Bayesian Analysis (2024) **19**, Number 3, pp. 843–867

Mixtures of Probit Regression Models with Overlapping Clusters[*]

Saverio Ranciati[†,‖], Veronica Vinciotti[‡], Ernst C. Wit[§], and Giuliano Galimberti[¶]

Abstract. Studies with binary outcomes on a heterogeneous population are quite common. Typically, the heterogeneity is modelled through varying effect coefficients within some binary regression setting combined with a clustering procedure. Most of the existing methods assign statistical units to distinct and non-overlapping clusters. However, there are scenarios where units exhibit a more complex organization and the clusters can be thought as partially overlapping. In this case, the standard approach does not work. In this paper, we define a mixture of regression models that allows overlapping clusters. This approach involves an overlap function that maps the regression coefficients, either at the unit or response level, of the parent clusters into the coefficients of the multiple allocation clusters. In order to deal with this intrinsic heterogeneity, regression analyses have to be stratified for different groups of observations or clusters. We present a computationally efficient Monte Carlo Markov Chain (MCMC) scheme for the case of a mixture of probit regressions. A simulation study shows the overall performance of the method. We conclude with two illustrative examples of modelling voting behavior, involving United States (US) Supreme Court justices over a number of topics and members of the United Kingdom (UK) parliament over divisions related to Brexit. These applications provide insights on the usefulness of the method in real applications. The method described can be extended to the case of a generic mixture of multivariate generalized linear models under overlapping clusters.

Keywords: heterogeneity, mixture models, overlapping clusters, Bayesian inference, binary data, probit regression.

1 Introduction

Clustering approaches are popular in many fields as they allow to identify unknown grouping structures from multivariate data. In a regression context, mixtures of Generalized Linear Models (GLMs) provide a natural model-based approach to account for the heterogeneity in the data due to the presence of heterogeneous clusters. These models have been developed extensively in the case of a single response variable, i.e., via a mixture of univariate GLMs (Grün and Leisch, 2008a). However, with the advent

[*] This project was partially supported by the European Cooperation in Science and Technology, COST Action “European Cooperation for Statistics of Network Data Science” (CA15109). EW acknowledges funding by the Swiss National Science Foundation (SNSF grants 188534 and 192549).

[†] Department of Statistical Sciences, University of Bologna, Bologna (Italy), saverio.ranciati2@unibo.it

[‡] Department of Mathematics, University of Trento, Trento (Italy)

[§] Institute of Computational Science, Universitá della Svizzera italiana, Lugano (Switzerland)

[¶] Department of Statistical Sciences, University of Bologna, Bologna (Italy)

[‖] Corresponding author: saverio.ranciati2@unibo.it

© 2024 International Society for Bayesian Analysis https://doi.org/10.1214/23-BA1372

of complex multivariate data both at the level of predictors and responses, multivariate regression models are now common in many fields (Fahrmeir and Tutz, 2013). Mixtures of multivariate GLMs were introduced by Wedel and DeSarbo (1995) with an implementation provided in the `R` package `FlexMix` (Grün and Leisch, 2008b). Further extensions to the high dimensional case have been studied recently by Price and Sherwood (2017) using penalized inferential approaches.

In the traditional formulation of a mixture model, a unit can belong to one and only one of the clusters, thus limiting the way we can build and discover more complex grouping structures of the units. Some extensions to allow for overlapping clusters have been proposed with respect to conventional mixtures of exponential family distributions and mixtures of univariate regression models. In an early contribution, Blei et al. (2003) defined overlapping clusters in term of a multivariate random variable whose univariate components belong to a traditional mixture distribution. To define overlapping clusters at a univariate level, Banerjee et al. (2005) and Fu and Banerjee (2008) proposed overlapping components in the mixture that arise as a product of conjugate densities from the exponential family. Heller and Ghahramani (2007) extended this approach in a nonparametric Bayesian fashion to allow for the selection of the number of components of the mixture, whereas Heller et al. (2008) provide a mixed membership model where parameters of some clusters are computed as weighted averages of allocation probabilities and parameters of the other clusters. Recently a further extension was provided in Hou-Liu and Browne (2021), where the authors describe a finite mixture of Gaussian distributions where parameters of overlapping or, in their terminology, *chimeral* clusters are defined as convex combinations of some prototype original groups.

In all the contributions above, the way the multiple allocation is handled induces a strict definition of the parameters of the resulting cluster, which are limited by the mathematical derivation of the product of the densities or the nature of the allocation vectors, and is implicit within the method. This hinders the flexibility of the model and the interpretability of the related parameters. An alternative view to account for overlap is that of re-parameterizing the model in such a way that parameters of the overlapping clusters are linked explicitly to those of the originating clusters. This idea was explored by Ranciati et al. (2017) in the context of univariate mixture models and by Ranciati et al. (2020) in the specific context of two-mode network data. In these approaches, the term overlap emphasizes the fact that the multiple membership or allocation vector of a statistical unit affects its parameter value. There is no inherent concept of a mixed membership. Indeed, once a choice is made about how the parameters of heir and parent clusters are linked, something that we refer to later on as overlap function, the methods perform a hard-clustering at the level of heir clusters, allowing for a unit to possibly belong with certainty to one of the "overlapping" clusters.

Following this latter idea, the main contribution of this work is to develop a mixture of regression models, that allows for the possibility of overlapping clusters and that accounts for covariate information, both at the level of the units and at the level of the responses. This was not considered by previous approaches and expands the applicability of these approaches significantly. Crucially, the approach involves an overlap function that maps the regression coefficients, either at the unit or response level, of the parent clusters into the coefficients of the multiple allocation clusters, providing additional model parsimony and enhanced interpretability of the resulting clustering. We will consider closely the case of multivariate binary data, motivated by two applications on the modelling of voting behavior. In the first application, US Supreme Court justices are clustered in two main classes and a joint allocation cluster, which determine their voting behavior on a number of rulings in 7 main topics. In the second application, members of the UK Parliament (MPs) are clustered in two clusters and an overlapping cluster in terms of their voting behavior on a number of divisions related to Brexit, considering covariates describing the MPs and the topic of the divisions. Our modelling framework is able to identify in both cases two main clusters, mostly associated to polarized opinions due to political affiliations, and a third joint allocation cluster of so called "swing votes".

The remainder of the manuscript is structured as follows. Section 2 defines the mixture of probit regression model and how it accounts for overlapping clusters, while Section 3 discusses its Bayesian implementation, including cluster allocation and model selection. Section 4 shows the computational and inferential performance of the proposed method through a simulation study, whereas Section 5 and Section 6 illustrate the method on the two different applications mentioned above. Finally, Section 7 is devoted to some final remarks and potential extensions of the method to the generic case of mixtures of multivariate generalized linear models under overlapping clusters.

