a higher longitude and variability. A huge amount of variability of the estimated effects can be noticed in the plots, especially when the functions reach large absolute values that correspond to 0 or 1 on the scale of the probability. This might be also due to the fact that the locations of the players are not uniformly distributed along the field. It is worth mentioning that this uneven distribution of the observations seems coherent with the specific roles of the two players considered in this analysis.

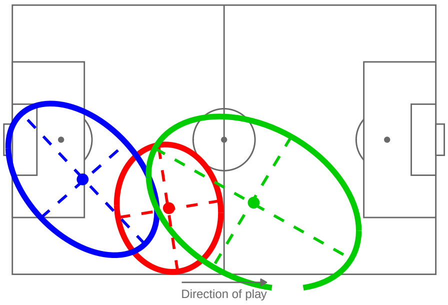


Figure 4.2: Estimated component density. The ellipse corresponding to the g-th cluster is centered in $\hat{\boldsymbol{\mu}}_g$, $g = 1, \ldots, 3$, while the size corresponds to a 0.9 confidence level.

4.3.2 Left-back results

On the opposite side, the best model, according to AICM, has $G = 5$ components. Figure 4.4 shows the locations of the two players during the game, allocated as the 5-component mixture of experts model indicates. Some overlap is present also between these clusters and, in fact, a finite mixture of normals with constant component weights keeps not being able to identify more than one cluster, according do AICM. As for the previous case study, an interpretation of these clusters as moments of the game can be provided: moving along the long side of the field, from the blue cluster to the violet one, different degrees of offensiveness (or defensiveness) are represented, while the green cluster may identify a phase when the game (or the ball) moves on the left side. For each component, posterior estimates of the mean and variance parameters $(\hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)$ are reported in Table 4.2. For illustrative purpose, the corresponding MCMC draws for Cluster 2 (green) are shown in Figures B.3 and B.4 in the Appendix B. The resulting estimated component densities $f_N(\hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)$, $g = 1, \ldots, 5$, are represented in Figure 4.5. Again, the ellipses tend to exceed the limits of the field, comfirming the previous remark about the appropriateness of the chosen form for the component densities.

Again, Cluster 1 (the red one) is taken as the reference, and the estimated effect of the location of Player 2 on the weights corresponding to the four remaining game phases of Player 8 are reported in Figures 4.6 to 4.9. With

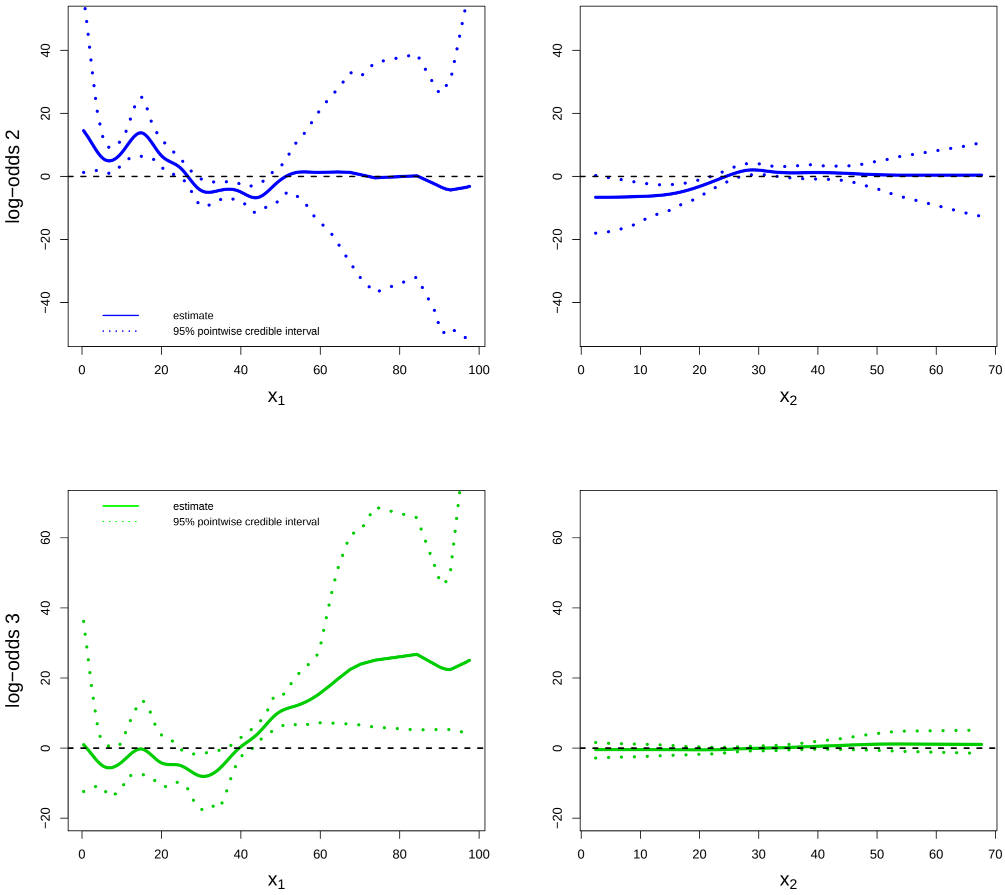


Figure 4.3: Estimated effect (and 95% pointwise credible interval) of the location of Player 13, (x_1, x_2) on the log-odds of the mixture weights, for Cluster 2 (upper, in blue) and Cluster 3 (lower, in green).

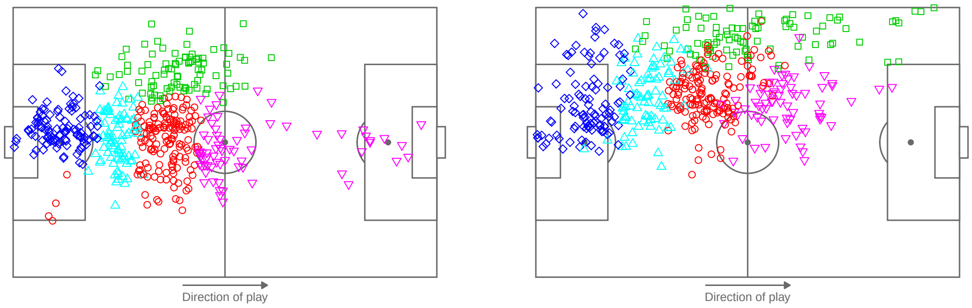


Figure 4.4: Locations of Player 2 (left plot) and Player 8 (right plot). Different colors and dot symbols correspond to different clusters.

Table 4.2: Estimated parameters of the component densities (and standard deviances).

	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_{12}$
Cluster 1	42.85	46.43	42.39	39.82	9.85
	(0.73)	(0.57)	(5.59)	(5.17)	(4.18)
Cluster 2	54.55	60.39	290.39	19.31	21.09
	(2.03)	(0.53)	(46.60)	(3.19)	(8.91)
Cluster 3	13.22	44.05	40.30	71.40	12.94
	(0.68)	(0.87)	(5.83)	(10.02)	(5.74)
Cluster 4	27.76	45.46	33.78	52.81	14.69
	(1.05)	(0.97)	(6.81)	(9.53)	(6.52)
Cluster 5	61.14	43.52	78.34	35.61	7.60
	(1.18)	(0.74)	(14.44)	(6.56)	(6.90)

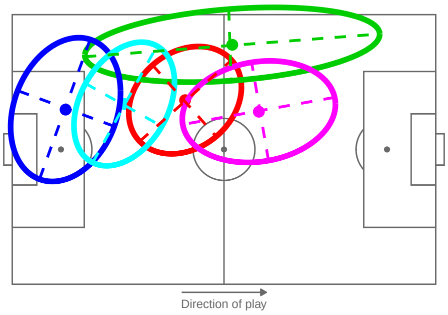


Figure 4.5: Estimated component density. The ellipse corresponding to the g-th cluster is centered in $\hat{\boldsymbol{\mu}}_g$, $g = 1, \ldots, G$, while the size corresponds to a 0.9 confidence level.

respect to the results of the previous study, now both dimensions of the field present a notable effect, especially for the green cluster (Figure 4.6). In particular, when Player 2 moves to the left side of the field, the odds of belonging to this cluster increase. The interpretation of the remaining estimated effects is in line with the one given in the previous Section. An increase in the width of the posterior pointwise credible intervals can still be noticed in the estimates. Also in this second analysis, this increase seems to be related with the uneven distribution of the observation on the field.

Finally, it is possible to conclude that the two full-backs, although playing specular positions, seem to react differently to the moves made by their respective nearby centre-back. In particular, this second analysis seems to highlight an offensive phase which take place on the left wing for Player 8, which has no analogue on the right for Player 9. This may be the result of some specific tactical choices made by the coach. However, it is worth mentioning that one of the limits of this analysis is that it does not allow to distinguish between active and passive phases of the game. This could be possible if the location of the ball was included too among covariates. Unfortunately, no information is available regarding the location of the ball for the examined game.

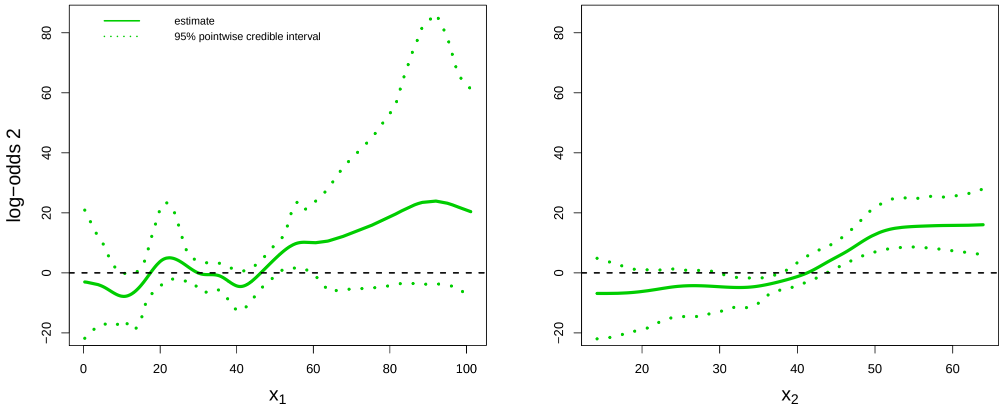


Figure 4.6: Estimated effect (and 95% pointwise credible interval) of the location of Player 2, $(\mathbf{x}_1, \mathbf{x}_2)$ on the log-odds of the mixture weights, for Cluster 2.

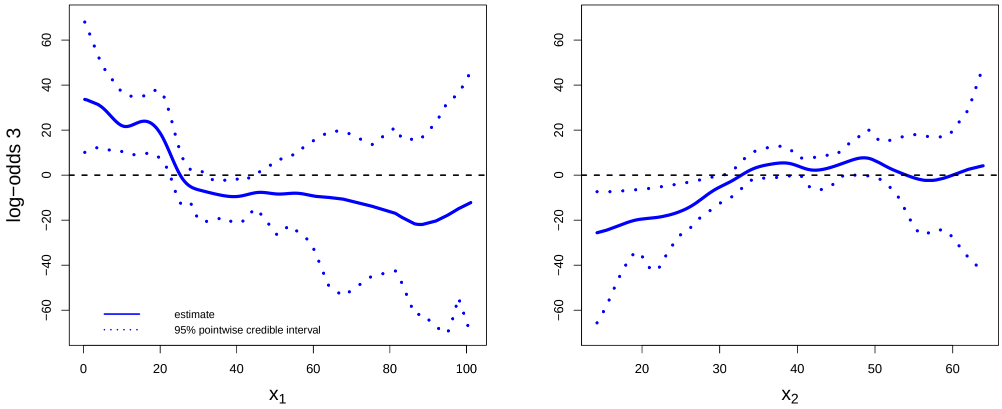


Figure 4.7: Estimated effect (and 95% pointwise credible interval) of the location of Player 2, $(\mathbf{x}_1, \mathbf{x}_2)$ on the log-odds of the mixture weights, for Cluster 3.

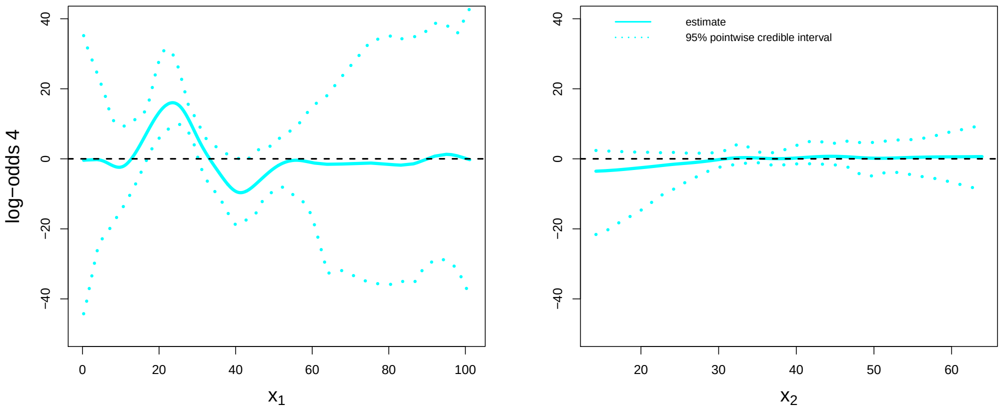


Figure 4.8: Estimated effect (and 95% pointwise credible interval) of the location of Player 2, $(\mathbf{x}_1, \mathbf{x}_2)$ on the log-odds of the mixture weights, for Cluster 4.

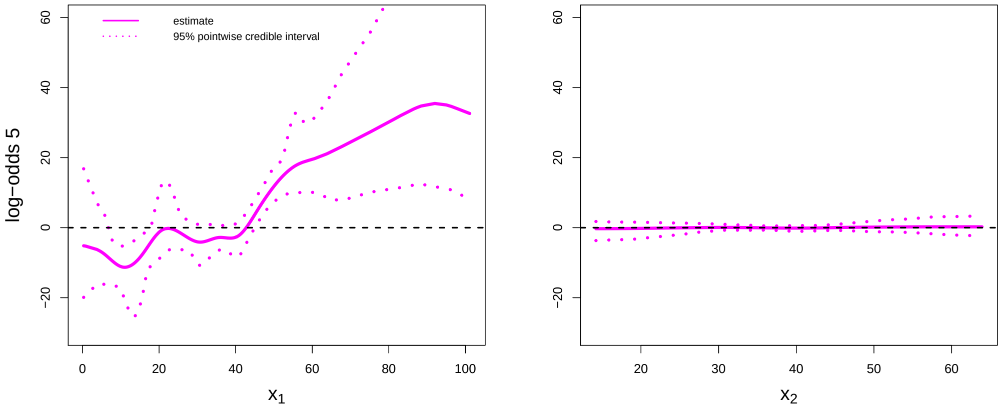


Figure 4.9: Estimated effect (and 95% pointwise credible interval) of the location of Player 2, $(\mathbf{x}_1, \mathbf{x}_2)$ on the log-odds of the mixture weights, for Cluster 5.

Chapter 5

Full mixture of experts models

5.1 Model specification

Suppose that $\{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, n$ is a random sample from a population clustered into G components. Throughout this Chapter, it is assumed that $\mathbf{y}$ is univariate, while $\mathbf{x}$ is $(J+1)$-dimensional, of which the last $J - J^*$ covariates are metrical, with $J^* \in \{0, 1, \ldots, J\}$. Let the latent class variable $\mathbf{c}$, introduced in Section1.1.3, have a discrete distribution $\Pr(c_i = g|\mathbf{x}_i) = \pi_g(\mathbf{x}_i)$, with $\pi_g(\mathbf{x}_i)$ defined as in Equation (2.14), $g = 1, \ldots, G-1$, and $i = 1, \ldots, n$. Conditioning on $\mathbf{c}$ and $\mathbf{x}$, y_i follows a normal distribution with mean $\mu_{c_i}(\mathbf{x}_i)$ and variance $\sigma^2_{c_i}$. It is further assumed that each $\mu_g(\cdot)$, $g = 1, \ldots, G$, is an unknown smooth function of the covariates $\mathbf{x}$. Hence, conditioning on $\mathbf{x}_i$, y_i follows a finite mixture of normals:

$$f(y_i|\mathbf{x}_i) = \sum_{g=1}^{G} \pi_g(\mathbf{x}_i) f_N\left(\mu_g(\mathbf{x}_i), \sigma^2_g\right). \tag{5.1}$$

The corresponding graphical representation is reported in Figure 5.1.

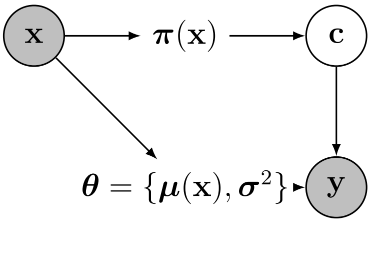

x
$\boldsymbol{\pi}(\mathbf{x})$
c
$\boldsymbol{\theta} = \{\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\sigma}^2\}$
y

Figure 5.1: Graphical model representation of the full mixture of experts model in Equation (5.1); grey-colored circle represent observed quantities.