2 Clustering under overlapping groups

In this Section, we describe a mixture of probit regression models that can accommodate for overlapping clusters. We call our proposal `miro`, as in **mi**xture of probit **r**egression models with **o**verlap. The starting point is n multivariate binary observations measured on a d-dimensional space spanned by $j = 1, \ldots, d$ columns (variables), so that each response vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{id})$ corresponds to a row of the $n \times d$ data matrix Y. The aim is to cluster these n observations based on their d variables, while accounting for possible additional information (covariates), either at the level of the units $i = 1, \ldots, n$ or of the variables $j = 1, \ldots, d$. We look first at some motivating examples before formalizing the approach.

2.1 Motivating examples

Examples of real world datasets falling in the modeling framework developed in this paper are those that will be described in Section 5 and Section 6. In the first one, the task is of clustering nine justices (n units) of the US Supreme Court according to their votings (d variables), while considering the nature of the topic they voted on (variable-specific covariate). While some degree of polarized agreement/disagreement with the majority vote is expected in the results of the Supreme Court decisions, we are interested in finding a grouping structure that is able to also highlight the so called 'swing votes', which stem from opinions on the discussed topics that are not entirely polarized, and should be reflected in voting patterns that are a mixture of more polarized clusters.

In the second one, the interest lies in modelling the voting behavior of MPs with regards to different divisions that are related to Brexit, following from the referendum that initiated Brexit. In this example, information is available both at the level of MPs, such as safeness of the seat of each MP and the general opinion on Brexit of the constituency they were elected into (unit-specific covariates), and at the level of the divisions, namely the more general topic of the division itself (variable-specific covariate). Given the controversy around Brexit, we expect, also in this case, a number of MPs whose behavior may well be explained by some overlap between the two main clusters of Labour versus Conservative voters.

In the field of network science, both these applications could be seen as a special case of two-mode networks (Wasserman and Faust, 1994, Chapter 8). Also known as bipartite or affiliation networks, these are networks consisting of two types of nodes, where links can occur only between nodes of a different type. An example of a two-node network is a group of *actors* attending or not attending a set of *events.* Thus, from an agent-based point of view, a two-mode network can be seen as a collection of d-dimensional multivariate binary response variables for n actors. As shown in Ranciati et al. (2020), allowing for cluster overlap in these settings can improve significantly the characterization of the clusters as well as leading to model parsimony. The approach developed in this paper applied on actor-event data would allow to cluster actors according to their pattern of attendance to events, while considering potential covariates, both at the level of the actors and of the events. These are often available, as in the examples described above.

2.2 A conventional mixture of regression models

In a mixture of regression framework for multivariate binary data, the usual assumption is that the response variables for each unit i come from a mixture of K-components (also called clusters) specified as

$$\mathbf{y}_i \sim \sum_{k=1}^{K} \alpha_k f\big(\mathbf{y}_i; \boldsymbol{\pi}_{ki}\big),$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ are the prior cluster probabilities, K is a positive integer, $f(\cdot)$ is the d-dimensional joint distribution for the vector $\mathbf{y}_i$ of binary observations, and $\boldsymbol{\pi}_{ki}$ is the collection of parameters governing the joint distribution, which are cluster-specific and a function of covariate information. An alternative hierarchical representation of the mixture model is achieved by introducing a unit-specific binary latent vector $\boldsymbol{\zeta}_i = (\zeta_{i1}, \dots, \zeta_{iK})$, made up of all zeroes with the exception of a single element $\zeta_{ik} = 1$ for unit i belonging to cluster k. Using these latent elements, the hierarchical formulation is given by

$$\begin{aligned} \boldsymbol{\zeta}_i | \boldsymbol{\alpha} \quad &\sim \quad \text{Multinom}(\alpha_1, \dots, \alpha_K), \qquad (1) \\ \mathbf{y}_i | \boldsymbol{\zeta}_i, \boldsymbol{\pi}_{ki} \quad &\sim \quad \prod_{k=1}^{K} \prod_{j=i}^{d} \big[\text{Bern}\big(y_{ij}; \pi_{kij}\big)\big]^{\zeta_{ik}}, \end{aligned}$$

where the joint $f(\cdot)$ has been factorized into a product of Bernoulli distributions by making the common assumption that y_{ij} are independent from one another for all pairs i, j, conditional on the allocation of each unit i into the groups. The cluster-specific probabilities π_{kij} are now indexed in both i and j because we will consider both covariate information that can be unit-specific, $\mathbf{x}_i = (\mathrm{x}_{i1}, \ldots, \mathrm{x}_{iL})$, or variable-specific, $\mathbf{w}_j = (\mathrm{w}_{j1}, \ldots, \mathrm{w}_{jQ})$. Here, 'variable-specific' refers to information pertaining to the d observed variables, which are usually called dependent or response variables in the regression literature, whereas 'unit-specific' is reserved for covariates pertaining to statistical units' characteristics. For example, the $n \times L$ matrix $\mathbf{X}$, with i-th row $\mathbf{x}_i$, could collects L characteristics of the n units, such as age, gender and education in a social study setting, whereas the $d \times Q$ matrix $\mathbf{W}$ describes the Q characteristics of the d variables, i.e., a time or location for the measurement of variable j.

The probabilities of each component are linked to the covariates via an appropriate link function applied to a linear combination of the predictors (Fahrmeir and Tutz, 2013). In particular, the probability for individual i, variable j, and cluster k is given by $\pi_{kij} = g(\eta_{kij})$ with $g(\cdot)$ a link function and η_{kij} the linear predictor. We choose $g(\cdot)$ to be the Gaussian cumulative distribution function $\pi_{kij} = \Phi\big(\eta_{kij}\big)$. This choice of a link function induces a probit regression model formulation that we consider closely in this paper. The linear predictor is defined as

$$\eta_{kij} = \mu_k + \mathbf{x}_i \boldsymbol{\beta}_k^{\intercal} + \mathbf{w}_j \boldsymbol{\gamma}_k^{\intercal}, \tag{2}$$

where $\{\mu_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k\}_{k=1,\ldots,K}$ are cluster-specific vectors of parameters. More precisely, for each cluster $k = 1, \ldots, K$, the model includes an intercept μ_k, an L-dimensional vector of regression coefficients $\boldsymbol{\beta}_k$ for the unit covariates and a Q-dimensional vector of regression coefficients $\boldsymbol{\gamma}_k$ pertaining to the variable-specific covariates. The latter can be thought as equivalent to random effects in multilevel models. According to (2), the effect of unit-specific covariates $\mathbf{x}_i$ is the same for all the d response variables within a particular cluster k: this means that the elements $\{\eta_{kij}\}_{k=1,\ldots,K}$ of the linear predictor differ, with respect to j, only due to the effect of the variable-specific covariates $\mathbf{w}_j$. We note that in the case $d = 1$ and $K > 1$, the model reverts back to a mixture of univariate GLMs (Wedel and DeSarbo, 1995; Grün and Leisch, 2008a); when $d > 1$ but $K = 1$, we obtain a multivariate probit, or more generally a multivariate GLM (Fahrmeir and Tutz, 2013), and if both $K = 1$ and $d = 1$, we are back in a simple GLM (McCullagh and Nelder, 1989).

2.3 Overlapping formulation of a mixture of regression model

In this Section, we describe our proposed extension of the classical mixture of regression models to the case of overlapping clusters. Using similar ideas to Ranciati et al. (2020), we modify the hierarchical model in (1) by relaxing conditions on the allocation vectors $\boldsymbol{\zeta}_i$ in order to allow for a multiple classification of the units. In particular, denoting with $\boldsymbol{z}_i = (\mathrm{z}_{i1}, \ldots, \mathrm{z}_{iK})$ the new allocation vector, we will allow $\boldsymbol{z}_i \in \{0, 1\}^K$, that is, the set of all sequences of zeroes and ones of length K. If there are K primary clusters, then there are $K^\star = 2^K$ multiple cluster allocations. Each of these $K^\star$ allocations is a non-overlapping *heir* cluster, which defines a new $K^\star$-dimensional allocation vector $\boldsymbol{z}_i^\star$ for each unit i. This new vector satisfies $\sum_{h=1}^{K^\star} z_{hi}^\star = 1$, and has a 1-to-1 correspondence with the $\boldsymbol{z}_i$, which allocates units into the overlapping *parent* clusters. The $z^\star$ re-parametrization can now be used as the basis of a traditional hierarchical model, namely

$$\begin{aligned} \boldsymbol{z}_i^\star | \boldsymbol{\alpha}^\star &\sim \text{Multinom}(\alpha_1^\star, \ldots, \alpha_{K^\star}^\star), \\ \mathbf{y}_i | \boldsymbol{z}_i^\star, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\gamma} &\sim \prod_{h=1}^{K^\star} \prod_{j=1}^{d} \left[\left(\pi_{hij}^\star\right)^{y_{ij}} \left(1 - \pi_{hij}^\star\right)^{1-y_{ij}} \right]^{z_{ih}^\star}, \end{aligned}$$

where the quantities $\pi_{hij}^\star$ are a probit transformation of the linear predictor $\eta_{hij}^\star$.

The key questions are: (i) how to connect the new model quantities, $\pi_{hij}^\star$ and $\eta_{hij}^\star$, to the original ones, π_{kij} and η_{kij}; (ii) how to connect the new mixture parameters $\boldsymbol{\alpha}^\star$ to the old mixture parameters $\boldsymbol{\alpha}$ and (iii) how to interpret the resulting multiple allocation groups and their corresponding quantities.

Overlap function: connecting heir parameters to parent parameters The first question can be answered with the choice of an appropriate function for linking the quantities of the $z^\star$ parametrization to those of the original one, which we call the *overlap function* and denote it with ψ. Various overlap functions are possible and each one of them determines the way we can interpret and build the potential overlap between the components of the mixture. Each specific choice leads to different models that have their own computational and inferential considerations. Although the overlap function can be applied directly at the level of the probabilities π_{kij}, essentially following Ranciati et al. (2020), in this new setting where covariates have been added, it is computationally more advantageous to define the function at the level of the linear predictors η_{kij}, or even the regression coefficients $\{\mu, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$, because of the added benefit of interpretation that such a definition can bring. Given the definition of the linear predictor $\eta_{kij} = \mu_k + \boldsymbol{\beta}_k \mathbf{x}_i^\intercal + \boldsymbol{\gamma}_k \mathbf{w}_j^\intercal$, we collect all K elements associated to the pair (i, j) into a vector $\boldsymbol{\eta}_{.ij}$. We can then see three natural choices for the function ψ that defines the linear predictor in the re-parametrized mixture, i.e., $\eta_{hij}^\star = \psi(\boldsymbol{\eta}_{.ij}, \boldsymbol{z}_i)$: (1) the minimum, (2) the maximum, and (3) the pointwise average of the linear predictors $\eta_{1ij}, \ldots, \eta_{Kij}$, according to the multiple allocations vector $\boldsymbol{z}_i$. In particular,

1. *minimum overlap function:* $\psi_s(\boldsymbol{\eta}_{.ij}, \boldsymbol{z}_i) = \min\limits_{k: z_{ik}=1} \{\eta_{kij}\}$,
2. *maximum overlap function:* $\psi_x(\boldsymbol{\eta}_{.ij}, \boldsymbol{z}_i) = \max\limits_{k: z_{ik}=1} \{\eta_{kij}\}$,
3. *mean overlap function:* $\psi_m(\boldsymbol{\eta}_{.ij}, \boldsymbol{z}_i) = \underset{k: z_{ik}=1}{\text{mean}} \{\eta_{kij}\}$.

Figure 1 shows an example of the three overlap functions applied to a single continuous covariate in a situation with three overlapping clusters. The mean overlap function $\psi_m(\cdot)$ is itself linear, whereas both the minimum and maximum overlap functions are only piecewise linear. Moreover, the mean overlap function can be seen as an average of the intercepts and regression coefficients across those clusters, to which the unit is

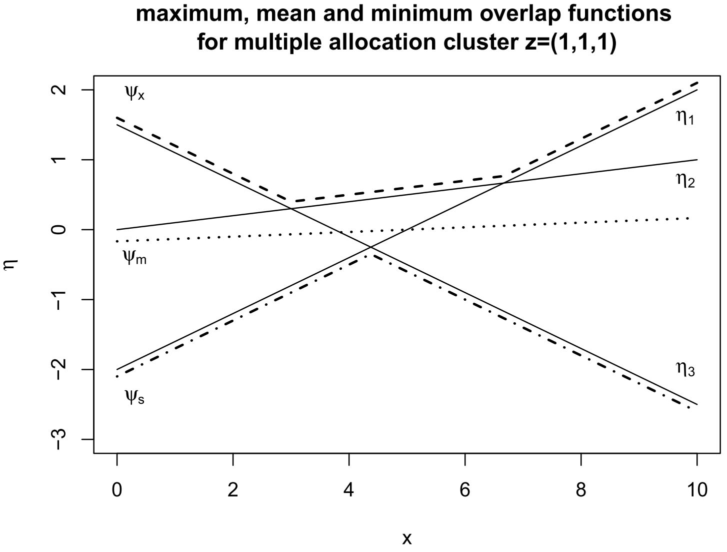


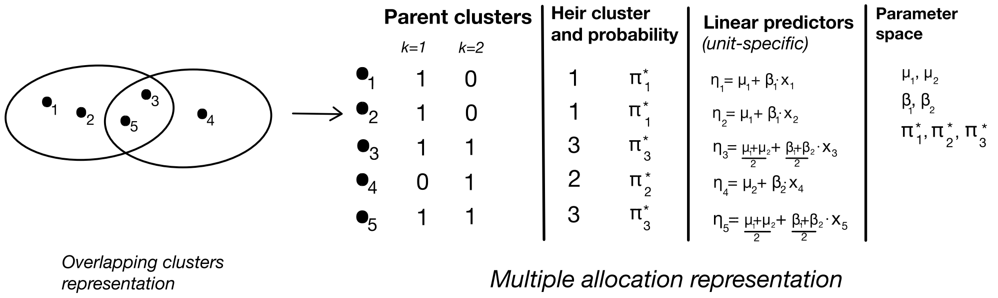
Figure 1: Example of the maximum, mean and minimum *overlap functions* for a multiple allocation cluster $\boldsymbol{z} = (1, 1, 1)$, for a single covariate x; solid lines refer to linear predictors' values, whereas dotted lines correspond to different *overlap functions.*