The additive paradigm is again exploited to define $\mu_g(\mathbf{x}_i)$ as a sum involving $J - J^*$ smooth functions of the metrical covariates $s^{(\mu)}_{gJ^*+1}(\cdot), \ldots, s^{(\mu)}_{gJ}(\cdot)$, together with the usual linear part of the predictor with coefficients $\boldsymbol{\gamma}_g = (\gamma_{g0}, \gamma_{g1}, \ldots, \gamma_{gJ^*})$:

$$\mu_g(\mathbf{x}_i) = \sum_{j=0}^{J^*} \gamma^{(\mu)}_{gj} x_{ij} + \sum_{j=J^*+1}^{J} s^{(\mu)}_{gj}(x_{ij}), \quad g = 1, \ldots, G. \tag{5.2}$$

Here, the superscript $^{(\mu)}$ is used to distinguish the parameters and functions involved from the ones relative to the weights in Equation (2.14). To ensure the indentifiability of the additive predictors in Equation (5.2), each function $s^{(\mu)}_{gj}(x_j)$ is constrained to have zero mean, that is

$$\frac{1}{\text{range}(x_j)} \int s^{(\mu)}_{gj}(x_j) \mathrm{d}x_j = 0, \quad j = J^* + 1, \ldots, J; \; g = 1, \ldots, G. \tag{5.3}$$

Model (5.1), with $\mu_g(\mathbf{x}_i)$ defined as in Equation (5.2), can be referred to as a semiparametric mixture of experts regression model. Similar models are reviewed in Xiang et al. (2019).

5.2 A note on identifiability

Huang and Yao (2012) study a semiparametric mixture of regression models with component weights as smooth functions of a covariate, and implement a modified expectation-maximization type estimation procedure using kernel regression. Huang et al. (2013) extend Huang and Yao (2012) by including smooth covariate effects in both parameters of the normal distribution. These models belong to the wider family of nonparametric mixtures of generalized linear models:

$$f(y_i|\mathbf{x}_i) = \sum_{g=1}^{G} \pi_g(\mathbf{x}_i) f\left(\theta_{g1}(\mathbf{x}_i), \theta_{g2}(\mathbf{x}_i)\right), \tag{5.4}$$

where $\theta_{g1}(\mathbf{x})$ is the component canonical (or natural) parameter, a function of the conditional expected value of Y given $\mathbf{x}$, while $\theta_{g2}(\mathbf{x})$ is the component dispersion parameter, $g = 1, \ldots, G$. Wang et al. (2014) provide the conditions that guarantee identifiability for the model in Equation (5.4):

- the domain $\mathcal{X}$ of $\mathbf{x}$ is an open set of $\mathcal{R}^{J+1}$;
- each weight $\pi_g(\mathbf{x}) > 0$ is a continuous function, and parameters $\theta_{g1}(\mathbf{x})$ and $\theta_{g2}(\mathbf{x})$ have continuous first derivative, for $g = 1, \ldots, G$;

- for any $\mathbf{x}$, and $1 \leq l \neq g \leq G$,

$$\sum_{k=0}^{1} ||\theta_{l1}^{(k)}(\mathbf{x}) - \theta_{g1}^{(k)}(\mathbf{x})||^2 + \sum_{k=0}^{1} ||\theta_{l2}^{(k)}(\mathbf{x}) - \theta_{g2}^{(k)}(\mathbf{x})||^2, \qquad (5.5)$$

  where $\theta_{g1}^{(k)}$ and $\theta_{g2}^{(k)}$, $g = 1, \dots, G$ denote the k-th derivatives of the component parameter functions;

- the parametric mixture model $\sum_{g=1}^{G} \pi_g f(\theta_{g1}, \theta_{g2})$ is identifiable.

This theorem can be considered valid also for to model (5.1) by taking into account that variance σ_g^2 is assumed independent – and, thus, constant – with respect to covariates $\mathbf{x}_j$, $j = 1, \dots, J$.

5.3 Bayesian inference

P-splines are used to approximate the smooth effects in Equation (5.2):

$$s_{gj}^{(\mu)}(x_{ij}) = \sum_{\rho=1}^{m} \mathbf{B}_{j\rho}(x_{ij}) \boldsymbol{\beta}_{gj}^{(\mu)}. \qquad (5.6)$$

As in Section 2.2.2, a first order random walk prior is assigned to the cofficients $\boldsymbol{\beta}_{gj}^{(\mu)}$:

$$\beta_{gj\rho}^{(\mu)} = \beta_{gj,\rho-1}^{(\mu)} + u_{gj\rho}^{(\mu)}, \quad u_{gj\rho}^{(\mu)} \sim N(0, \tau_{gj}^{2\ (\mu)}) \quad \rho = 1, \dots, m \qquad (5.7)$$

or, equivalently,

$$\boldsymbol{\beta}_{gj}^{(\mu)} | \tau_{gj}^{2\ (\mu)} \propto \exp\left(-\frac{1}{2\tau_{gj}^{2\ (\mu)}} \boldsymbol{\beta}_{gj}^{(\mu)'} \mathbf{K}_j \boldsymbol{\beta}_{gj}^{(\mu)} \right), \quad j = J^*, \dots, J;\ g = 1, \dots, G, \qquad (5.8)$$

with penalty matrix $\mathbf{K}_j$ defined as in Equation (2.8). Coherently with Section 2.2.2, for $g = 1, \dots, G$, it is assumed that:

- $\tau_{gj}^{2\ (\mu)} \sim IG(a_j, b_j)$, $j = J^* + 1, \dots, J$;

- $\boldsymbol{\gamma}_g^{(\mu)} \sim MVN\left(\mathbf{0}, v\mathbf{I}_{J^*+1}\right)$;

- $\sigma_g^{2(\mu)} \sim IG(a_\sigma, b_\sigma)$.

5.4 The MCMC algorithm

Throughout this Section, the superscript $^{(g)}$ is applied 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. Step 7 of the MCMC algorithm in Section 2.2.3 can be modified by sampling the component parameters $(\mu_g(\mathbf{x}_i), \sigma_g^2)$, $g = 1, \dots, G$ conditional on the component indicator $\mathbf{D}_i, \dots, \mathbf{D}_n$ as follows:

- sample the regression coefficients $\boldsymbol{\beta}_j$, $j = J^* + 1, \dots, J$, from a multivariate normal density with covariance matrix $\mathbf{V}_{gj}^{(\mu)}$ and mean $\mathbf{m}_{gj}^{(\mu)}$

  $$\mathbf{V}_{gj}^{(\mu)} = \left(\frac{1}{\sigma_g^2} \mathbf{B}_j^{(g)\prime} \mathbf{B}_j^{(g)} + \frac{1}{\tau_{gj}^{2\ (\mu)}} \mathbf{K}_j \right)^{-1}, \quad \mathbf{m}_{gj}^{(\mu)} = \mathbf{V}_{gj}^{(\mu)} \mathbf{B}_j^{(g)\prime} \left(\mathbf{y}^{(g)} - \tilde{\boldsymbol{\eta}}_{g,-j}^{(\mu)} \right), \tag{5.9}$$

  where $\tilde{\boldsymbol{\eta}}_{g,-j}^{(\mu)}$ is the part of the predictor associated with all effects in the model, but the j-th;

- center each smooth function $s_{gj}^{(\mu)}(x_j)$. Imposing the constraint in Equation (5.3) is equivalent to sampling $\boldsymbol{\beta}_{gj}^{(\mu)} | (\mathbf{1}_n^{\prime(g)} \mathbf{B}_j^{(g)} \boldsymbol{\beta}_{gj}^{(\mu)} = \mathbf{0})$, $j = J^* + 1, \dots, J$. As in (2.20), this can be done by trasforming each vector of coefficients $\boldsymbol{\beta}_{gj}^{(\mu)}$ as follows:

  $$\tilde{\boldsymbol{\beta}}_{gj}^{(\mu)} = \boldsymbol{\beta}_{gj}^{(\mu)} - \mathbf{V}_{gj}^{(\mu)} \mathbf{B}_j^{(g)\prime} \mathbf{1}_n^{(g)\prime} \left(\mathbf{1}_n^{(g)\prime} \mathbf{B}_j^{(g)} \mathbf{V}_{gj}^{(\mu)} \mathbf{B}_j^{(g)\prime} \mathbf{1}_n \right)^{-1} \mathbf{1}_n^{(g)\prime} \mathbf{B}_j^{(g)} \boldsymbol{\beta}_{gj}^{(\mu)}; \tag{5.10}$$

- sample the fixed effects parameters $\boldsymbol{\gamma}_g^{(\mu)}$ from a multivariate normal distribution with covariance matrix $\mathbf{V}_{\boldsymbol{\gamma}_g}^{(\mu)}$ and the mean $\mathbf{m}_{\boldsymbol{\gamma}_g}^{(\mu)}$ obtained as

  $$\mathbf{V}_{\boldsymbol{\gamma}_g}^{(\mu)} = \left(\frac{1}{\sigma_g^2} \mathbf{X}^{(g)\prime} \mathbf{X}^{(g)} + \frac{1}{v} \mathbf{I}_{J^*+1} \right)^{-1}, \quad \mathbf{m}_{\boldsymbol{\gamma}_g}^{(\mu)} = \mathbf{V}_{\boldsymbol{\gamma}_g}^{(\mu)} \mathbf{X}^{(g)\prime} \left(\mathbf{y}^{(g)} - \tilde{\boldsymbol{\eta}}_{g,-\gamma}^{(\mu)} \right). \tag{5.11}$$

  Here, $\tilde{\boldsymbol{\eta}}_{g,-\gamma}^{(\mu)}$ represents the nonlinear part of the g-th predictor;

- sample the parameter $\tau_{gj}^{2\ (\mu)}$ conditional on $\boldsymbol{\beta}_{gj}^{(\mu)}$:

  $$\tau_{gj}^{2\ (\mu)} | \boldsymbol{\beta}_{gj}^{(\mu)} \sim IG\left(a_j + \frac{\operatorname{rank}(\mathbf{K}_j)}{2}, b_j + \frac{1}{2} \boldsymbol{\beta}_{gj}^{(\mu)\prime} \mathbf{K}_j \boldsymbol{\beta}_{gj}^{(\mu)} \right) \tag{5.12}$$

- compute $\mu_g(\mathbf{x})$ as

$$\mu_g(\mathbf{x}) = \mathbf{X}\boldsymbol{\gamma}_g^{(\mu)} + \sum_{j=J^*+1}^{J} \mathbf{B}_j \boldsymbol{\beta}_{gj}^{(\mu)}; \tag{5.13}$$

- sample the variance parameter σ_g^2 conditional on $\mu_g^{(g)}(\mathbf{x})$:

$$\begin{aligned} \sigma_g^2|\mu_g^{(g)}(\mathbf{x}), \sim IG\Big(& a_\sigma^{(\mu)} + \frac{\sum_{i=1}^n D_{gi}}{2}, \\ & b_\sigma^{(\mu)} + \frac{1}{2}\left(\mathbf{y}^{(g)} - \mu_g^{(g)}(\mathbf{x})\right)' \left(\mathbf{y}^{(g)} - \mu_g^{(g)}(\mathbf{x})\right)\Big). \end{aligned} \tag{5.14}$$

5.5 Simulation study

The performance of the proposed approach is investigated in a simulated environment. In particular, two scenarios are considered, differing for the true number of components and the distribution of the manifest variables. 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 are evaluated through a comparison between the true effects and the estimated posterior effects, after fitting each of the following mixture of regression models:

- semiparametric mixture of experts regression model (SMoERm), with flexible specification of both the mixture weights $\pi_g(x)$ and the conditional means $\mu_g(x)$, $g = 1, \ldots, G$;

- mixture of semiparametric regression model (MoSRm), with constant mixture weights π_g and flexible specification of the conditional means $\mu_g(x)$, $g = 1, \ldots, G$;

- parametric mixture of experts regression model (PMoERm), 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 $\mu_g(x)$, $g = 1, \ldots, G$;

- mixture of parametric regression model (MoPRm), with constant mixture weights π_g and linearity assumption for the effect of x on the conditional means $\mu_g(x)$, $g = 1, \ldots, G$;

with G 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 $\{x_i^*\}, i = 1, \dots, n$, are grid points taken evenly in the range of covariate x. To quantitatively assess the performance of the estimators of the unknown regression functions $\mu_g(x)$, their square root of the average squared errors (RASE) is compared:

$$\text{RASE}_{\mu_g} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\hat{\mu}_g(x_i^*) - \mu_g(x_i^*)\right)^2}, \quad g = 1, \dots, G. \tag{5.15}$$

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 semiparametric and the parametric mixture of experts regression model.

Regarding the clustering performance, a comparison between the true allocations and the estimated ones is made in terms of Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) and soft ARI (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. The optimal choice for the number of components is based on AICM, and the number of non-empty components $\tilde{G}$ is computed according to Equation (2.27). For each of the competing 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.

5.5.1 First simulation experiment: $G = 2$

A batch 100 independent datasets are generated with $n = 1000$ from a 2-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: $x_i \sim Unif(0, 1)$, $i = 1, \dots, 1000$. The functional form of $\eta_1(x)$, coupled with the specific range of values for x_i, leads to a non-monotonic concave log-odd. Conditional on x and the component indicators, each component density is a normal distribution, with means $\mu_1(x), \mu_2(x)$ and variances σ_1^2, σ_2^2, respectively. Two alternatives are considered for σ_2, leading to as many different levels of overlap between the groups:

$$\begin{aligned} \mu_1(x) &= 15(x - 0.5)^2 + 1, \quad \sigma_1 = 0.3; \\ \mu_2(x) &= 5(x - 0.5)^2, \quad \sigma_2 = 0.2, 0.25. \end{aligned}$$

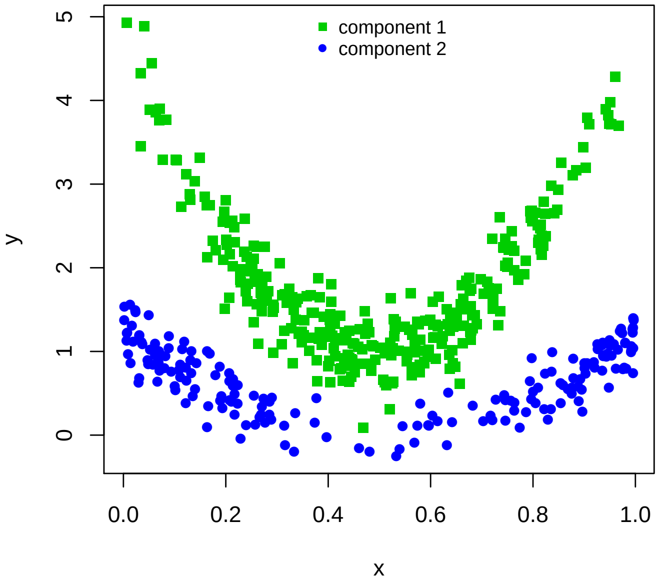


Figure 5.2: Example of a simulated dataset, with $G = 2$ and $\sigma_2 = 0.2$.

The simulation experiment starts with $\sigma_2 = 0.2$. Figure 5.2 shows one of the 100 independent samples with low overlap between the groups, which look quite separated. Figure 5.3 highlights the limits of the parametric approach when a non monotonic function, symmetric about x has to be approximated. In particular, for fixed number of components $G = 2$, the parametric mixture of experts regression models tends to fit a constant function with an average RASE_η equal to 6.921 (standard deviation = 11.439) over the 100 simulations. Conversely, the corresponding semiparametric approach seems to catch the underlying trend, even though some oversmoothing is present around the peak. For this model, the RASE_η drops to 0.173, with standard deviation of 0.488.

Regarding the estimated conditional means, the SMoER model shows good performances in Figure 5.4, 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 Figure 5.3. Thus, here, most of the observations comes 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

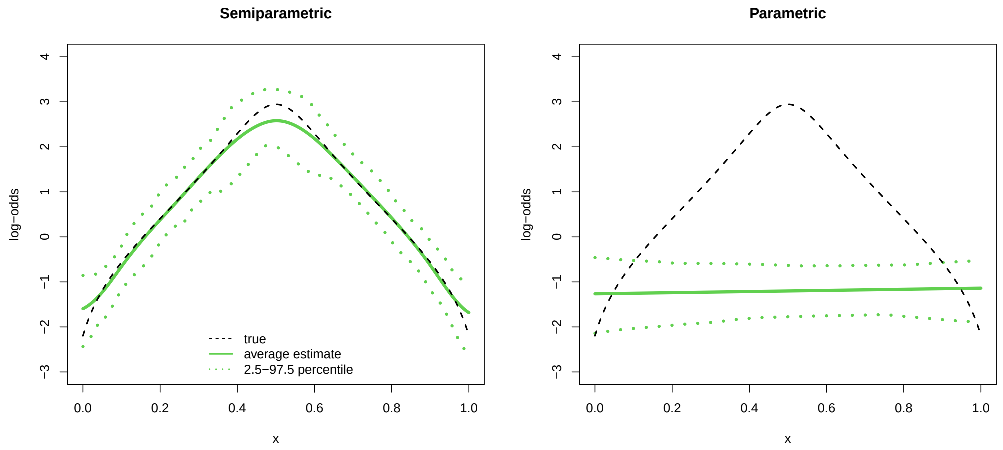


Figure 5.3: Pointwise average and 2.5 – 97.5 percentiles of the log-odds of the mixture weight η_1 estimated by both the mixture of experts regression models over 100 simulated datasets.