allocated, since

$$\psi_m(\boldsymbol{\eta}_{.ij}; \boldsymbol{z}_i) = \frac{\boldsymbol{z}_i \boldsymbol{\mu}^{\mathsf{T}}}{||\boldsymbol{z}_i||_1} + \left(\frac{\boldsymbol{z}_i \boldsymbol{B}}{||\boldsymbol{z}_i||_1}\right)\mathbf{x}_i^{\mathsf{T}} + \left(\frac{\boldsymbol{z}_i \boldsymbol{\Gamma}}{||\boldsymbol{z}_i||_1}\right)\mathbf{w}_j^{\mathsf{T}},$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$ is the vector containing all the K intercepts, $\boldsymbol{B}$ the $K \times L$ matrix whose rows are $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, and $\boldsymbol{\Gamma}$ the $K \times Q$ matrix with rows given by $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K$. So, units in multiple allocation clusters have the average effects for each covariate with respect to their parent clusters, providing enhanced interpretability compared to the other choices of overlap functions.

For the case of being allocated to none of the primary parent clusters, $\boldsymbol{z}_i = 0$, a few options are possible. First, one could follow an 'agnostic' approach and simply consider $K^\star$ to be $2^K - 1$, effectively dropping the special cluster where units have all zero elements from the allocation vector $\boldsymbol{z}_i$. This is indeed the general recommendation if there is no prior information or no explicit interest in clustering units into this residual group. The second option would be to elicit a special definition for the overlap function. In this sense, this definition depends on the individual situation: sometimes it might be possible to define the 'null' allocation as an *a priori* interpretable constant, whereas most other times it can be defined in a data-driven way. In particular, natural choices would be to use the overall minimum, $\psi(\boldsymbol{\eta}_{.ij}, 0) = \min\limits_{\mathrm{x,w}}\{\eta_{kij}(\mathbf{x}, \mathbf{w})\}$, overall maximum, $\psi(\boldsymbol{\eta}_{.ij}, 0) = \max\limits_{\mathrm{x,w}}\{\eta_{kij}(\mathbf{x}, \mathbf{w})\}$ and overall average, $\psi(\boldsymbol{\eta}_{.ij}, 0) = \underset{\mathrm{x,w}}{\mathrm{mean}}\{\eta_{kij}(\mathbf{x}, \mathbf{w})\}$, across all observed values of x and w, corresponding to the ψ_s, ψ_x and ψ_m cases, respectively.


1
2
3
5
4
Overlapping clusters representation
Multiple allocation representation

	Parent clusters		Heir cluster and probability		Linear predictors *(unit-specific)*	Parameter space
	$k=1$	$k=2$				
●1	1	0	1	π^*_1	$\eta_1 = \mu_1 + \beta_1 \cdot x_1$	μ_1, μ_2
●2	1	0	1	π^*_1	$\eta_2 = \mu_1 + \beta_1 \cdot x_2$	β_1, β_2
●3	1	1	3	π^*_3	$\eta_3 = \frac{\mu_1+\mu_2}{2} + \frac{\beta_1+\beta_2}{2} \cdot x_3$	$\pi^*_1, \pi^*_2, \pi^*_3$
●4	0	1	2	π^*_2	$\eta_4 = \mu_2 + \beta_2 \cdot x_4$	
●5	1	1	3	π^*_3	$\eta_5 = \frac{\mu_1+\mu_2}{2} + \frac{\beta_1+\beta_2}{2} \cdot x_5$	

Figure 2: Schematic representation of the proposed model for the case of two parent clusters, one unit-specific covariate and the mean overlap function.

For the rest of this manuscript, we focus on the *mean overlap function* ψ_m. In addition to the advantages of interpretability mentioned above, this function offers also computational advantages. Assuming that the d response variables are conditionally independent, given the cluster allocation, the mean overlap function translates the multivariate problem into a univariate regression and allows estimations of all the coefficients of the K clusters simultaneously. We provide the details of the implementation in the next Section, as well as a full `R` implementation of the proposed method available at `https://github.com/savranciati/miro`.

Allocation probabilities: connecting heir parameters to parent parameters Similarly to the regression coefficients, a connection between the allocation probabilities $\boldsymbol{\alpha}$ of the primary clusters $\boldsymbol{z}$ and the allocation probabilities $\boldsymbol{\alpha}^\star$ of the heir clusters $\boldsymbol{z}^\star$ can be elicited, in order to establish a correspondence between the two parametrizations. The most intuitive choice is an independent allocation mechanism among clusters, in the sense that the heir allocation probabilities are defined as $\alpha^\star_{z^\star} = \prod_{k:z_k=1} \alpha_k$. In practical applications, this is generally too restrictive and not readily verifiable. Moreover, it also indirectly implies further constraints on the size of each cluster. Therefore, in our proposed approach the heir allocation probabilities are estimated directly, without resorting to their corresponding version of the parent clusters.

Figure 2 provides a schematic representation of the proposed model on a small example with two parent clusters and one unit-specific covariate. The mean overlap function links explicitly the regression parameters associated to the overlapping cluster from those associated to the parent clusters. On the right of the plot, the parameter space is identified. The next Section focuses on statistical inference for the proposed model.

3 Bayesian implementation of `miro`

We approach inference by following a Bayesian paradigm, which requires specification of prior distributions for all parameters in our model. First, we assume the prior cluster sizes $\boldsymbol{\alpha}^\star$ to come from a Dirichlet distribution with hyper-parameters $a_1, \ldots, a_{K^\star}$, where we set $a_h = K^\star$ if $\sum_{h=1}^{K^\star} u_h = 1$ and 1 otherwise. This choice satisfies the constraints identified by previous works on the contraction rate of posterior distributions when using Dirichlet priors in the context of mixture models (Rousseau and Mengersen, 2011; Malsiner-Walli et al., 2016), and therefore offers some protection against overfitting in the case of $K^\star$ being potentially large. The intercepts $\{\mu\}_{k=1,\ldots,K}$ and regression coefficients $\{\boldsymbol{\beta}_k, \boldsymbol{\gamma}_k\}_{k=1,\ldots,K}$ are assumed to be a priori independent, normally distributed, centered at zero and with scalar variance parameters $(\sigma^2_\mu, \sigma^2_\beta, \sigma^2_\gamma)$ treated as hyper-parameters. The set of priors is summarized below

$$\begin{aligned} \boldsymbol{\alpha}^\star &\sim \mathrm{Dir}(a_1, \ldots, a_{K^\star}), \\ \mu_k &\sim \mathrm{N}(0, \sigma^2_\mu), \\ \boldsymbol{\gamma}_k &\sim \mathrm{N}_Q(\mathbf{0}_Q, \sigma^2_\gamma\, I_Q), \\ \boldsymbol{\beta}_k &\sim \mathrm{N}_L(\mathbf{0}_L, \sigma^2_\beta\, I_L) \end{aligned} \tag{3}$$

and coupled with the complete likelihood

$$L_{y,z^\star} \propto \prod_{i=1}^{n} \prod_{j=1}^{d} \left[\Phi(\eta^\star_{ij})\right]^{y_{ij}} \left[1 - \Phi(\eta^\star_{ij})\right]^{1-y_{ij}}.$$

Once one writes the complete joint posterior distribution of all the parameters and latent quantities in the model, an MCMC algorithm can be defined to sample from it. The pseudo-code for `miro` is an MCMC algorithm with Gibbs samplers, that iterates the following instructions across $t = 1, \ldots, T$ iterations:

1. Use the full conditional of $\boldsymbol{z}^\star_i$, which is

$$f(z^\star_{ih} = 1 | Y, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^\star) = \frac{\alpha^\star_h \cdot \prod_{j=1}^{d} \left[\left(\pi^\star_{hij}\right)^{y_{ij}} \left(1 - \pi^\star_{hij}\right)^{1-y_{ij}}\right]^{z^\star_{ih}}}{\sum_{h'=1}^{K^\star} \alpha^\star_{h'} \cdot \prod_{j=1}^{d} \left[\left(\pi^\star_{h'ij}\right)^{y_{ij}} \left(1 - \pi^\star_{h'ij}\right)^{1-y_{ij}}\right]^{z^\star_{ih'}}}$$

   to sample the allocation vectors $\{\boldsymbol{z}^\star_i\}_{i=1,\ldots,n}$, and then compute the updated cluster sizes $n_h = \sum_{i=1}^{n} z^\star_{ih}$;

2. Sample new values for $\{\alpha^\star_h\}$ from their full conditional

$$\boldsymbol{\alpha}^\star | Y, \boldsymbol{z}_{i=1,\ldots,n}, \mathbf{a} \sim \mathrm{Dir}(a_1 + n_1, \ldots, a_h + n_h, \ldots, a_{K^\star} + n_{K^\star});$$

3. Sample intercepts and regression coefficients $(\boldsymbol{\mu}, \boldsymbol{B} = \{\boldsymbol{\beta}\}, \boldsymbol{\Gamma} = \{\boldsymbol{\gamma}\})$ from their full conditional distributions, with $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ being matrices collecting all the regression coefficients associated to, respectively, the individual-specific and variable-specific covariates.

3.1 The specific case of the mean overlap function

Of particular notice is the fact that, (i) using the probit link and (ii) choosing the mean overlap function, one is able to cast the inferential procedure as a single Bayesian probit regression model, effectively collapsing the mixture structure during the sampling of the regression coefficients. This provides significant computational gains in sampling the set of regression coefficients (Step 3), as we further explain here.

First, we let $\tilde{\mathbf{y}}^{\mathsf{T}}$ be an $\tilde{n} \times 1$ vector obtained by stacking the columns of the data matrix Y, with $\tilde{n} = n \cdot d$. Furthermore, we stack together the cluster-specific vector of intercepts and regression coefficients for individual covariates into a vector $\tilde{\boldsymbol{\beta}} = [\mu_1 \;\; \boldsymbol{\beta}_1 \;\; \mu_2 \;\; \boldsymbol{\beta}_2 \;\; \dots \;\; \mu_K \;\; \boldsymbol{\beta}_K]$, and we do the same with variable-specific covariates coefficients as a vector $\tilde{\boldsymbol{\gamma}} = [\boldsymbol{\gamma}_1 \quad \boldsymbol{\gamma}_2 \;\; \dots \;\; \boldsymbol{\gamma}_K]$. We can then define a single probit regression model, simultaneously for all components, as follows

$$\tilde{y}_i | \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \boldsymbol{z}_i \sim \mathrm{Bern}\big(\Phi(\tilde{\eta}_i)\big)$$

with $\tilde{\eta}_i$ being the i-th element of $\tilde{\boldsymbol{\eta}} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}^{\mathsf{T}} + \tilde{\mathbf{W}}\tilde{\boldsymbol{\gamma}}^{\mathsf{T}}$. The design matrices $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{W}}$ are built by filtering the corresponding matrices of covariates' values $\mathbf{X}$ and $\mathbf{W}$ through the allocation of each unit, and building a block structure to reflect the possible configurations of $\boldsymbol{z}_i$. In particular, we collect in a matrix $\mathbf{X}_{[h]}$ the predictors' recorded values for the $n_h = \sum_{i=1}^{n} z^{\star}_{hi}$ units allocated into cluster h, and we stack them vertically d times. In addition, we append a column unitary vector of length n_h to account for the intercept. The matrix $\mathbf{W}_{[h]}$ is simply built by stacking $\mathbf{W}$ exactly n_h times, in order to have conforming dimensions. The resulting matrices have $n_h \cdot d$ rows and $(L+1)$ and Q number of columns, respectively. This process is repeated for $h = 1, \dots, K^{\star}$. Finally, the block structures of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{W}}$ are defined such that

$$\tilde{\mathbf{X}} = \begin{bmatrix} \frac{\mathbf{u}_1}{||\mathbf{u}_1||_1} \otimes \mathbf{X}_{[1]} \\ \vdots \\ \frac{\mathbf{u}_h}{||\mathbf{u}_h||_1} \otimes \mathbf{X}_{[h]} \\ \vdots \\ \frac{\mathbf{u}_{K^\star}}{||\mathbf{u}_{K^\star}||_1} \otimes \mathbf{X}_{[K^\star]} \end{bmatrix}, \qquad \tilde{\mathbf{W}} = \begin{bmatrix} \frac{\mathbf{u}_1}{||\mathbf{u}_1||_1} \otimes \mathbf{W}_{[1]} \\ \vdots \\ \frac{\mathbf{u}_h}{||\mathbf{u}_h||_1} \otimes \mathbf{W}_{[h]} \\ \vdots \\ \frac{\mathbf{u}_{K^\star}}{||\mathbf{u}_{K^\star}||_1} \otimes \mathbf{W}_{[K^\star]} \end{bmatrix},$$

where $\mathbf{u}_h$ is the h-th row of U and $\otimes$ is the Kronecker product, and $|| \cdot ||_1$ is the sum of the absolute values of the elements of the vector. The final design matrices $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{W}}$ have $\tilde{n} = n \cdot d$ rows and $(L+1) \times K$ and $Q \times K$ columns, respectively. The sub-matrices involving $h = 1$ are not used to sample the regression coefficients if the corresponding cluster $(0, \dots, 0)$ is not considered in the model or has been assigned some fixed parameter values. The matrix U is defined to contain all the possible configurations of 1s and 0s of length K, so that we have $z^{\star}_{hi} = \mathbb{1}_{[\mathbf{u}_h = \boldsymbol{z}_i]}$ with $\mathbf{u}_h$ denoting the h-th row of U and $\mathbb{1}_{[\cdot]}$ the indicator function.