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 MoSR model; see Figure 5.5. Again, the main problem regards mostly the lower component, whose true mean is now barely included in the bands, even though they widen considerably in the overlap region.

No assumption of constant weights is made when fitting the PMoER model, but, as previously shown in Figure 5.3, this model estimates a constant effect of the covariate, making it practically equivalent to a MoPR model. This is evident in Figure 5.6, where the conditional means estimated by the two models are compared. Since these two functions are generated to be quadratic and symmetric about x, both parametric models fit horizontal lines, effectively collapsing to a simple mixture of normals not involving the effect of the covariate for the marginal distribution of the dependent variable. Moreover, because of the non-capability to detect the underlying unit-specific mixture weights and, consequently, the true class membership through these approaches, the fitted constant means are not even centered around the true average group means. Table 5.1 summarizes this comparison among the conditional mean functions estimated by the four models from a quantitative

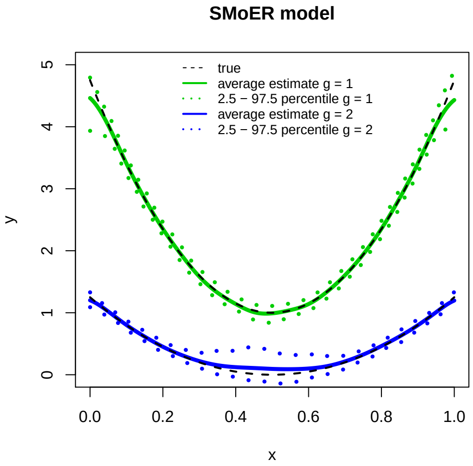


Figure 5.4: Pointwise average and 2.5 – 97.5 percentiles of the conditional means estimated by the semiparametric mixture of experts regressions model over 100 simulated datasets.

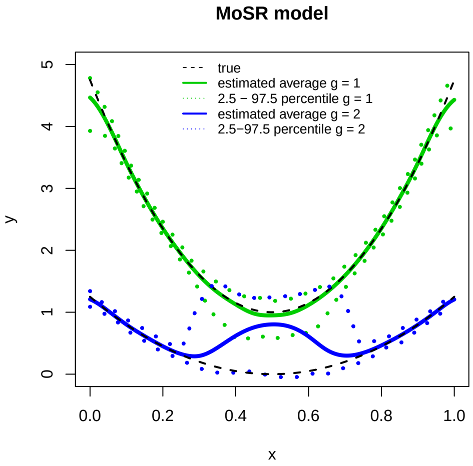


Figure 5.5: Pointwise average and 2.5 – 97.5 percentiles of the conditional means estimated by the mixture of semiparametric regressions model over 100 simulated datasets.

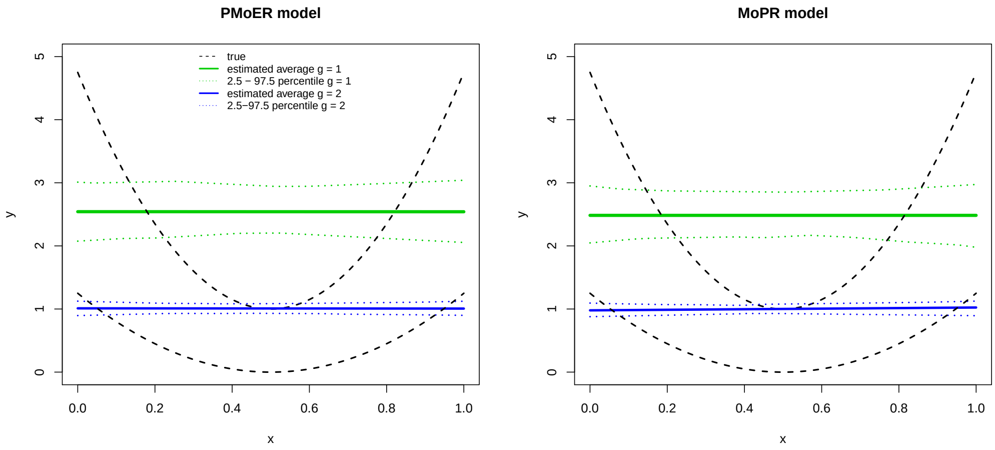


Figure 5.6: Pointwise average and 2.5 – 97.5 percentiles of the conditional means estimated with both parametric approaches over 100 simulated datasets.

Table 5.1: Mean (and standard deviation) of the RASE scores computed on the estimated conditional means over 100 simulated datasets.

RASE$_{\mu_g}$	$\mu_1(x)$	$\mu_2(x)$
SMoERm	0.100 (0.058)	0.080 (0.080)
MoSRm	0.131 (0.063)	0.391 (0.188)
PMoERm	1.187 (0.095)	0.700 (0.049)
MoPRm	1.167 (0.061)	0.700 (0.105)

point of view, by displaying, for each combination of method and component $g = 1, 2$, the average RASE$_{\mu_g}$ and the corresponding standard deviation over 100 simulated datasets. Quality of the estimates are strictly related to the quality of the allocations, as Table 5.2 confirms. The SMoER model, in fact, outperforms its competitors in terms of AICM, ARI and sARI for fixed number of components $G = 2$, followed by the MoSR model. The parametric approaches prove to be not satisfactory in this simulation setting.

A comparison among the four competing models is performed also by examining the best models selected according to AICM when considering a number of components ranging from 1 to 4. Table 5.3 reports the distribution of the number of non-empty components $\tilde{G}$ selected. $\tilde{G} = 2$ is always the best choice, according to the SMoER model, while 7 times out of 100 the MoSR

Table 5.2: Average AICM, ARI and sARI (number of times each model ranks first) over 100 simulated datasets, for fixed $G = 2$.

	AICM	(best)	ARI	(best)	sARI	(best)
SMoERm	144.5	(94)	0.971	(96)	0.957	(98)
MoSRm	261.5	(6)	0.673	(4)	0.658	(2)
PMoERm	1487.9	(0)	-0.004	(0)	0.008	(0)
MoPRm	1553.1	(0)	0.001	(0)	0.001	(0)

Table 5.3: Number of non-empty component selected for each method, according to AICM.

	$\tilde{G} = 1$	$\tilde{G} = 2$	$\tilde{G} = 3$	$\tilde{G} = 4$
SMoERm	-	100	-	-
MoSRm	-	93	7	-
PMoERm	85	15	-	-
MoPRm	93	5	2	-

model provides a better AICM with an additional component. Conversely, the parametric approaches tend to perform better with a single component.

By comparing the best models (according to AICM) fitted with each method for each simulation, rather than fixing the number of components, the results do not change much. All the AICMs reported in Table 5.2 improve, also for the SMoER, because sometimes adding an extra component, even if it is emptied during the posterior allocation, slightly decreases the AICM. For the same reason, both the average ARI and sARI appear to slightly improve for the SMoER model, while it worsen for models that tend to pick the wrong number of components; see Table 5.4.

For $\sigma_2 = 0.25$ the overlap between the two components increases, involving a higher number of units; Figure 5.7 shows one of the 100 independent samples with this setting. This does not seem to affect much the quality of the estimated unit specific log-odds of the mixture weights via semiparametric approach, apart from a slighly increased oversmoothing; see Figure 5.8). In particular, the average RASE_η increases to 0.209, with a standard deviation of 0.665. The same conclusion can be drawn, from the first plot in Figure 5.9, about the average conditional means estimated by the SMoER model. In the second plot, the effects of the augmented overlap on the es-

Table 5.4: 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)
SMoERm	140.0	(94)	0.983	(97)	0.970	(95)
MoSRm	252.53	(6)	0.671	(3)	0.651	(5)
PMoERm	1337.9	(0)	-0.003	(0)	-0.002	(0)
MoPRm	1341.0	(0)	0.001	(0)	-0.000	(0)

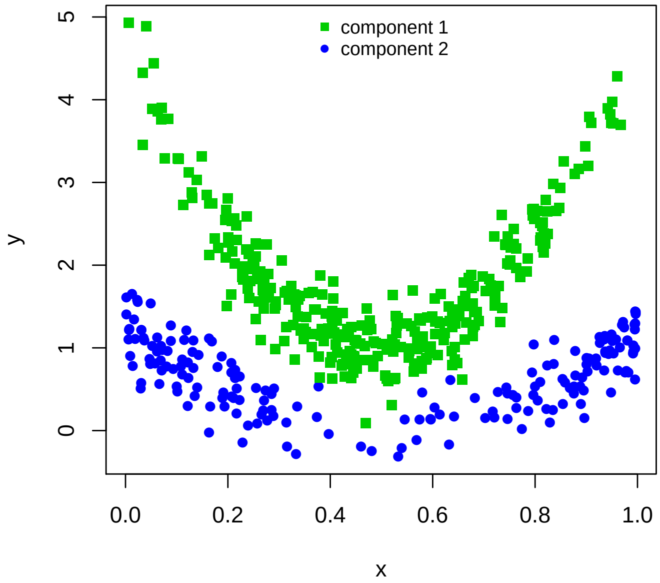


Figure 5.7: Example of a simulated dataset, with $G = 2$ and $\sigma_2 = 0.25$.

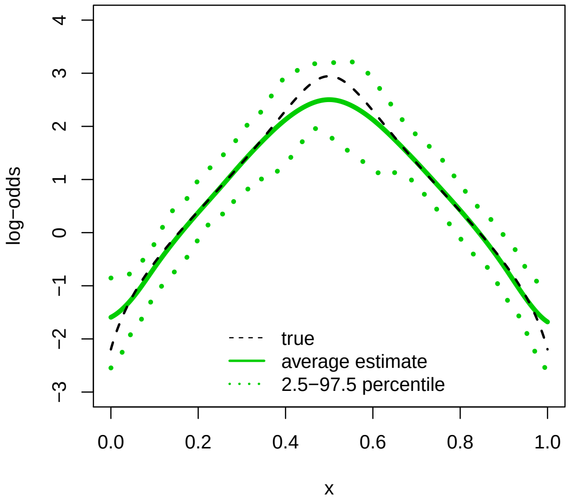


Figure 5.8: Pointwise average and 2.5 – 97.5 percentiles of the log-odds of the mixture weight η_1 estimated by the semiparametric mixture of experts regression model over 100 simulated datasets.

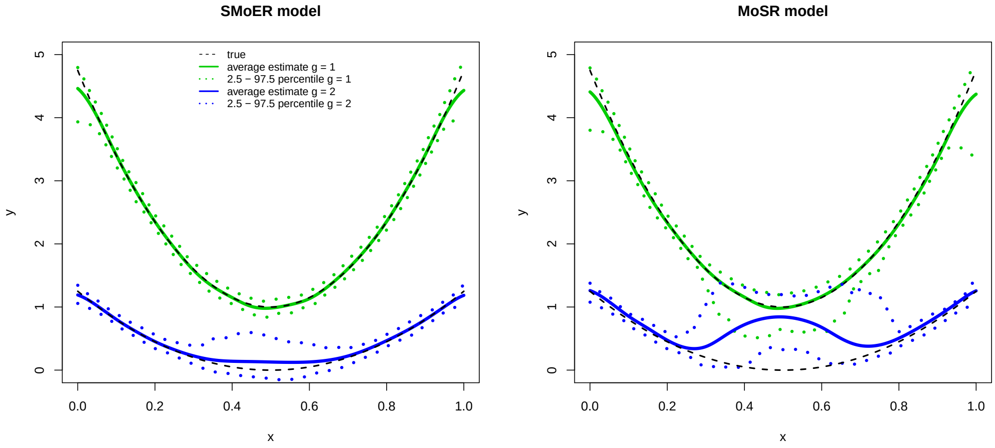


Figure 5.9: Pointwise average and 2.5 – 97.5 percentiles of the conditional means estimated with both semiparametric approaches over 100 simulated datasets.

timates produced by the MoSR model are more evident: in particular, the estimates of the lower component seem to be attracted by the upper component, leading to an increased bias around central values of x, if compared to the first setting, and the bands do not contain the true function in this area. These results are reflected in the RASE scores reported in Table 5.5. The reduced quality in the estimates affects the goodness of fit and the accuracy of the allocations, for fixed $G = 2$. However, the SMoER model still produces satisfying results and outperforms the competitors. Details are provided in Table 5.6.

Finally, when considering the best models selected according to the AICM,

Table 5.5: Mean (and standard deviation) of the RASE scores computed on the estimated conditional means over 100 simulated datasets with $\sigma_2 = 0.25$.

RASE$_{\mu_g}$	$\mu_1(x)$	$\mu_2(x)$
SMoERm	0.102 (0.066)	0.099 (0.086)
MoSRm	0.162 (0.232)	0.450 (0.237)
PMoERm	1.222 (0.199)	0.716 (0.062)
MoPRm	1.179 (0.078)	0.719 (0.133)

Table 5.6: Average AICM, ARI and sARI (number of times each model ranks first) over 100 simulated datasets, for $\sigma_2 = 0.25$ and fixed $G = 2$.

	AICM	(best)	ARI	(best)	sARI	(best)
SMoERm	232.0	(99)	0.960	(98)	0.940	(98)
MoSRm	453.9	(1)	0.627	(2)	0.580	(2)
PMoERm	1521.9	(0)	-0.012	(0)	-0.000	(0)
MoPRm	1607.8	(0)	-0.007	(0)	0.006	(0)

Table 5.7: Number of non-empty groups selected for each method, according to AICM, with $\sigma_2 = 0.25$.

	$\tilde{G} = 1$	$\tilde{G} = 2$	$\tilde{G} = 3$	$\tilde{G} = 4$
SMoERm	-	100	-	-
MoSRm	-	74	26	-
PMoERm	90	10	-	-
MoPRm	93	2	5	-

with G ranging from 1 to 4, all the competing models tend to be more prone to errors in terms of the number of non-empty components, apart from the SMoER model; see Table 5.7. This creates some further separation, with respect to the competitors, about the quality of the allocation, as shown by the average values of ARI and sARI reported in Table 5.8.

Table 5.8: Average AICM, ARI and sARI (counting how many times each model prevails) over 100 simulations, for $\sigma_2 = 0.25$ and optimal G, according to AICM.

	AICM	(best)	ARI	(best)	sARI	(best)
SMoERm	228.4	(100)	0.975	(100)	0.955	(100)
MoSRm	430.7	(0)	0.587	(3)	0.533	(5)
PMoERm	1342.0	(0)	-0.003	(0)	-0.002	(0)
MoPRm	1344.17	(0)	-0.002	(0)	-0.001	(0)

5.5.2 Second simulation expertiment: $G = 3$

A batch of 100 independent datasets is generated with $n = 1000$ from a 3-component mixture of regression models, with log-odds of mixture weights $\eta_g = \log \pi_g(x)/\pi_3(x)$, $g = 1, 2$, defined as:

$$\eta_1(x) = 2 \sin \, \pi(x + 0.5),$$
$$\eta_2(x) = 2 \sin \, \pi(x + 1.5),$$

where x is the only covariate, sampled from a uniform distribution: $x_i \sim Unif(0, 1)$, $i = 1, \ldots, 1000$. In this second experiment, the functional forms of $\eta_1(x)$ and $\eta_2(x)$, coupled with the range of x, lead to log-odds that are monotonically decreasing and increasing, respectively. Since these effects are quite smooth, a second scenario is also considered:

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

Conditional on x and the component indicators, y follows a univariate normal distribution, with means $\mu_1(x), \mu_2(x), \mu_3(x)$, and variances $\sigma_1^2, \sigma_2^2, \sigma_3^2$, 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 5.16 shows one of the 100 independently generated samples for this first scenario.

Figure 5.11 shows that the semiparametric mixture of experts regression is able to catch almost perfectly the effects of the covariate x on both predictors η_1 and η_2. Since nonlinearity is not much evident, the linear approximation made by the parametric mixture of experts regression model fits quite well: in particular, the true effects never exceed the bands. Because of the flexibility, the bands associated to the effects estimated by the SMoER model are wider, and this affects the performances in terms of RASE, as reported in Table 5.9.

Regarding the estimates of the conditional means, the SMoER model seems to outperform the competitors with quite appropriate fitting, despite some overlap present between Cluster 1 and Cluster 3 for lower values of x and between Cluster 2 and Cluster 3 for higher values of x; see Figure 5.12.