As a simple example, when $K = 2$ and only unit-specific covariates are considered, the relevant quantities are

$$U = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \\ \mathbf{u}_4 \end{bmatrix}, \qquad \tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{0}_{1+L} & \mathbf{0}_{1+L} \\ \mathbf{X}_{[2]} & \mathbf{0}_{1+L} \\ \mathbf{0}_{1+L} & \mathbf{X}_{[3]} \\ \frac{1}{2}\mathbf{X}_{[4]} & \frac{1}{2}\mathbf{X}_{[4]} \end{bmatrix}, \qquad \tilde{\boldsymbol{\beta}}^{\intercal} = \begin{bmatrix} \mu_1 \\ \boldsymbol{\beta}_{[1]} \\ \mu_2 \\ \boldsymbol{\beta}_{[2]} \end{bmatrix}$$

and each unit i may be assigned:

- to none of the two clusters, $\boldsymbol{z}_i = \mathbf{u}_1 = (0, 0)$, corresponding to $\boldsymbol{z}_i^{\star} = (1, 0, 0, 0)$;
- only to the first *parent* cluster, $\boldsymbol{z}_i = \mathbf{u}_2 = (1, 0)$, corresponding to $\boldsymbol{z}_i^{\star} = (0, 1, 0, 0)$;
- only to the second *parent* cluster, $\boldsymbol{z}_i = \mathbf{u}_3 = (0, 1)$, corresponding to $\boldsymbol{z}_i^{\star} = (0, 0, 1, 0)$;
- to both of them, $\boldsymbol{z}_i = \mathbf{u}_4 = (1, 1)$, the *heir* cluster corresponding to $\boldsymbol{z}_i^{\star} = (0, 0, 0, 1)$.

We can now employ this formulation of the linear predictor, which accounts simultaneously for all K clusters and their overlappings, into the probit model framework of Holmes and Held (2006) as a single regression procedure. More specifically, in Step 3 of the pseudo-code, the set of regression coefficients, $\boldsymbol{\mu}$, $\boldsymbol{B}$, and $\boldsymbol{\Gamma}$, is sampled together in a single probit regression step. We adopt the framework suggested by Holmes and Held (2006), where a random latent utility r is introduced, such that $y_{ij} = 1$ if $r_{ij} > 0$, and zero otherwise, for $i =, \dots, n$ and $j = 1, \dots d$. Then the utility is defined as $r_{ij} = \eta_{ij}^{\star} + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$ and the linear predictor $\eta_{ij}^{\star}$ is the same of a probit regression on the original y_{ij}. The approach leads to the following update mechanism for the regression coefficients $\boldsymbol{\theta}^{\intercal} = [\boldsymbol{B} \;\; \boldsymbol{\Gamma}]$ in $\eta_{ij}^{\star}$:

i) $D \leftarrow 0$;

ii) for $i = 1, \dots, n$ and $j = 1, \dots, d$

$$\begin{aligned} m_{ij} &\leftarrow A_{ij}\,\boldsymbol{\theta}^{(t-1)}, \\ r_{ij} &\leftarrow \texttt{truncNorm}(y_{ij}; m_{ij}, 1), \\ D &\leftarrow D + r_{ij} S_{ij} \end{aligned}$$

with truncNorm($\cdot$) denoting sample values from a truncated Normal distribution;

iii) $C \leftarrow \mathrm{N}(\mathbf{0}, I)$;

iv) $\boldsymbol{\theta}^{(t)} \leftarrow D + \mathrm{Chol}(V)^{\top} C$,

where V is the prior block-covariance matrix of the coefficients in $\boldsymbol{\theta}$; A is the full design matrix obtained by stacking side-by-side both $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{W}}$, $S = V\,A^{\top}$ and Chol($\cdot$) extracts the lower-triangular matrix from the Cholesky decomposition.

3.2 Model selection, posterior allocation, and label switching

In the context of our approach, model selection is equivalent to choosing the number of primary clusters K. A fully Bayesian specification would prescribe a prior on this quantity and, due to the nature of the model, would require the implementation of a trans-dimensional MCMC version of our algorithm, such as a reversible jump MCMC (Green, 1995). For two reasons we avoid this approach. First, sampling K from its posterior distribution might lead to computational issues within our framework. In fact, given that $K^\star$ scales exponentially with K, the support of the prior distribution would have to be rather narrow to avoid unfeasible values of $K^\star$, which defeats the main purpose of using a distribution for K. Second, we expect the 'true' number of primary clusters K to be small in practice, as even $K = 4$ can accommodate for up to $K^\star = 16$ clusters with the proposed approach.

On the basis of these considerations, we opt for a more heuristic model selection approach, in which we fit our model for different values of K and then select the optimal one through an information criterion. In particular, we rely on the Bayesian Information Criterion MCMC (BIC-MCMC) (Frühwirth-Schnatter, 2011), recently employed to select infinite mixtures of infinite factor analyzers models (Murphy et al., 2019). The BIC-MCMC criterion typically encourages the selection of parsimonious models, and it is defined as

$$\text{BIC-MCMC} = -2l_{\max} + \log(n \cdot d) \cdot p_\theta,$$

where $l_{\max}$ is the maximum value of the log-likelihood across the MCMC chain (after burn-in) and p_θ is the number of parameters in the model effectively sampled and not computed (Wit et al., 2012). In particular, p_θ refers to the parameters involved in the K original clusters and not their *heir* counterpart, which are instead computed using the selected overlap function and do not contribute to effective number of parameters.

After the choice of K and, implicitly, $K^\star$ is made, units are allocated into clusters according to their average posterior probabilities and using the Maximum-A-Posteriori (MAP) rule. In particular, unit i will be assigned to the cluster h that attains the highest value for

$$\bar{\mathrm{P}}(\boldsymbol{z}_i^\star = h | Y, \boldsymbol{\alpha}^\star, \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{P}(\boldsymbol{z}_i^\star = h | Y, \boldsymbol{\alpha}^\star, \boldsymbol{\theta}),$$

computed after the initial burn-in window.

A well-known problem with mixture models in the Bayesian paradigm is the label switching phenomenon (Celeux, 1998; Stephens, 2000; Sperrin et al., 2010). Although, in theory, a desirable property of the formulation, as it allows the MCMC chain to visit all the modes of the target distribution, the label switching arises from the invariance property of the likelihood with respect to the order of the cluster labels. From a practical point of view, this is reflected in unwanted difficulties while summarizing posterior quantities, i.e., posterior means, posterior standard deviations, etc., for some parameters of interest. To tackle this issue, we reorder the MCMC output of the algorithm through the geometrically-based Pivotal Reordering Algorithm (PRA) (Marin et al., 2005; Marin and Robert, 2007), available as a function in the `R` package `label.switching` (Papastamoulis, 2016). The procedure needs a pivotal labelling, which we select according to the strategy proposed in Carmona et al. (2018). In particular, we first compute the matrix of co-occurrences $C^{(t)}$ at each iteration $t = 1, \ldots, T$ of the MCMC (after burn-in). This is an $n \times n$ matrix where a generic element c_{ij} is equal to one if unit i is in the same cluster of unit j for that iteration, and zero otherwise. Then, an average of these matrices is computed, denoted by $\bar{C}$. Finally, we select the labelling of iteration $t_{\min}$ as our pivotal quantity, where

$$t_{\min} = \underset{t}{\operatorname{argmin}} \big\{ \big[C^{(t)} - \bar{C} \big]^2 \big\}.$$

Once samples have been relabelled according to the algorithm, we can compute posterior quantities of interest.