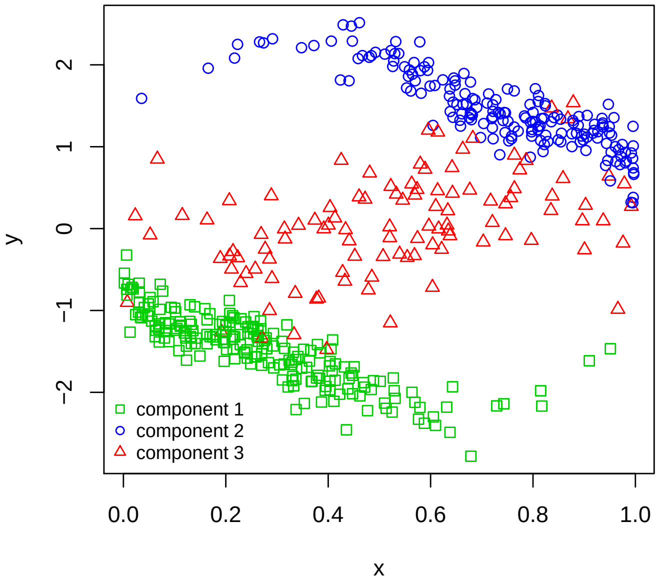


Figure 5.10: Example of a simulated dataset with smooth effect of the covariate x on the log-odds of the mixture weights.

Table 5.9: Mean (and standard deviation) of the RASE scores computed on the estimated log-odds of the mixture weights over 100 simulated datasets.

RASE_{η_g}	$\eta_1(x)$	$\eta_2(x)$
Semiparametric MoERm	0.451 (0.258)	0.447 (0.262)
Parametric MoERm	0.439 (0.207)	0.401 (0.175)

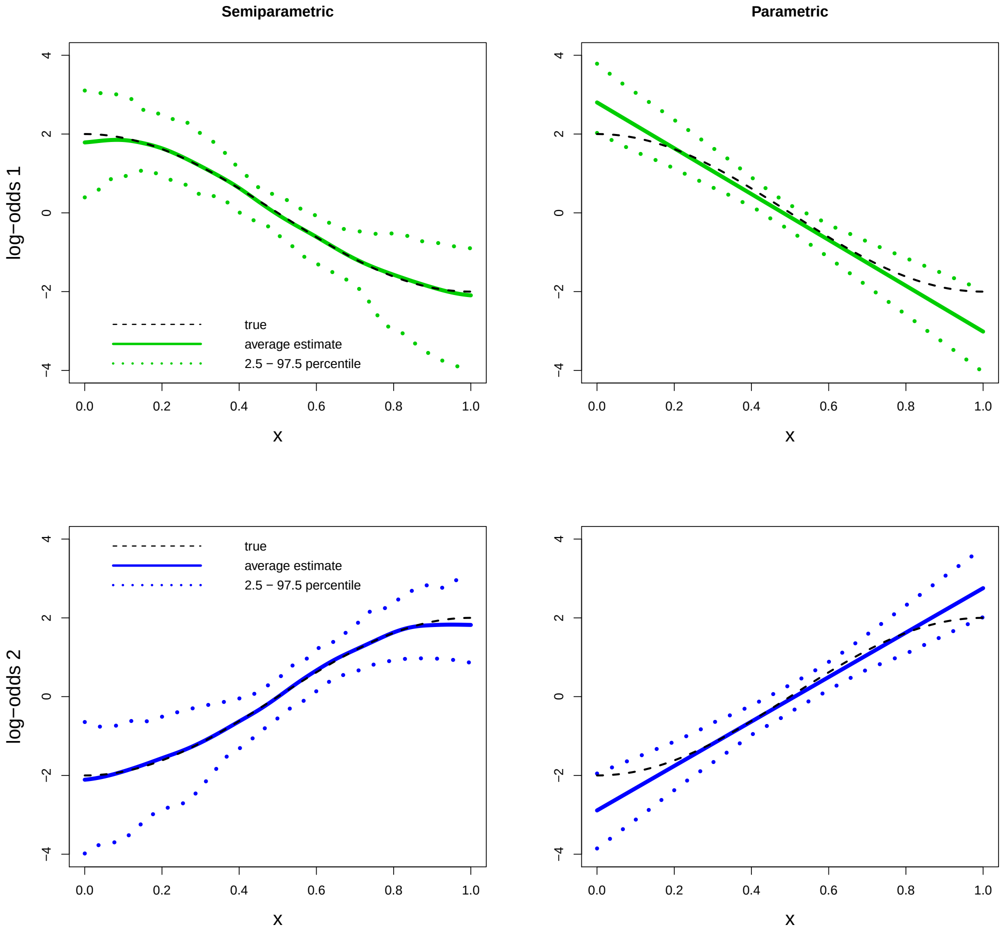


Figure 5.11: Comparison between the log-odds of the mixture weights estimated by the semiparametric (left) and parametric (right) full MoE models over 100 simulated datasets.

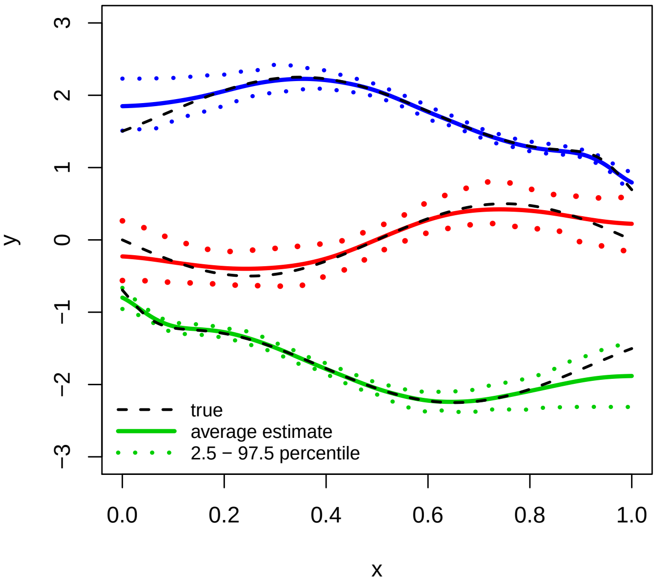


Figure 5.12: Conditional means estimated by the semiparametric mixture of experts regression model over 100 simulations.

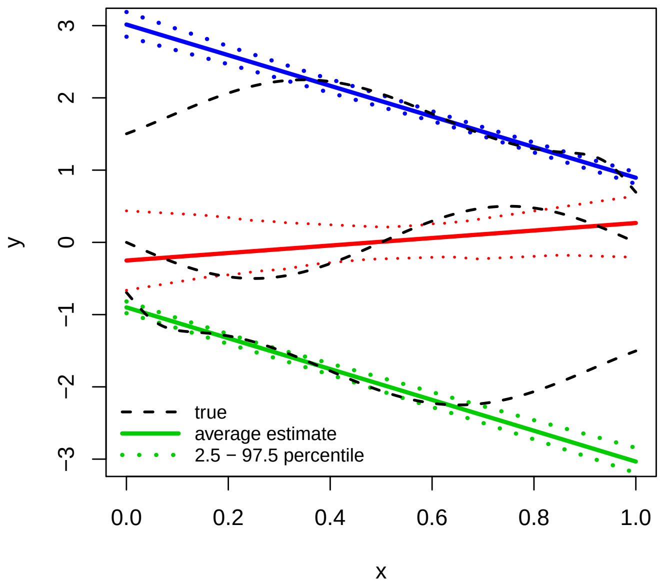


Figure 5.13: Conditional means estimated by the parametric mixture of experts regression model over 100 simulated datasets.

The PMoER model is unable to properly approximate the nonlinear trends, especially in the low probability regions, where the observations are sparse (i.e. in Cluster 1 for high values of x, or in Cluster 2 for low values of x). Nevertheless, thanks to the good estimates of the mixture weights, Figure 5.13 shows that the PMoER model discriminates almost correctly among groups in the aforementioned overlapping areas.

The mixture of semiparametric regression model needed Figure 5.14 to be decomposed into 3 different plots because of its bad performance. In particular, the flexibility allowed for the estimates of the conditional means, combined together with the impossibility to include the effect of covariate x into the estimates of the mixture weights, results into overlapping estimated functions and wide bands. Here, the performance of the mixture of parametric regression model is just slightly worse with respect to the ones obtained by the parametric mixture of experts regression model. The main differences can be observed in the overlapping regions of Figure 5.15, where the estimated conditional means intersect each other.

All the conclusions drawn from a graphical point of view are confirmed by

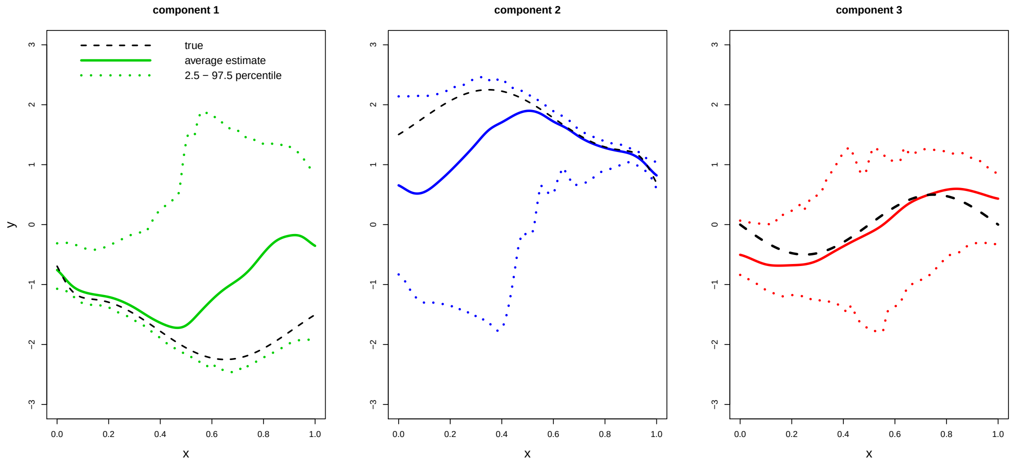


Figure 5.14: Conditional means estimated by the mixture of semiparametric regression model over 100 simulated datasets.

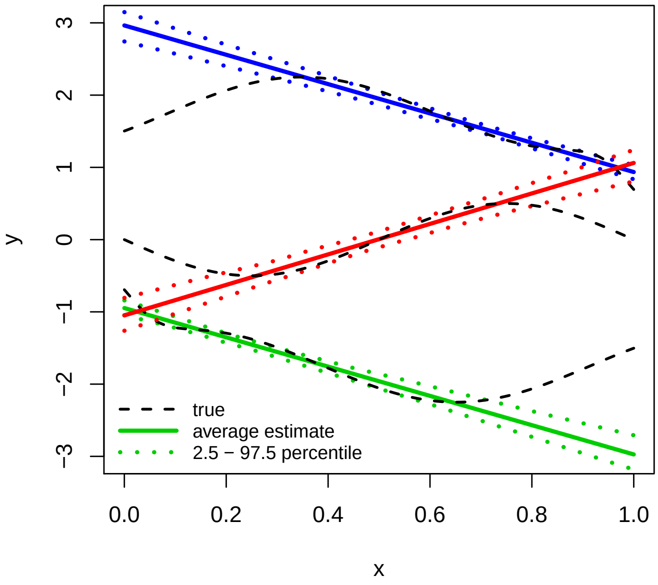


Figure 5.15: Conditional means estimated by the mixture of parametric regression model over 100 simulated datasets.

Table 5.10: Mean (and standard deviation) of the RASE scores computed on the estimated conditional means over 100 simulated datasets.

RASE$_{\mu_g}$	$\mu_1(x)$	$\mu_2(x)$	$\mu_3(x)$
SMoERm	0.125 (0.080)	0.115 (0.072)	0.161 (0.057)
MoSRm	1.087 (0.981)	0.811 (0.846)	0.506 (0.227)
PMoERm	0.498 (0.038)	0.490 (0.036)	0.293 (0.086)
MoPRm	0.475 (0.051)	0.471 (0.043)	0.407 (0.049)

Table 5.11: Average AICM, ARI and sARI (number of times each method ranks first) over 100 simulated datasets, for fixed $G = 3$.

	AICM	(best)	ARI	(best)	sARI	(best)
SMoERm	234.4	(100)	0.912	(83)	0.854	(82)
MoSRm	703.72	(0)	0.388	(0)	0.301	(0)
PMoERm	1160.7	(0)	0.895	(17)	0.840	(18)
MoPRm	1162.1	(0)	0.798	(0)	0.551	(0)

the quantitative results in terms of RASE$_\mu$ reported in Table 5.10. Table 5.11 shows that, in terms of AICM, the SMoER model is evidently better than its competitors, for fixed $G = 3$. 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, both the mixture of experts regression models perform well, even though the semiparametric one slightly prevails.

Regarding model selection, for each method and each sample, the best model (according to AICM) is considered among different mixture models with $G = 1, \ldots, 5$. Table 5.12 shows that the SMoER model is the only one which is able to pick the correct number of non-empty groups. This leads to more favorable results for the SMoER model, if the ARI and sARI computed with reference to the best models selected by each method are compared; see Table 5.13.

Figure 5.16 shows one of the 100 simulated datasets generated with more pronounced nonlinear effect of the covariate x on the predictors $\eta_1(x)$ and $\eta_2(x)$. According to Figure 5.17, the SMoER model is still able to catch almost perfectly these effects On the contrary, because of the increased nonlinearity, the linear approximation by the PMoER model is worse, so that the true effects exceed the bands at the boundaries of the range of x. In this case, the average RASE scores for the SMoER model (together with the

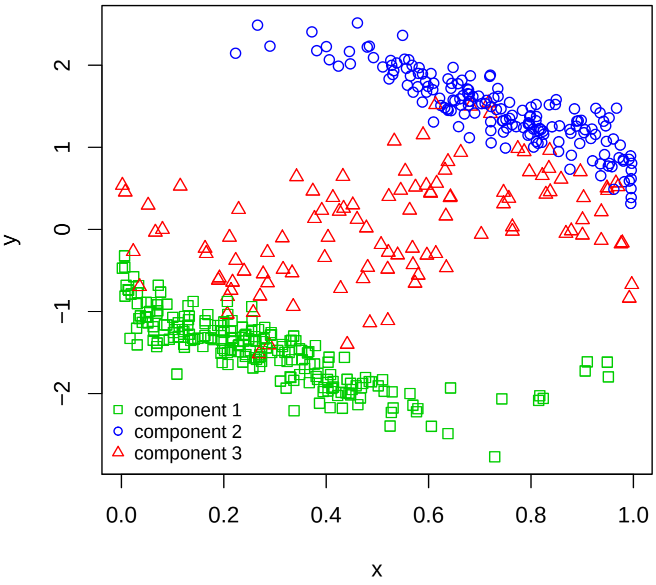


Figure 5.16: Example of a simulated dataset with sharp effect of the covariate x on the log-odds of the mixture weights.

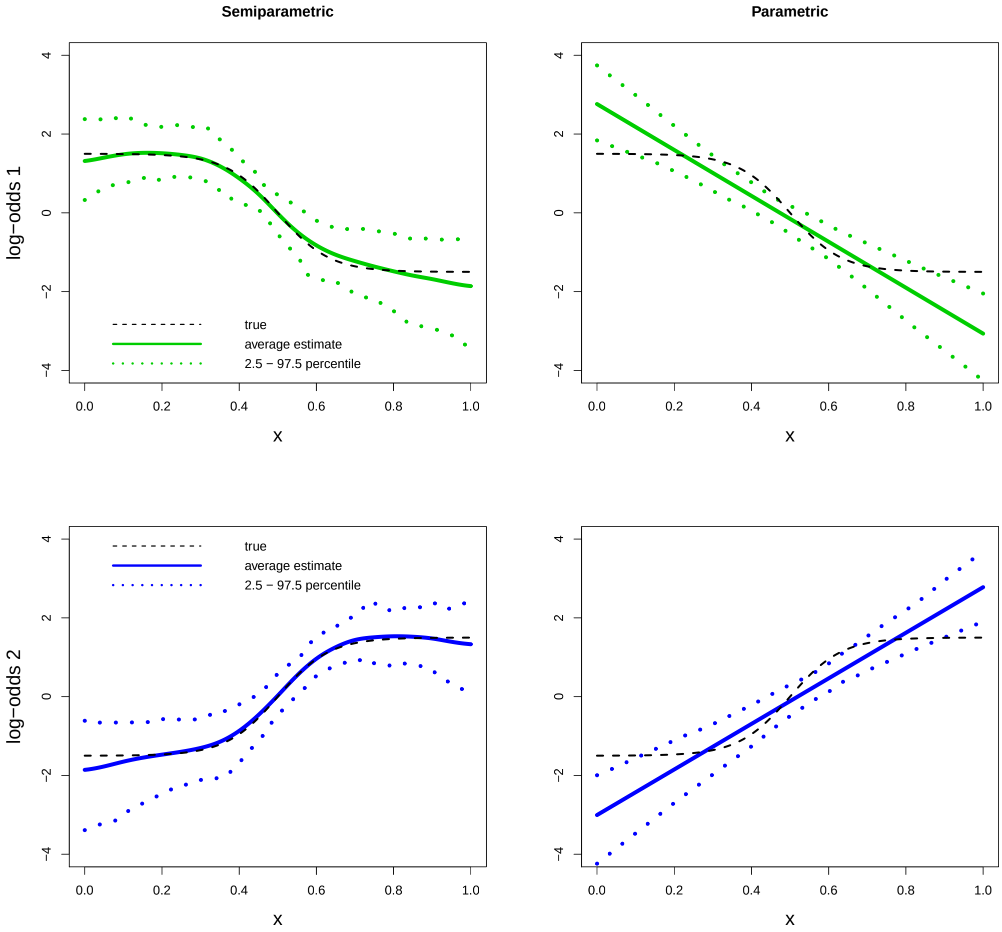


Figure 5.17: Comparison between the log-odds of the mixture weights estimated by the semiparametric (left) and parametric (right) full MoE models over 100 simulated datasets.

Table 5.12: Number of non-empty components selected for each method, according to AICM.

	$\tilde{G}=1$	$\tilde{G}=2$	$\tilde{G}=3$	$\tilde{G}=4$	$\tilde{G}=5$
SMoERm	-	-	100	-	-
MoSRm	-	-	26	74	-
PMoERm	-	36	25	20	3
MoPRm	2	-	77	20	1

Table 5.13: Average AICM, ARI and sARI (number of times each method ranks first) over 100 simulated datasets, for optimal G.

	AICM	(best)	ARI	(best)	sARI	(best)
SMoERm	228.16	(100)	0.911	(84)	0.854	(86)
MoSRm	586.5	(0)	0.301	(0)	0.262	(0)
PMoERm	1065.6	(0)	0.810	(16)	0.775	(14)
MoPRm	1140.1	(0)	0.778	(0)	0.534	(0)

associated standard deviations) reported in Table 5.14 are clearly lower, than those of the competitors considered.

Regarding the estimates of the conditional means, all the remarks made for the first scenario are confirmed by Figures 5.18, to 5.21, and Table 5.15. As far as the allocations are concerned, the differences in terms of ARI and sARI are now more evident, even though the parametric approaches continue to perform decently; see Table 5.16. In terms of model selection, the results obtained (not reported here) are in line with those of the first scenario. For this simulation study, a comparison is made also in terms of execution time of the algorithms estimating the four competing models. Figure 5.22 shows through a box plot how the different computational complexity between the

Table 5.14: Mean (and standard deviation) of the RASE scores computed on the estimated log-odds of the mixture weights over 100 simulated datasets.

RASE_{η_g}	$\eta_1(x)$	$\eta_2(x)$
Semiparametric MoERm	0.421 (0.153)	0.431 (0.169)
Parametric MoERm	0.679 (0.210)	0.663 (0.213)

Table 5.15: Mean (and standard deviation) of the RASE scores computed on the estimated conditional means over 100 simulated datasets.

RASE_{μ_g}	$\mu_1(x)$	$\mu_2(x)$	$\mu_3(x)$
SMoERm	0.086 (0.035)	0.086 (0.038)	0.145 (0.034)
MoSRm	1.073 (1.078)	0.995 (1.041)	0.472 (0.188)
PMoERm	0.509 (0.042)	0.507 (0.048)	0.101 (0.472)
MoPRm	0.474 (0.123)	0.459 (0.069)	0.343 (0.164)

Table 5.16: Average AICM, ARI and sARI (number of time each method ranks first) over 100 simulated datasets, for fixed $G = 3$.

	AICM	(best)	ARI	(best)	sARI	(best)
SMoERm	252.8	(100)	0.906	(99)	0.845	(96)
MoSRm	755.0	(0)	0.326	(0)	0.260	(0)
PMoERm	1641.7	(0)	0.854	(1)	0.797	(4)
MoPRm	1463.8	(0)	0.804	(0)	0.568	(0)

four algorithms affect the time employed by each of them, for each of the 100 samples, with the current setting. The reported times refer to analyses performed using an IBM x3750 M4 server with 4 Intel Xeon E5-4620 processors with 8 cores and 128GB RAM. It is worth noting that all the four algorithms require less time to complete a higher number of prefixed iteration (8000 versus 5000) if compared to both the latent class models with covariates considered in the simulation study reported in Section 3.2.2. The introduction of covariate effects on the component density has a lower impact, in terms of time, with respect to the higher number of components ($G = 6$ versus $G = 3$) and increased dimensionality of the manifest variable itself ($Q = 12$ versus $Q = 1$). Furthermore, the impact of the increase in complexity due the use of Bayesian P-spline is evident, since the semiparametric approaches are slower than their parametric analogues. Nevertheless, the MoSR model is faster than the PMoER model, indicating that the inclusion of covariate effects on the component weights, although assuming linearity, has a higher cost than allowing for flexible specification of the conditional means as nonlinear functions of the only covariate considered in this setting. Similar conclusions can be drawn by comparing the execution times for the other experimental settings considered in this Chapter.

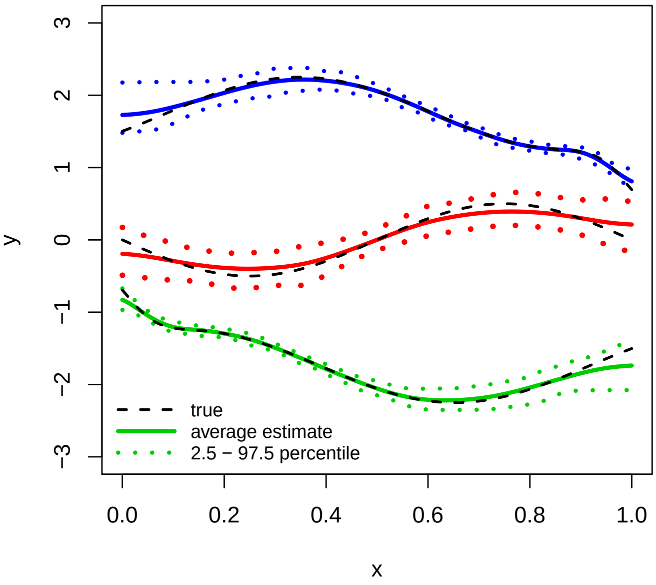


Figure 5.18: Conditional means estimated by the semiparametric mixture of experts regression model over 100 simulated datasets.

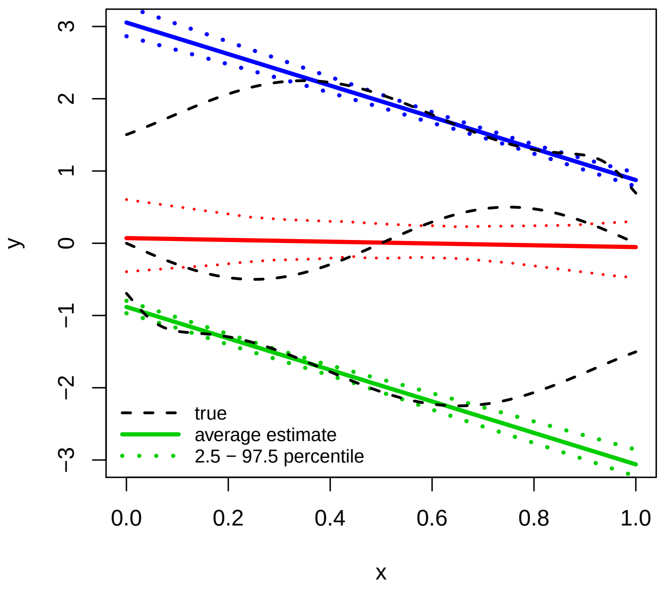


Figure 5.19: Conditional means estimated by the parametric mixture of experts regression model over 100 simulated datasets.

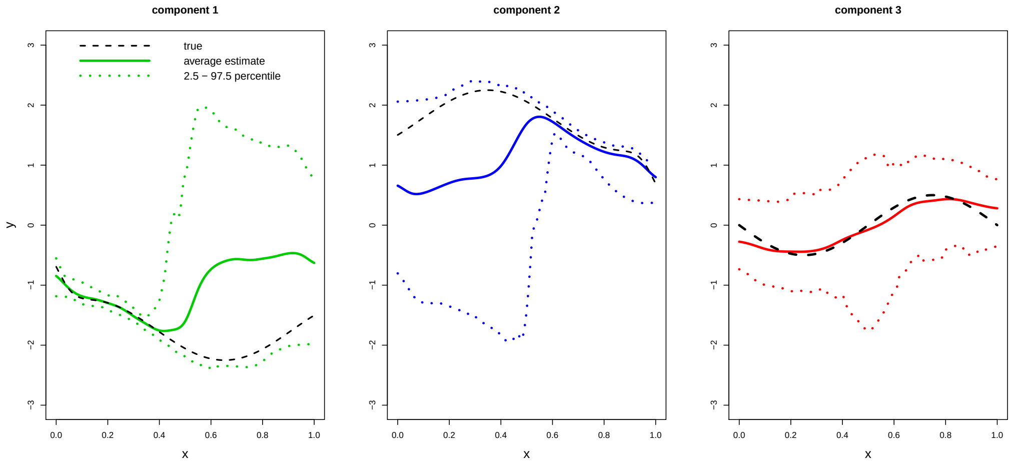


Figure 5.20: Conditional means estimated by the mixture of semiparametric regression model over 100 simulated datasets.

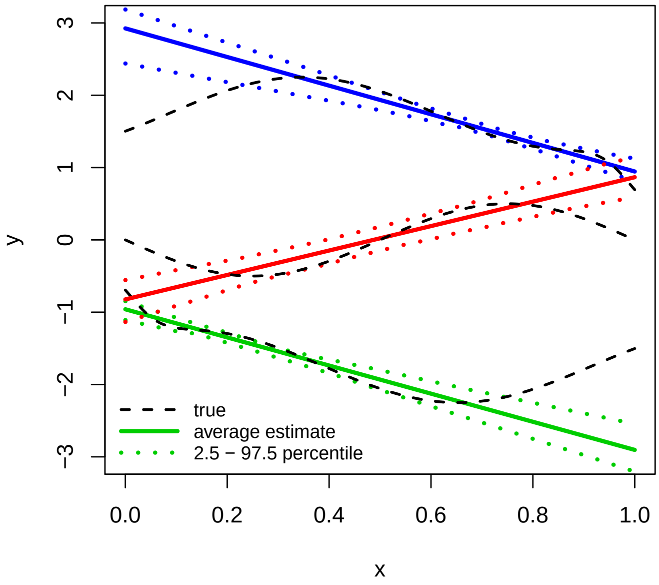


Figure 5.21: Conditional means estimated by the mixture of parametric regression model over 100 simulated datasets.

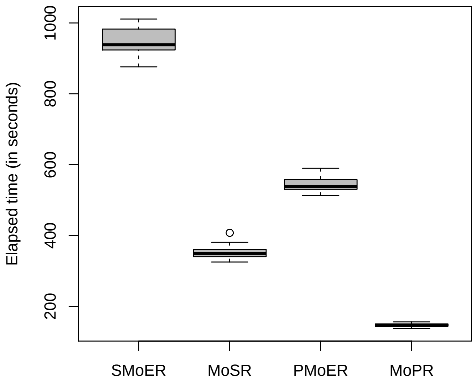


Figure 5.22: Time employed by each algorithm to complete 8000 iterations (initialization and posterior inference included), for each of the 100 replications, with fixed G=3.

5.6 Baseball salaries data

Watnik (1998) provides a dataset consisting of information about players for the 1992 Major League Baseball season. In particular their 1992 salaries are considered as the response, along with numerous measures of the 337 players' previous year's performances. Notice that this dataset is already well known in the mixture of experts literature; see, e.g., Khalili and Chen (2007) and Chamroukhi and Huynh (2018). For simplicity, the proposed analysis starts by focusing on one of the metrical 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 four different mixture of regression models considered in Section 5.5 for a fixed number of components ranging from 1 to 4. As suggested by Watnik (1998), due to asimmetry, the response is previously tranformed by taking the natural logarithm. The results are presented in Section 5.6.1.

For illustrative purpose, a second explanatory variable is then added to the analysis: the number of walks. This covariate is highly correlated (correlation 0.685) with the number of runs. However, it is worth noting that most of the metrical variables in the original dataset are strongly correlated with each other, or even function of some others. The results of this second analysis are presented in 5.6.2.

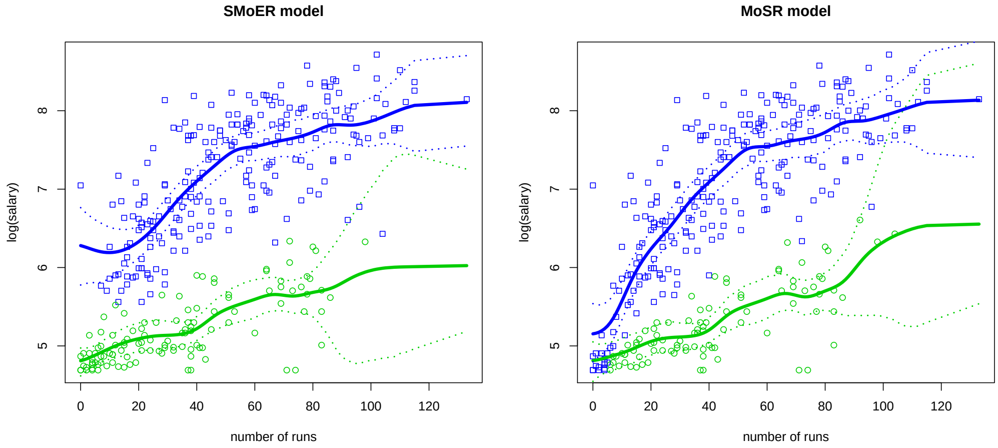


Figure 5.23: Estimated posterior conditional means (and pointwise 95% posterior credible bands) obtained from the SMoER Model (left panel) and the MoSR model (right panel).

5.6.1 Number of runs

After choosing the optimal value for G according to the AICM, the optimal number of non-empty components $\tilde{G}$ resulted to be equal to 2 for the two semiparametric models (semiparametric mixture of experts model and mixture of semiparametric regression model), and equal to 1 for the two parametric models (parametric mixture of experts regression model and mixture of parametric regression model). Among this four models, the SMoER model presents the absolute best AICM (663.4), followed by the MoSR model (733.7), while the remaining best parametric models, having $G = 1$, collapse to the same model, with the highest AICM (888.1).

As Figure 5.23 shows, the main difference between the semiparametric models seems to be related to the allocation of the players with a low number of runs. In particular the SMoER model keeps the two clusters well separated, by assigning all of these units to the lower one, while the MoSR model creates some overlap, such that the functions describing the conditional means, $\hat{\mu}_1(x)$ and $\hat{\mu}_2(x)$, almost interesect each other. Figure 5.23 confirms, 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.

The partition induced by the SMoER model identifies a cluster, the lower one (in green), which might be broadly interpreted as the cluster of "un-

Table 5.17: Comparison between the resulting allocations of the SMoER model and the free agency or arbitration elegibility.

	free agency or arbitration		
cluster	not eligible	eligible	
lower (green)	109	6	115
upper (blue)	29	193	222
	138	199	337

derrated" (or "underpaid", with respect to the others) baseball players. In fact, while it is obvious players with better performances get paid more, as is comfirmed by the increasing trends of both means, there seems to be a group of players whose salary is substantially lower than that of players with 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 Figure 5.23 appear almost parallel. A partial explanation of this result can be given thanks some additional information present in the dataset. In particular, there is a variable indicating the "free agency elegibility" of each player, i.e. if that player could have gone to a team of his choice in 1992. At the time – Watnik (1998) explain – 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 agent 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 free agents. For interpretation purpose, the two above described categories, "free agent eligible" and "arbitration eligible" players are merged, Table 5.17 compares the partition made by the SMoER model with the one obtained by distinguishing between (free agency or artbitration) eligible and non-eligible players. The resulting ARI (0.626) is the highest observed among the four 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 apparentely have been able to obtain an "adequate" salary without probing the market.

Fixing the number of components $G = 2$, the MoPR model allocates the players similarly to the SMoER model. In particular, only 9 units out of 337 (ARI = 0.894) are allocated differently. Focusing on the paramet-

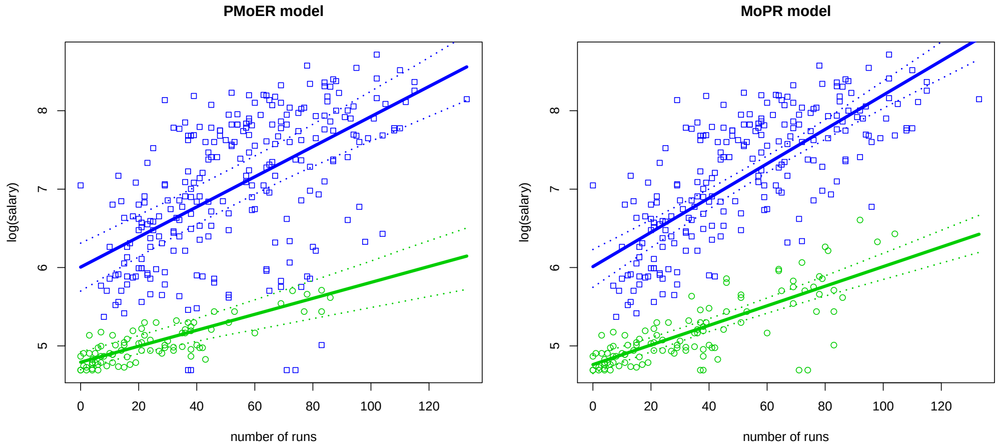


Figure 5.24: Estimated posterior conditional means (and pointwise 95% posterior credible bands) obtained from the PMoER Model (left panel) and the MoPR model (right panel).

ric approaches, the main difference between the two 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 PMoER model. This probably induces variability in the estimates of the latter, whose estimated mean functions present wider bands, if compared to the ones estimated by PMoER model; see Figure 5.24.