4 Simulation study

We investigate the performance of our model under two different data generating processes: (i) data coming from our proposed model, described in Section 2; (ii) data obtained from a mixture of non-overlapping components, with covariates and a logit link function. For each simulation, we simulate 25 independent datasets, and we average the results across the replicates. The performance is measured via: (i) the misclassification error rate (*MER*), which is the fraction of wrongly allocated units with respect to the true labeling, and (ii) the Adjusted Rand Index (*ARI*), that measures how much the true and inferred labellings agree with each other. As both simulation settings are characterized by four non-overlapping components, the benchmarks values for *MER* and *ARI* are, respectively, 68.5% and 0%. The first is computed as $\sum_k [\alpha_k(1 - \alpha_k)]$, where $\boldsymbol{\alpha} = (0.10, 0.45, 0.25, 0.20)$ is the chosen vector of cluster sizes: thus, it is the probability of a wrong allocation under random assignments of units to clusters. The second is by definition the rand index obtained under a random allocation, with maximum value of 100.0% indicating a perfect match in classification. In both simulation studies, we perform inference using four competing models: (i) `mixtbern`, a conventional mixture of Bernoulli distributions; (ii) `manet`, a mixture of Bernoulli distributions with overlapping clusters (Ranciati et al., 2020); (iii) `mixtprobit`, a classical mixture of probit regression models; (iv) `miro`, our proposed model. For `manet` and `miro`, the prior distributions on the cluster sizes are set to $\mathrm{P}(\alpha_1^\star, \alpha_2^\star, \ldots, \alpha_h^\star, \ldots, \alpha_{K^\star}^\star) = \mathrm{Dir}(a_1, a_2, \ldots, a_h, \ldots, a_{K^\star})$ with $a_h = K^\star$ if $\sum_{h=1}^{K^\star} u_h = 1$ and 1 otherwise. Moreover, for these methods, the MER and ARI values are calculated at the level of the $K^\star$ heir clusters. In each scenario, Bayesian inference is conducted by running the algorithms for 10000 MCMC iterations with a 50% burn-in window. In the following Sections, we give more details about each setting and discuss the results.

4.1 Synthetic data from `miro` model

Data are simulated from $K = 2$ overlapping clusters, using the hierarchical formulation of the `miro` model as the data generating process, namely a mixture model with overlapping components and a probit regression formulation. We consider 5 scenarios in

Type of covariates	Sample size n	MER (%)			
	# of variables d	`mixtbern`	`manet`	`mixtprobit`	`miro`
unit-specific	$n = 50,\ d = 10$	36.16	**25.60**	34.88	27.44
unit-specific	$n = 50,\ d = 20$	30.24	21.28	24.16	**14.72**
variable-specific	$n = 50,\ d = 15$	26.08	19.44	20.24	**15.60**
variable-specific	$n = 150,\ d = 15$	21.87	20.56	20.69	**13.73**
unit & variable	$n = 250,\ d = 21$	22.13	21.79	22.42	**18.00**

Table 1: Simulation study with data generated from `miro`: Misclassification Error Rate (MER), averaged across 25 replicated datasets, for each of the four competing models. Performance values associated to the best model are reported in bold font.

Type of covariates	Sample size n	ARI (%)			
	# of variables d	`mixtbern`	`manet`	`mixtprobit`	`miro`
unit-specific	$n = 50,\ d = 10$	31.29	**50.41**	30.57	43.09
unit-specific	$n = 50,\ d = 20$	41.44	61.77	53.66	**68.51**
variable-specific	$n = 50,\ d = 15$	51.60	65.98	62.50	**67.62**
variable-specific	$n = 150,\ d = 15$	59.16	61.70	62.48	**70.61**
unit & variable	$n = 250,\ d = 21$	59.89	55.15	59.23	**66.00**

Table 2: Simulation study with data generated from `miro`: Adjusted Rand Index (ARI), averaged across 25 replicated datasets, for each of the four competing models. Performance values associated to the best model are reported in bold.

total, according to the type of covariates considered, and by varying either the sample size n or the number of binary variables d. In particular, we consider:

- Settings with unit-specific covariates only: sample size $n = 50$ units, $d = \{10, 20\}$ variables, $L = 1$ continuous covariate x sampled from a Standard Normal distribution;
- Settings with variable-specific covariates only: sample size $n = \{50, 150\}$ units, $d = 15$ variables, $Q = 2$ binary covariates $(\mathrm{w}_1, \mathrm{w}_2)$ from one categorical covariate with three levels;
- Setting with unit- and variable-specific covariates: sample size $n = 250$ units, $d = 21$ variables, $L = 1$ continuous covariate x sampled from a Standard Normal distribution, $Q = 2$ binary covariates $(\mathrm{w}_1, \mathrm{w}_2)$ from one categorical covariate with three levels.

The results are reported in Table 1 and Table 2 for MER and ARI, respectively. With the exception of the scenario with the lowest sample size and number of variables, $n = 50$ and $d = 10$, respectively, the proposed model (`miro`) outperforms the competitors in every other scenario considered, both in terms of *MER* and *ARI*. This is expected, given that we are simulating from the same model which is then used for inference. We further notice that not only sample size but also increasing the number of variables leads to better performances. In particular, increasing the number of variables d has a positive effect for the setting with only unit-specific covariates: this is due to the fact that, having more attendances to events, using an actor-event terminology, is analogous to having more time points in a repeated measures model framework. A similar argument can be made for the effect of sample size n on the performances in scenarios where there are only variable-specific covariates.

4.2 Synthetic data from misspecified model

We now simulate data from $K = 4$ non-overlapping groups via a mixture model of binary regression models, where probabilities $\{\pi_{kij}\}$ are computed as a logit transformation of the linear predictor

$$\eta_{kij} = \mu_k + \beta_{k1}\mathrm{x}_{i1} + \gamma_{k1}\mathrm{w}_{j1}.$$

Here, we simulate $n = 300$ units and $d = 20$ variables. This data generating process differs from `miro` due to the fact that: (i) logit is used instead of the probit link function; (ii) units belong only to one component at a time, as in a conventional mixture model. As covariates, we use: $L = 1$ continuous unit-specific covariate x sampled from a standard Normal distribution; $Q = 1$ binary variable-specific covariate w. The results are reported in Table 3.

Performances degrade in this setting with respect to the previous simulation (Section 4.1), due to the misspecification of all the models that we fit. However, the results for `miro` with $K = 3$ and $K = 4$ are less affected than those for the two competing models, `mixtbern` and `manet`. As expected, `mixtprobit` performs on par or slightly better than `miro` in terms of ARI and MER. Indeed, this is the model closest to the data generating process, with the only difference being in the use of a probit link in place of the logit link.