Both mixture of experts regression models 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 method estimates a nonlinear function that cannot be approximated properly by a straight line (Figure 5.23). The ability to pick this underlying effect is the main reason for the differences observed between the performances of the two semiparametric approaches.

5.6.2 Number of walks

Including the number of walks in the model does not seem to add much information with respect to the one already provided by the number of runs to this analysis. In terms of model selection, nothing changes with respect to the number of component picked by each method. However, the addition of a non significant covariate in the model causes more variability and affects the performance of the models, increasing the AICM for the semiparametric

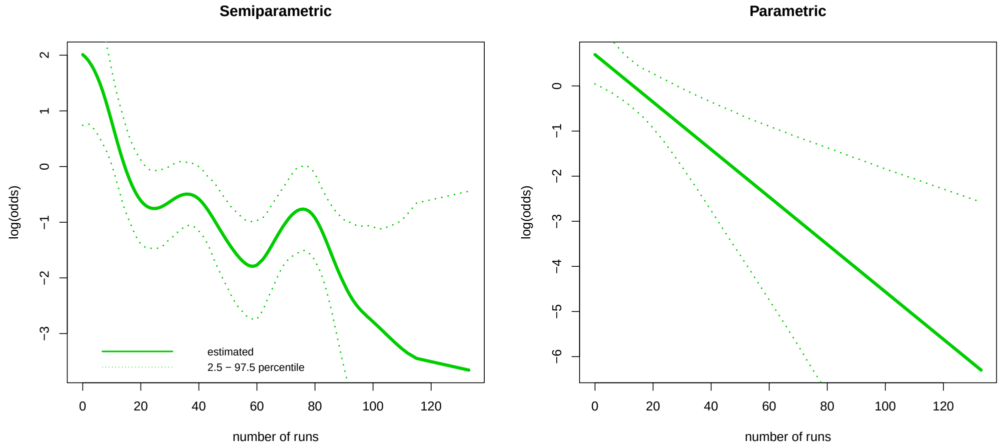


Figure 5.25: Estimated posterior effects on the log-odds (and pointwise 95% posterior credible bands) obtained from the SMoER model (left panel) and the PMoER model (right panel).

2-component mixture of experts regression model (724.8), which continues to prevail, even though the advantage with respect to the 2-components mixture of semiparametric regression model is close to zero (AICM = 725.8). The parametric approaches produce similar results (AICM = 889.1) with respect to ones obtained by considering the number of runs as the only explanatory varible.

Comparing the first plot in Figure 5.26 with the first one in Figure 5.23 leads to the conclusion that no clear changes happen, in terms of the partition obtained by the SMoER model; indeed, only 5 units are allocated differently (ARI = 0.941). Moreover, in Figure 5.26, it is easy to notice a similar structure of the clusters, conditional on each covariate.

The number of runs seems to explain most of the variability of both the conditional means $\mu_1(\mathbf{x})$ and $\mu_2(\mathbf{x})$ (Figure 5.27), and the log-odds of the mixture weight of the green component $\eta_1(\mathbf{x})$ (Figure 5.28). In fact, the bands of all the estimated effects of x_2 always include the constant function in zero, while the effects estimated for x_1 are very close to the ones observed in the previous analysis, without taking into account the number of walks. This is consistent with the fact that, according to the AICM, the model considering the number of runs as the only explanatory variable should be preferred to the model including both covariates.

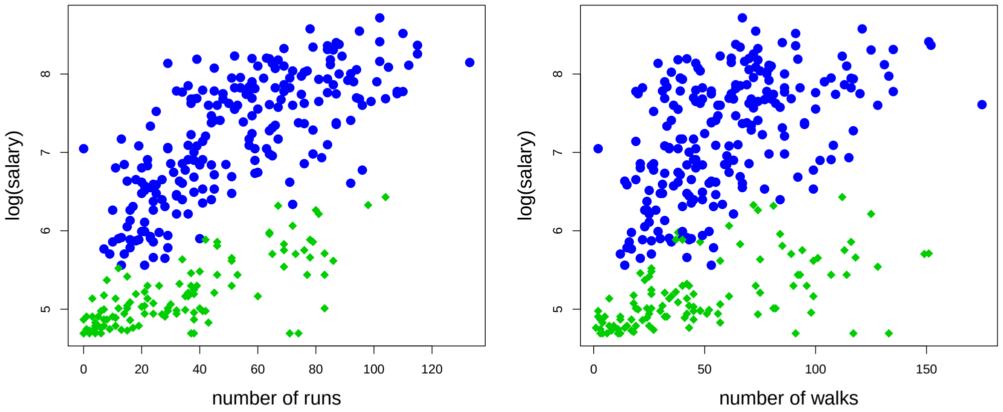


Figure 5.26: Allocations provided by the 2-component semiparametric mixture of experts regression model with respect to both covariates.

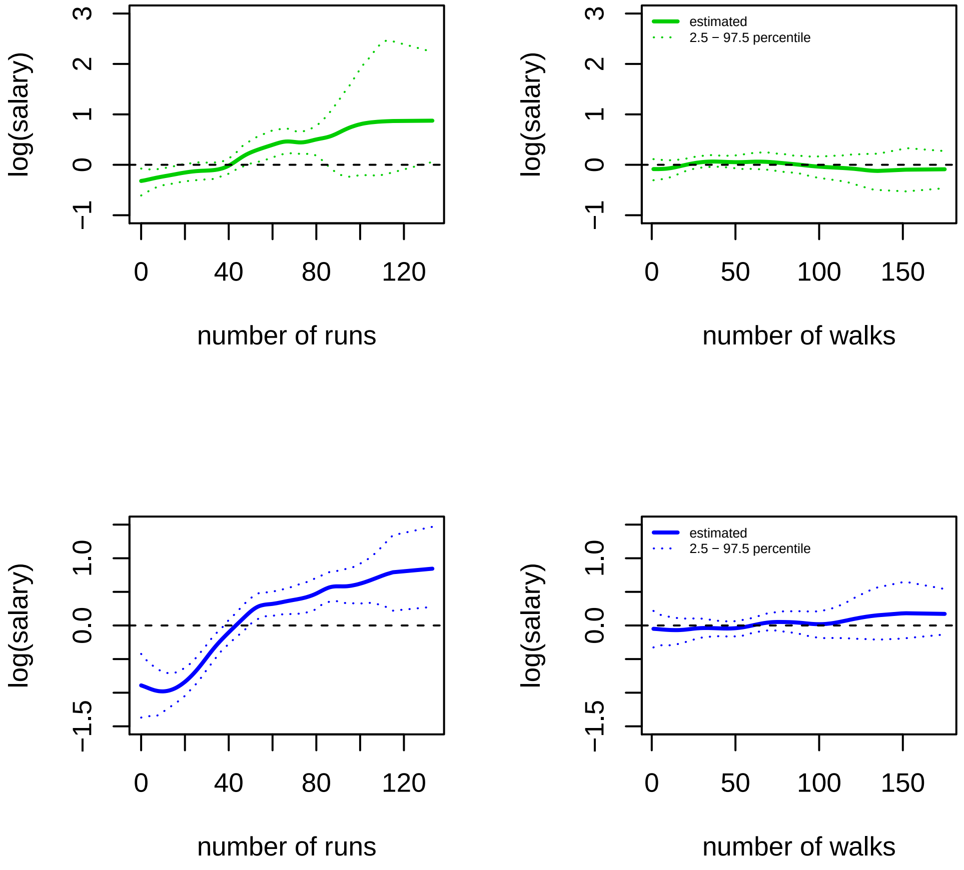


Figure 5.27: Estimated posterior conditional means (and pointwise 95% posterior credible bands) for the two components (one per row), obtained by the SMoER Model.

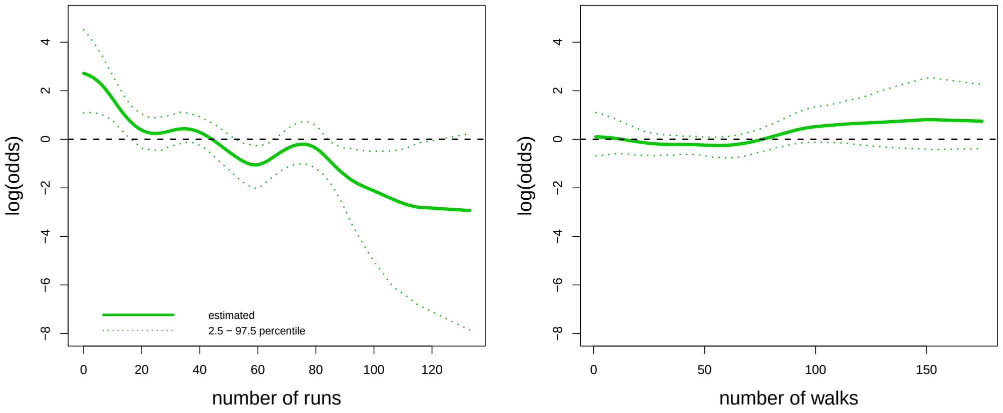


Figure 5.28: Estimated posterior effects on the log-odds (and pointwise 95% posterior credible bands) obtained by the SMoER model.

Chapter 6

Conclusions

In this Thesis, a general specification of a mixture model is proposed, allowing the weights to be nonlinear functions of some covariates. This general approach exploits an additive structure for the log-odds, and resort to spline functions for approximating the smooth effect of the concomitant variables. Parameter estimation is based on a formal Bayesian approach through MCMC machinery. Although a similar result, in principle, could be emulated also through a fully parametric approach, e.g. by considering a monomial set of bases to represent the map between component probabilities and covariates, resorting to a parametric representation, flexible enough to catch nonlinearity, would require some arbitrary choices, such as the maximum degree for monomial bases, or the definition of an automatic selection criterion. The approach advanced in this Thesis 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 proves to be a useful tool for recovering the underlying relation between component weights and concomitant variables – especially when it is 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 is illustrated also through applications to real data. The analysis of the determinants of cluster membership of the United Kingdom MPs shows that political party membership is not enough to explain their position towards Brexit, because both the safeness of seat (in terms of effective number of competing candidates at the previous general election) and the share of leave vote in the constituency they represent have a strong – sometimes nonlinear – effect. The study about soccer player position highlights the potentiality of the proposed methodology to deal with continuous outcome variable. In particular, the results reveals different patterns of interaction between pairs of players, suggesting specific playing strategies of the team. Finally, the methodology is extended to include the effects of the covariates not only on the component weights, but also on the component-specific distributions, focusing on mixture of regression models.

Although the results shown in this Thesis seem encouraging, there are some issues that might deserve further investigation. First of all, the focus is on estimation when manifest variables are categorical or continuous, but the proposed methodology can be adapted to any other type of response variables by choosing an appropriate form for the component density $f(\mathbf{y}|\boldsymbol{\theta}_g)$. Furthermore, the semiparametric latent class model presented in Section 3.1 is based on conditional independence, whereas different specifications for such a multivariate model with categorical margins might be possible, e.g. based on the so-called underlying random variables (URV) approaches (Ranalli and Rocci, 2017) in case of ordinal categorical manifest variables. Regarding the semiparametric mixture of normals in Section 4.1, problems related to the possible high dimensionality of the manifest variable are not addressed in this Thesis. It is worth mentioning that such problems are common to any mixture model based on multivariate Gaussian components and are not related to the introduction of covariate effects on the component weights. Remedies to this issue, such as the use of component-specific factor models to reduce the free parameters of the covariance matrices (see, for example, Fokoué and Titterington (2003)) could be included in the MCMC algorithm described in Chapter 4.1. Moreover, the semiparametric full mixture of experts in Section 5.1 assumes the manifest variable to be univariate, since the adaptation to the multivariate case would require particular attention to deal with the presence of component-specific to the covariance matrices.

As far as the computational implementation is concerned, one of the main advantage of the proposed MCMC algorithm is the absence of MH steps. On the other hand, the use of mixture of Gaussians to approximate the logistic distribution introduces an additional latent variable that can increase the computational burden, and the implemented MCMC algorithm requires the number of components to be fixed as input. 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 high. One solution could be incorporating the choice of the number of components within the algorithm itself. As observed in the simulation studies, it might happen the proposed MCMC algorithm converges to a solution that is characterised by empty components. This peculiar behaviour could be exploited to devise a strategy similar to the one proposed by Malsiner-Walli et al. (2016), which circumvents the issue of choosing the optimal value for G and focuses the attention on the posterior distribution of the number of non-empty components, by combining a large value for G with appropriate prior distributions. Alternatively, a reversible jump MCMC algorithm could be exploited (Richardson and Green, 1997), by designing appropriate dimension-changing moves, such as split-and-merge moves (Green and Richardson, 2001) and birth-and-death moves (Stephens, 2000).

Appendix A

Brexit data further results

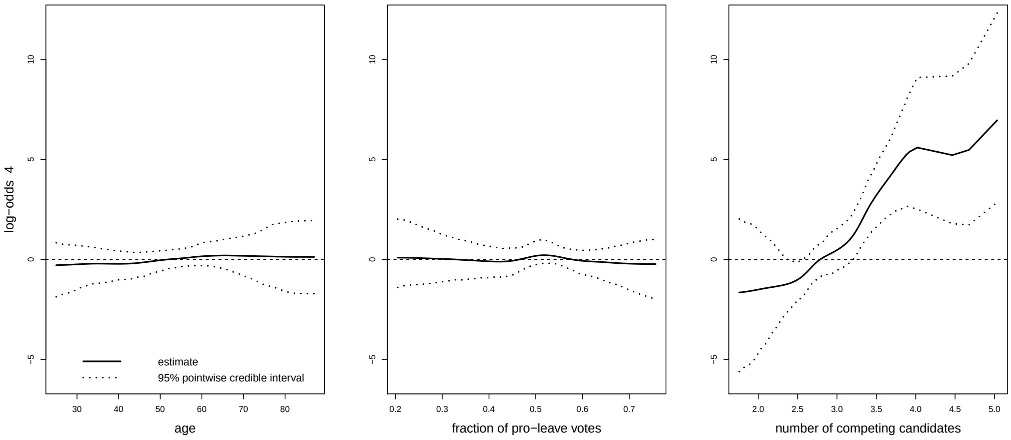


Figure A.1: Estimated smooth effects for Cluster 4

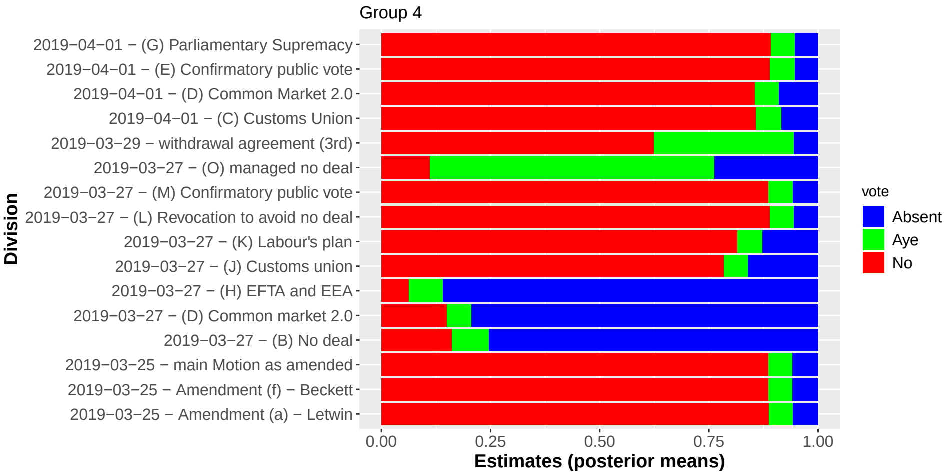


Figure A.2: Vote estimates (posterior means) for Cluster 4

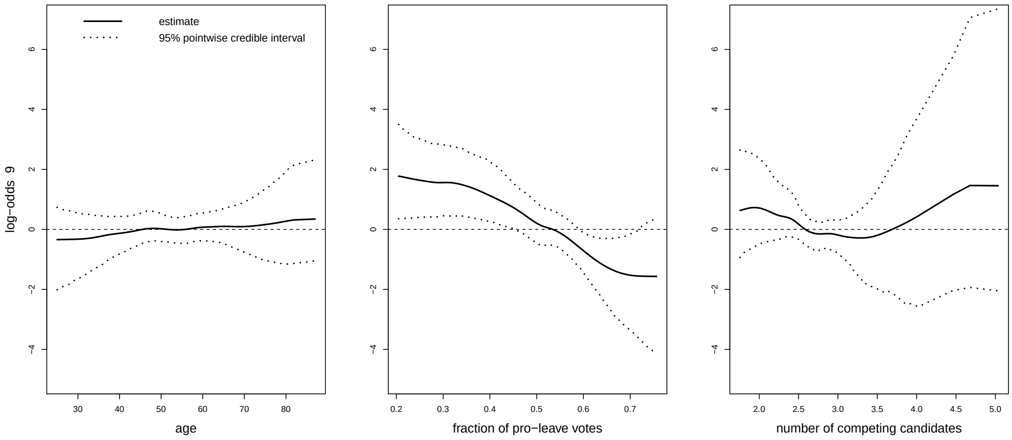


Figure A.3: Estimated smooth effects for Cluster 9

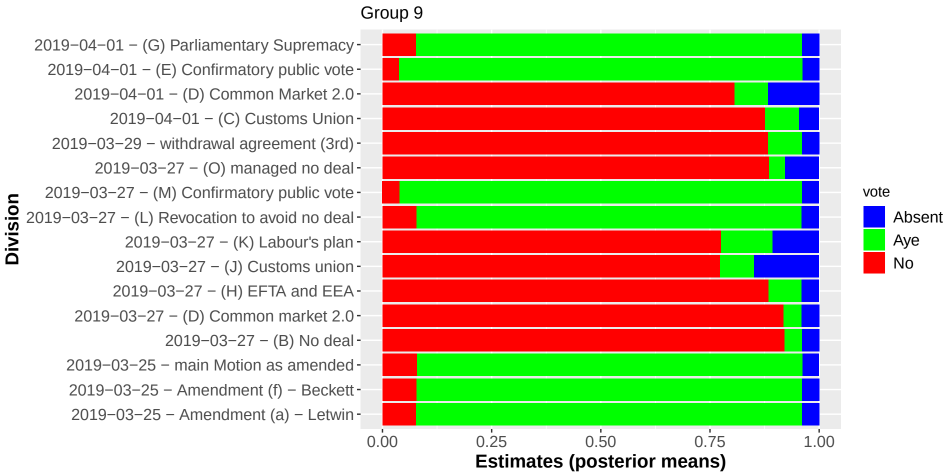


Figure A.4: Vote estimates (posterior means) for Cluster 9

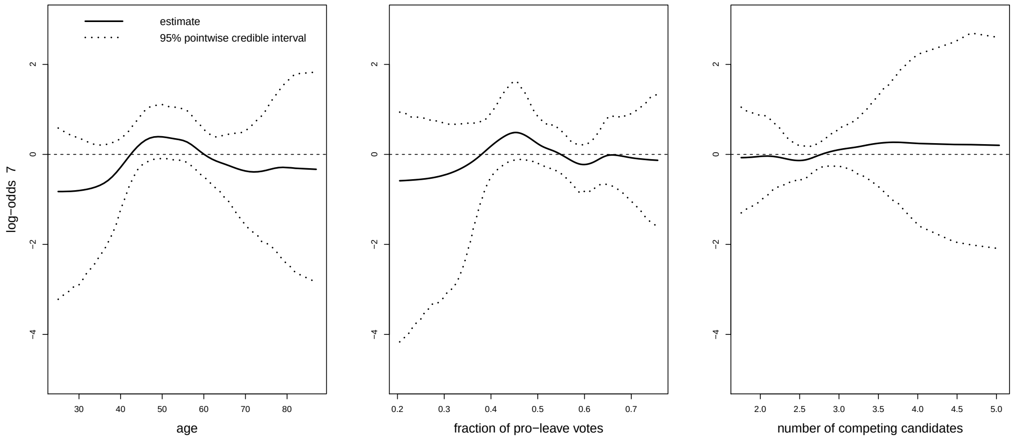


Figure A.5: Estimated smooth effects for Cluster 7

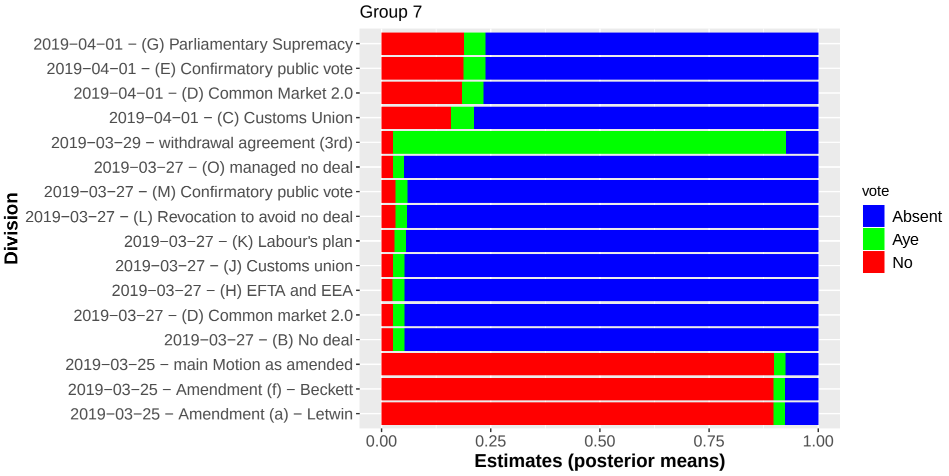


Figure A.6: Vote estimates (posterior means) for Cluster 7

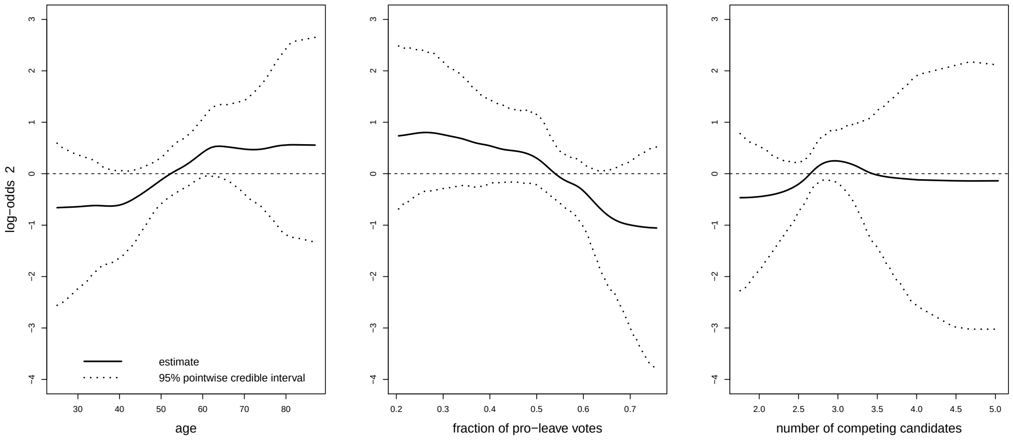


Figure A.7: Estimated smooth effects for Cluster 2

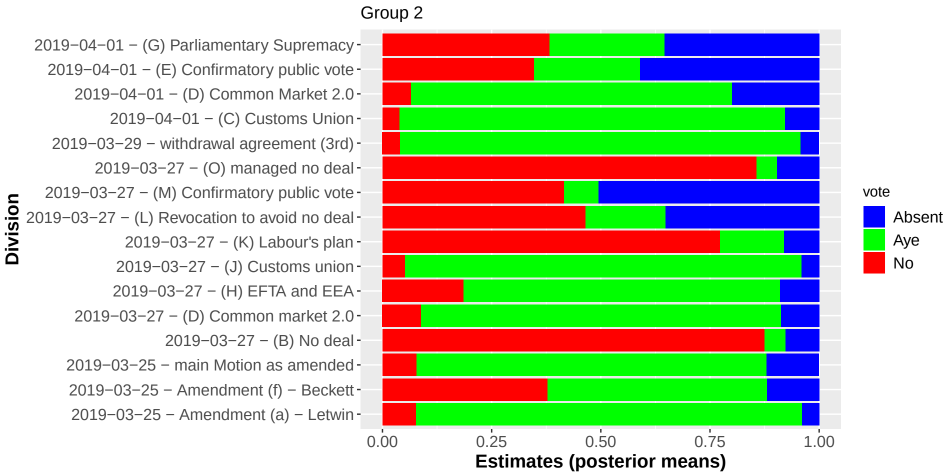


Figure A.8: Vote estimates (posterior means) for Cluster 2

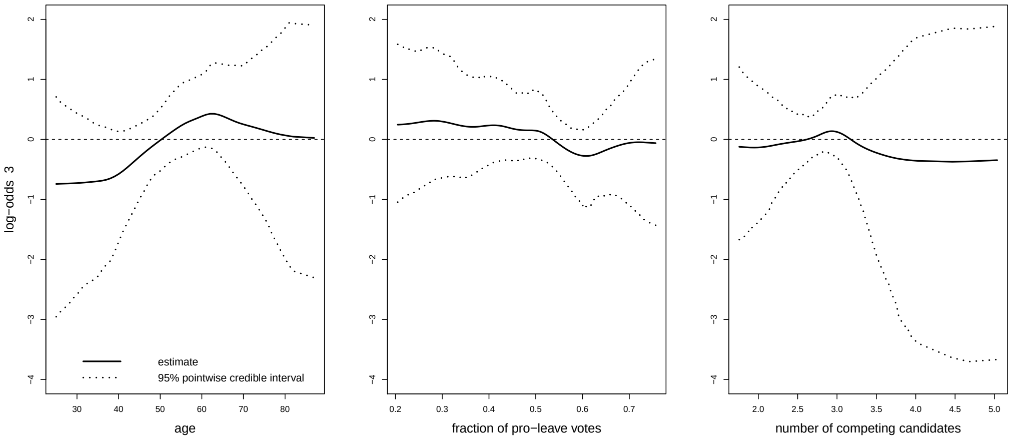


Figure A.9: Estimated smooth effects for Cluster 3

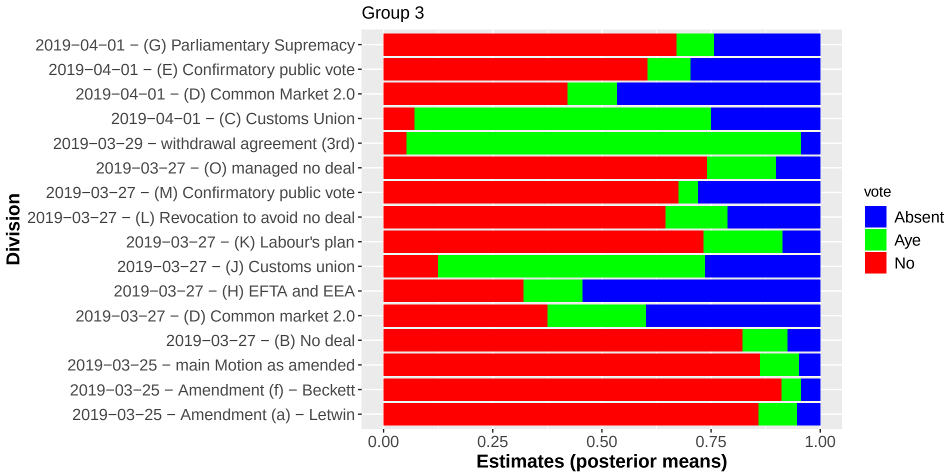


Figure A.10: Vote estimates (posterior means) for Cluster 3

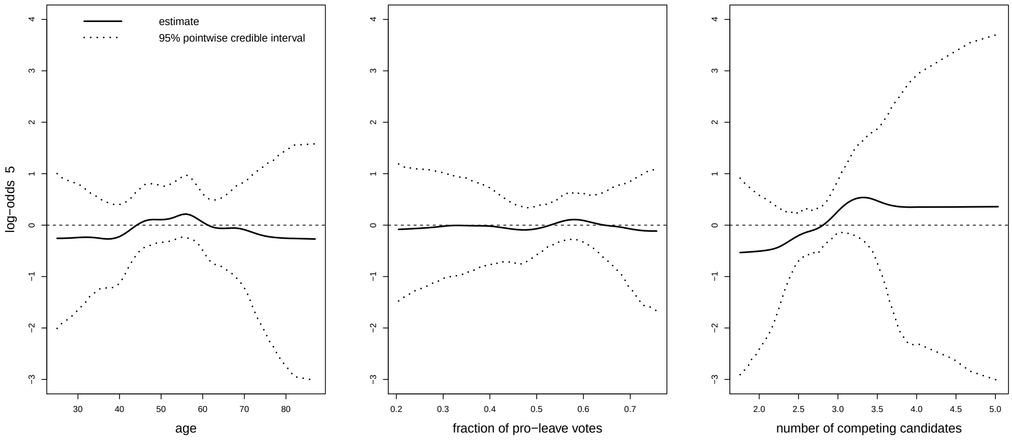


Figure A.11: Estimated smooth effects for Cluster 5

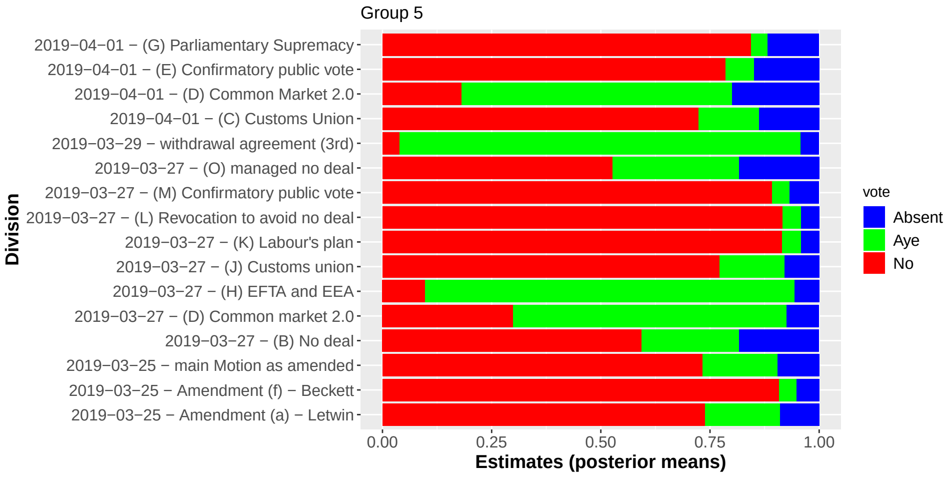


Figure A.12: Vote estimates (posterior means) for Cluster 5

Appendix B

Soccer player positions data: MCMC chain plots

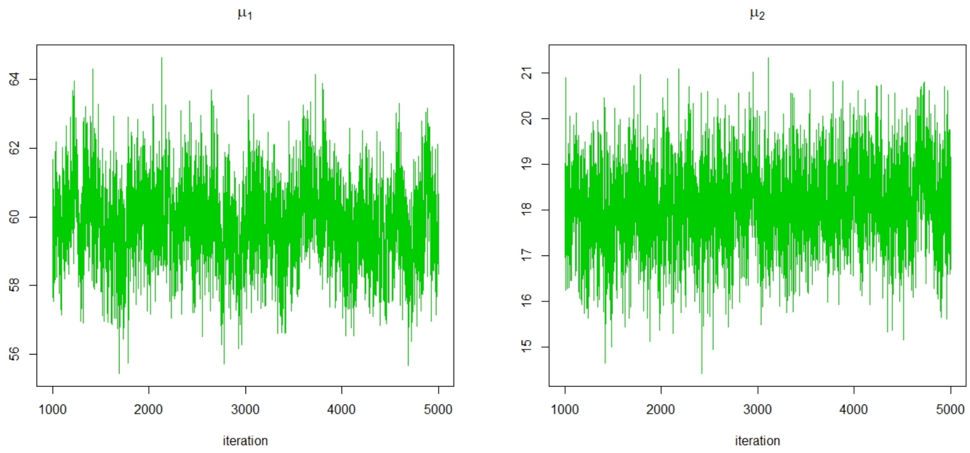


Figure B.1: MCMC draws of the mean parameters for Cluster 3 (in green) in Section 4.3.1. Light green denotes the discarted draws during the burn-in phase, delimited by a dashed line.

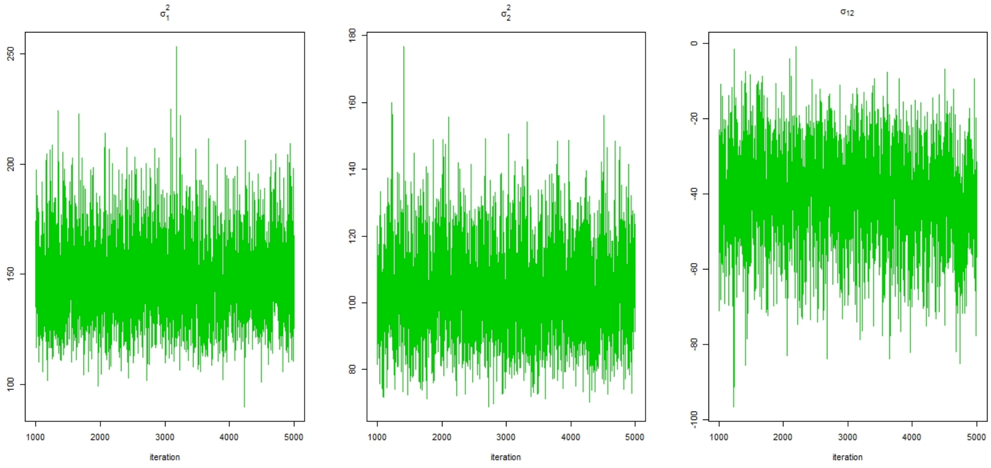


Figure B.2: MCMC draws of the (co)variance parameters for Cluster 3 (in green) in Section 4.3.1. Light green denotes the discarted draws the during burn-in phase, delimited by a dashed line.

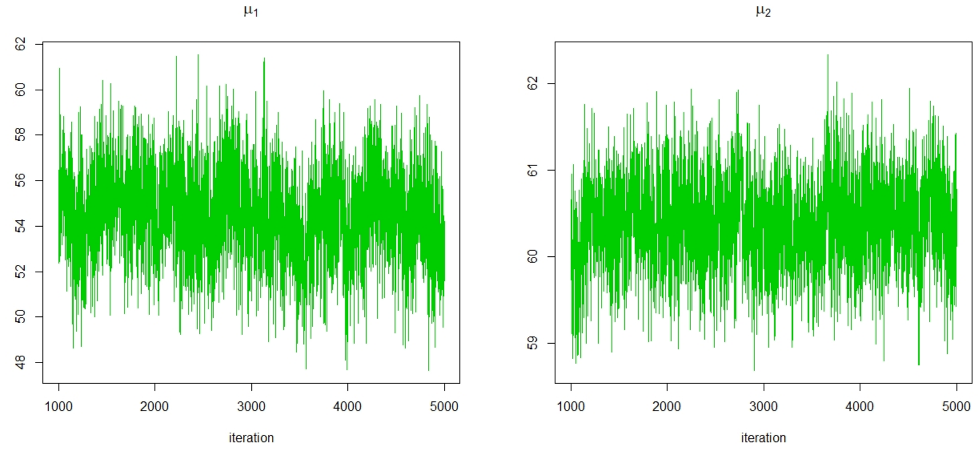


Figure B.3: MCMC draws of the mean parameters for Cluster 2 (in green) in Section 4.3.2. Light green denotes the discarted draws during the burn-in phase, delimited by a dashed line.

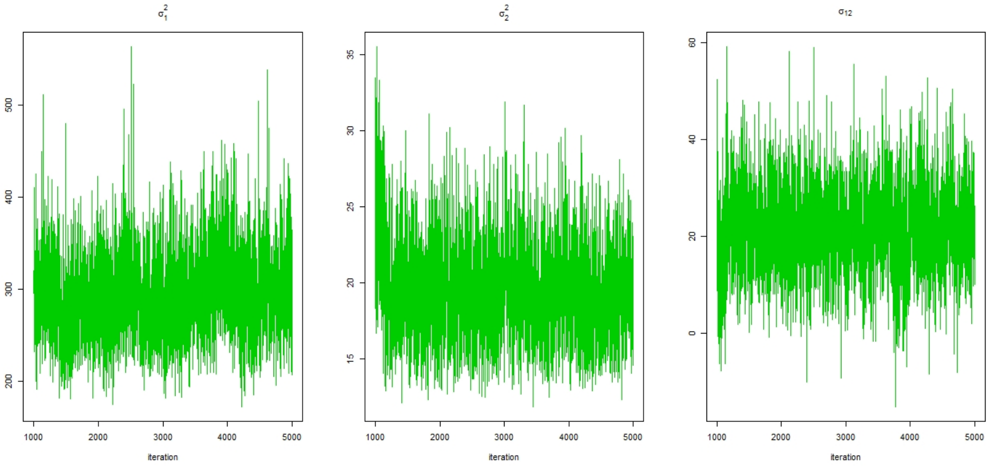


Figure B.4: MCMC draws of the (co)variance parameters for Cluster 2 (in green) in Section 4.3.2. Light green denotes the discarted draws during the burn-in phase, delimited by a dashed line.

Bibliography

Aitkin, M. and I. Aitkin (1996). A hybrid em/gauss-newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing 6*(2), 127–130.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control 19*(6), 716–723.

Althoen, S. C. and R. Mclaughlin (1987). Gauss-jordan reduction: A brief history. *The American mathematical monthly 94*(2), 130–142.

Apostolova, V., L. Audickas, C. Baker, A. Bate, R. Cracknell, N. Dempsey, O. Hawkins, R. McInnes, T. Rutherford, and E. Uberoi (2017). General election 2017: results and analysis. *Briefing Paper no CBP 7979.*

Atienza, N., J. Garcia-Heras, J. Munoz-Pichardo, and R. Villa (2007). On the consistency of mle in finite mixture models of exponential families. *Journal of Statistical Planning and Inference 137*(2), 496–505.

Baudry, J.-P. et al. (2015). Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electronic journal of statistics 9*(1), 1041–1077.

Bensmail, H., G. Celeux, A. E. Raftery, and C. P. Robert (1997). Inference in model-based cluster analysis. *statistics and Computing 7*(1), 1–10.

Berger, J. O., J. K. Ghosh, and N. Mukhopadhyay (2003). Approximations and consistency of bayes factors as model dimension grows. *Journal of Statistical Planning and Inference 112*(1-2), 241–258.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence 22*(7), 719–725.

Biernacki, C. and G. Govaert (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, 451–457.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* springer.

Bishop, C. M. and M. Svensén (2012). Bayesian hierarchical mixtures of experts. *arXiv preprint arXiv:1212.2447*.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association 112*(518), 859–877.

Böhning, D. (1999). *Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping and others*, Volume 81. CRC press.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika 52*(3), 345–370.

Brezger, A. and S. Lang (2006). Generalized structured additive regression based on bayesian p-splines. *Computational Statistics & Data Analysis 50*(4), 967–991.

Casella, G. and R. L. Berger (2002). *Statistical inference*, Volume 2. Duxbury Pacific Grove, CA.

Celeux, G., F. Forbes, C. P. Robert, D. M. Titterington, et al. (2006). Deviance information criteria for missing data models. *Bayesian analysis 1*(4), 651–673.

Celeux, G., M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*(451), 957–970.

Chamroukhi, F. (2015). Non-normal mixtures of experts. *arXiv preprint arXiv:1506.06707*.

Chamroukhi, F. and B. T. Huynh (2018). Regularized maximum-likelihood estimation of mixture-of-experts for regression and clustering. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE.

Chandra, S. (1977). On the mixtures of probability distributions. *Scandinavian Journal of Statistics*, 105–112.

Coppersmith, D. and S. Winograd (1987). Matrix multiplication via arithmetic progressions. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pp. 1–6.

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society: Series A (General) 134*(3), 321–353.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika 56*(3), 463–474.

Dayton, C. M. and G. B. Macready (1988). Concomitant-variable latent-class models. *Journal of the american statistical association 83*(401), 173–178.

De Boor, C. (1972). On calculating with b-splines. *Journal of Approximation theory 6*(1), 50–62.

De Boor, C., C. De Boor, E.-U. Mathématicien, C. De Boor, and C. De Boor (1978). *A practical guide to splines*, Volume 27. springer-verlag New York.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

DeSarbo, W. S. and W. L. Cron (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification 5*(2), 249–282.

Diebolt, J. and C. P. Robert (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological) 56*(2), 363–375.

Eilers, P. H. and B. D. Marx (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 89–102.

Erosheva, E. A., S. E. Fienberg, and C. Joutard (2007). Describing disability through individual-level mixture models for multivariate binary data. *The annals of applied statistics 1*(2), 346.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association 90*(430), 577–588.

Evans, M., I. Guttman, and I. Olkin (1992). Numerical aspects in estimating the parameters of a mixture of normal distributions. *Journal of Computational and Graphical Statistics 1*(4), 351–365.

Everitt, B. and D. Hand (1981). Finite mixture distributions, chapman and hall. *London. xi.*

Everitt, B. S., S. Landau, and M. Leese (2001). Cluster analysis, 4”ˆ edition. *Edward Amold, New York I 993.*

Flynt, A., N. Dean, and R. Nugent (2019). sari: a soft agreement measure for class partitions incorporating assignment probabilities. *Advances in Data Analysis and Classification 13*(1), 303–323.

Fokoué, E. and D. Titterington (2003). Mixtures of factor analysers. bayesian estimation and inference by stochastic simulation. *Machine Learning 50*(1), 73–94.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models.* Springer Science & Business Media.

Frühwirth-Schnatter, S. and R. Frühwirth (2010). Data augmentation and mcmc for binary and multinomial logit models. In *Statistical modelling and regression structures*, pp. 111–132. Springer.

Frühwirth-Schnatter, S., C. Pamminger, A. Weber, and R. Winter-Ebmer (2012). Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts markov chain clustering. *Journal of Applied Econometrics 27*(7), 1116–1137.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.

Geweke, J. and M. Keane (2007). Smoothly mixing regressions. *Journal of Econometrics 138*(1), 252–290.

Gormley, I. C. and T. B. Murphy (2008). Exploring voting blocs within the irish electorate: A mixture modeling approach. *Journal of the American Statistical Association 103*(483), 1014–1027.

Gormley, I. C. and T. B. Murphy (2010a). Clustering ranked preference data using sociodemographic covariates. In *Choice modelling: the state-of-the-art and the state-of-practice*. Emerald Group Publishing Limited.

Gormley, I. C. and T. B. Murphy (2010b). A mixture of experts latent position cluster model for social network data. *Statistical methodology 7*(3), 385–405.

Gormley, I. C. and T. B. Murphy (2011). Mixture of experts modelling with social science applications. *Mixtures: Estimation and Applications*, 101–121.

Gormley, I. C., T. B. Murphy, et al. (2008). A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics 2*(4), 1452–1477.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika 82*(4), 711–732.

Green, P. J. (2003). Trans-dimensional markov chain monte carlo. *Oxford Statistical Science Series*, 179–198.

Green, P. J. and S. Richardson (2001). Modelling heterogeneity with and without the dirichlet process. *Scandinavian journal of statistics 28*(2), 355–375.

Green, P. J. and B. W. Silverman (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman and Hall/CRC.

Hanretty, C. (2017). Areal interpolation and the uk's referendum on eu membership. *Journal of Elections, Public Opinion and Parties 27*(4), 466–483.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics 8*(3), 431–444.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized additive models*, Volume 43. CRC press.

Huang, M., R. Li, and S. Wang (2013). Nonparametric mixture of regression models. *Journal of the American Statistical Association 108*(503), 929–941.

Huang, M. and W. Yao (2012). Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association 107*(498), 711–724.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of classification 2*(1), 193–218.

Hunter, D. R. and K. Lange (2004). A tutorial on mm algorithms. *The American Statistician 58*(1), 30–37.

Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G. E. Hinton (1991). Adaptive mixtures of local experts. *Neural computation 3*(1), 79–87.

Jordan, M. I. and R. A. Jacobs (1994). Hierarchical mixtures of experts and the em algorithm. *Neural computation 6*(2), 181–214.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, 49–66.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the american Statistical association 102*(479), 1025–1038.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The annals of mathematical statistics 22*(1), 79–86.

Lang, S. and A. Brezger (2004). Bayesian p-splines. *Journal of computational and graphical statistics 13*(1), 183–212.

Lange, K. (2016). *MM optimization algorithms.* SIAM.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 1350–1360.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pp. i–163. JSTOR.

Linzer, D. A., J. B. Lewis, et al. (2011). polca: An r package for polytomous variable latent class analysis. *Journal of statistical software 42*(10), 1–29.

Malsiner-Walli, G., S. Frühwirth-Schnatter, and B. Grün (2016). Model-based clustering based on sparse finite gaussian mixtures. *Statistics and computing 26*(1-2), 303–324.

Marin, J.-M., K. Mengersen, and C. P. Robert (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics 25*, 459–507.

Marx, B. D. and P. H. Eilers (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis 28*(2), 193–209.

McGrory, C. A. and D. Titterington (2007). Variational approximations in bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis 51*(11), 5352–5367.

McLachlan, G. J. and K. E. Basford (1988). *Mixture models: Inference and applications to clustering*, Volume 38. M. Dekker New York.

McLachlan, G. J. and D. Peel (2004). *Finite mixture models.* John Wiley & Sons.

McNicholas, P. D. (2016). *Mixture model-based classification.* CRC press.

Meng, X.-L. and D. B. Rubin (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika 80*(2), 267–278.

Mengersen, K. L., C. Robert, and M. Titterington (2011). *Mixtures: estimation and applications*, Volume 896. John Wiley & Sons.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics 21*(6), 1087–1092.

Mollica, C. and L. Tardella (2017). Bayesian plackett–luce mixture models for partially ranked data. *Psychometrika 82*(2), 442–458.

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc. 26*, 394–395.

Murphy, K. and T. B. Murphy (2019). Gaussian parsimonious clustering models with covariates and a noise component. *Advances in Data Analysis and Classification*, 1–33.

Nguyen, H. D. and F. Chamroukhi (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8*(4), e1246.

Nguyen, H. D., F. Chamroukhi, and F. Forbes (2019). Approximation results regarding the multiple-output gaussian gated mixture of linear experts model. *Neurocomputing 366*, 208–214.

Nguyen, H. D. and G. J. McLachlan (2016). Laplace mixture of linear experts. *Computational Statistics & Data Analysis 93*, 177–191.

Olteanu, M. and J. Rynkiewicz (2011). Asymptotic properties of mixture-of-experts models. *Neurocomputing 74*(9), 1444–1449.

Parliament, U. (2018). European union (withdrawal) act 2018. *Parliament UK, retrieved 10*, 2017–19.

Pauler, D. K. (1998). The schwarz criterion and related methods for normal linear models. *Biometrika 85*(1), 13–27.

Peng, F., R. A. Jacobs, and M. A. Tanner (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association 91*(435), 953–960.

Pettersen, S. A., D. Johansen, H. Johansen, V. Berg-Johansen, V. R. Gaddam, A. Mortensen, R. Langseth, C. Griwodz, H. K. Stensland, and P. Halvorsen (2014). Soccer video and player position dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pp. 18–23.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American statistical association 67*(338), 306–310.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raftery, A., M. Newton, J. Satagopan, and P. Krivitsky (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (Eds.), *Bayesian Statistics*, Volume 8, pp. 1–45. Oxford University Press.

Ranalli, M. and R. Rocci (2017). Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis 110*, 87–102.

Redner, R. et al. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Annals of Statistics 9*(1), 225–228.

Redner, R. A. and H. F. Walker (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review 26*(2), 195–239.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische mathematik 10*(3), 177–183.

Richardson, S. and P. J. Green (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology) 59*(4), 731–792.

Robert, C. P., T. Ryden, and D. M. Titterington (1999). Convergence controls for mcmc algorithms, with applications to hidden markov chains. *Journal of Statistical Computation and Simulation 64*(4), 327–355.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica: Journal of the Econometric Society*, 577–591.

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

Schwarz (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

Scott, S. L. (2011). Data augmentation, frequentist estimation, and the bayesian analysis of multinomial logit models. *Statistical Papers 52*(1), 87–109.

Smith, A. F. and G. O. Roberts (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological) 55*(1), 3–23.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology) 64*(4), 583–639.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Annals of statistics*, 40–74.

Tang, X. and A. Qu (2016). Mixture modeling for longitudinal data. *Journal of Computational and Graphical Statistics 25*(4), 1117–1137.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association 82*(398), 528–540.

Teicher, H. (1963). Identifiability of finite mixtures. *The annals of Mathematical statistics*, 1265–1269.

Titterington, D. (1990). Some recent research in the analysis of mixture distributions. *Statistics 21*(4), 619–641.

Titterington, D. M., A. F. Smith, and U. E. Makov (1985). *Statistical analysis of finite mixture distributions*. Wiley,.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, 450–469.

Wang, B., D. M. Titterington, et al. (2006). Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis 1*(3), 625–650.

Wang, P., M. L. Puterman, I. Cockburn, and N. Le (1996). Mixed poisson regression models with covariate dependent rates. *Biometrics*, 381–400.

Wang, S., W. Yao, and M. Huang (2014). A note on the identifiability of nonparametric and semiparametric mixtures of glms. *Statistics & Probability Letters 93*, 41–45.

Watnik, M. R. (1998). Pay for play: Are baseball salaries based on performance? *Journal of Statistics Education 6*(2).

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

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

Xu, L., M. I. Jordan, and G. E. Hinton (1995). An alternative model for mixtures of experts. In *Advances in neural information processing systems*, pp. 633–640.

Yakowitz, S. J. and J. D. Spragins (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 209–214.