Model	MER (%)	ARI (%)	BIC-MCMC
`mixtbern`, $K = 2$	46.29	21.46	9049.91
`mixtbern`, $K = 3$	44.43	16.59	8765.40
`mixtbern`, $K = 4$	44.69	16.11	8954.96
`manet`, $K = 2$	46.61	20.61	7462.24
`manet`, $K = 3$	44.92	16.62	7442.89
`manet`, $K = 4$	48.87	15.35	7513.01
`mixtprobit`, $K = 2$	41.88	29.74	7171.69
`mixtprobit`, $K = 3$	27.09	49.35	6973.28
`mixtprobit`, $K = 4$	25.25	50.86	6948.71
`miro`, $K = 2$	41.65	25.67	7022.59
`miro`, $K = 3$	**26.37**	**50.26**	**6944.91**
`miro`, $K = 4$	29.01	45.37	7019.25

Table 3: Simulation study from a misspecified setting, with data generated with $K = 4$ non-overlapping components, mixture of logistic regression models, $n = 300$ units, $d = 20$ variables. Misclassification error rate (MER), Adjusted Rand Index (ARI), and BIC-MCMC values are averaged across 25 replicated datasets. Performance values associated to the best model are reported in bold.

As a further comparison between the methods, we evaluate the fit of the models to the data via BIC-MCMC. The last column of Table 3 reports the value for each model, as an average out of the 25 replicated datasets. These results point to `miro` with $K = 3$ clusters as the best model. Looking at the individual BIC values for each dataset, we find that almost 70% of the time `miro` with $K = 3$ clusters was the preferred choice, with the remaining cases selecting `mixtprobit` with $K = 4$ as the best model. We also compute the proportion of times each K was selected according to BIC-MCMC within each model class, and find that (i) $K = 3$ was the optimal choice 96% of the times for `miro`; (ii) $K = 4$ was the selected number of clusters 92% of the times for `mixtprobit`; (iii) `manet` performed better in terms of the BIC-MCMC 64% of the times with $K = 3$ and approximately 36% of the times with $K = 2$; (iv) 96% of the times, $K = 3$ was the best model within the `mixtbern` class. In conclusion, `miro` was often the optimal choice according to BIC-MCMC, with a clear choice of K for achieving the best trade-off between fitness and parsimony.

5 US Supreme Court: agreement and polarization

In this section, we describe how during the 2000–2001 term the 9 US Supreme Court justices could be clustered with respect to their decisions on 26 important cases (Greenhouse, 2001). These data were analyzed in Doreian and Fujimoto (2003) and then further explored in Doreian et al. (2004). The $n = 9$ units in the data are justices Breyer, Ginsburg, Souter, Stevens, O'Connor, Kennedy, Rehnquist, Scalia and Thomas, whereas the decisions represent the $d = 26$ binary variables. According to Greenhouse (2001), the decisions can be categorized into 7 main topics, which can be used as a categorical covariate W with the following levels: "Presidential Election", "Criminal law", "Federal authority", "Civil rights", "Immigration law", "Speech and Press", "Labor and Properties". Each observation is coded as $y_{ij} = 1$ when justice i was part of the majority decision j, while $y_{ij} = 0$ stands for the situation where the justice voted with the minority. The goal of the analysis is to be able to cluster the nine justices according to their voting patterns, considering a situation where both patterns of polarized opinions and overlapping agreements/disagreement with the majority vote can exist.

We apply four clustering algorithms on this dataset: `mixtbern`, `manet`, `mixtprobit` and `miro`, as in the simulations. For all of them, we opt for 10,000 MCMC iterations with a generous 5,000 burn-in window. Model fit comparison is conducted quantitatively in terms of BIC-MCMC and qualitatively using the clustering output. In Table 4 we report the selected number of clusters K according to the BIC-MCMC value. We also

Model	K, # of cluster	BIC-MCMC	# of parameters
`mixtbern`	3	571.06	81
`manet`	2	438.73	56
`mixtprobit`	3	368.15	24
`miro`	2	**332.28**	18

Table 4: Model selection criterion reported for the three competing algorithms; K is the number of cluster selected for each model.

provide a proxy for the complexity of the models by reporting the number of parameters. The results show how `miro` achieves the lowest BIC-MCMC. Moreover, `manet` and `miro` produce the same classification of the justices, which suggests that no real better fit is provided by `manet` at the expense of increased model complexity. In particular, these models allocate justices Breyer, Ginsburg, Souter, and Stevens into a primary cluster $(1, 0)$, whereas Rehnquist, Scalia, and Thomas are grouped together in the second primary cluster (0,1). Appropriately, O'Connor and Kennedy are allocated into the multiple allocation cluster $(1, 1)$. Indeed, it has been well-documented that Kennedy and O'Connor constituted the swing vote in the Supreme Court (Toobin, 2008). On the other hand, `mixtbern` identifies three separate clusters, where Kennedy is put together with Rehnquist, Scalia, and Thomas, while O'Connor is allocated alone into a third group. Finally, BIC-MCMC suggests a value of $K = 3$ in the case of `mixtprobit`, although the posterior classification of the units leaves one cluster empty, with Breyer, Ginsburg, Souter and Stevens in one cluster and O'Connor, Kennedy, Rehnquist, Scalia and Thomas in the other. This clustering seems less meaningful.

Unlike `manet` and `mixtprobit`, `miro` is able to make use of additional covariate information to aid the clustering, while allowing for potential overlapping of the resulting groups. Figure 3 visualizes the results for `miro` in terms of clustered data and posterior means of regression coefficients for the two primary clusters. Coherently with the allocations of the model, there are some decision types that better discriminate between the voting behavior of the two primary clusters. In particular, those decisions belonging to the categories "Federal Authority", "Presidential Election" and "Labor and Properties" on the right side of the bottom plot in Figure 3 clearly discriminate the liberal judges in cluster 2 from the conservative judges in cluster 1. For the other four decision categories, the 9 justices are in much closer agreement.

6 Brexit: divisions and parties

We present here an analysis of parliamentary votes of UK MPs on motions related to Brexit. We use data originally studied by Berrettini et al. (2021) and retrieved using the `R` package `hansard` (Odell, 2017). The votes, also known as "meaningful votes", are divisions taken under the terms of Section 13 of the United Kingdom's European Union – Withdrawal – Act 2018 (Parliament, 2018), which required UK to discuss parliamentary motions at the end of the negotiations that followed Brexit. Our sample size concerns a subset of 281 MPs, who voted either "Aye" (coded with $y_{ij} = 1$) or "No" ($y_{ij} = 0$) on each voting occasion during the time frame of 25/03/2019 to 01/04/2019; for this reason, MPs with at least one abstention or not present during at least one of the divisions are excluded from the data. Of all the votings, we focus on eight divisions, described as follows: (1) "No deal": Conservative MP Mr John Baron's proposal to immediately leave the European Union (EU) without any deal; (2) "Common market 2.0": the proposal to join the Single Market and a customs union; (3) "European Free Trade Association (EFTA) and European Economic Area (EEA)": the proposal to remain in the Single Market outside of a customs union; (4) "Customs union": the proposal for a permanent customs union; (5) "Labour's plan": Labour's alternative position proposed by MP Mr Jeremy Corbyn, including a comprehensive set of arrangements and agreements

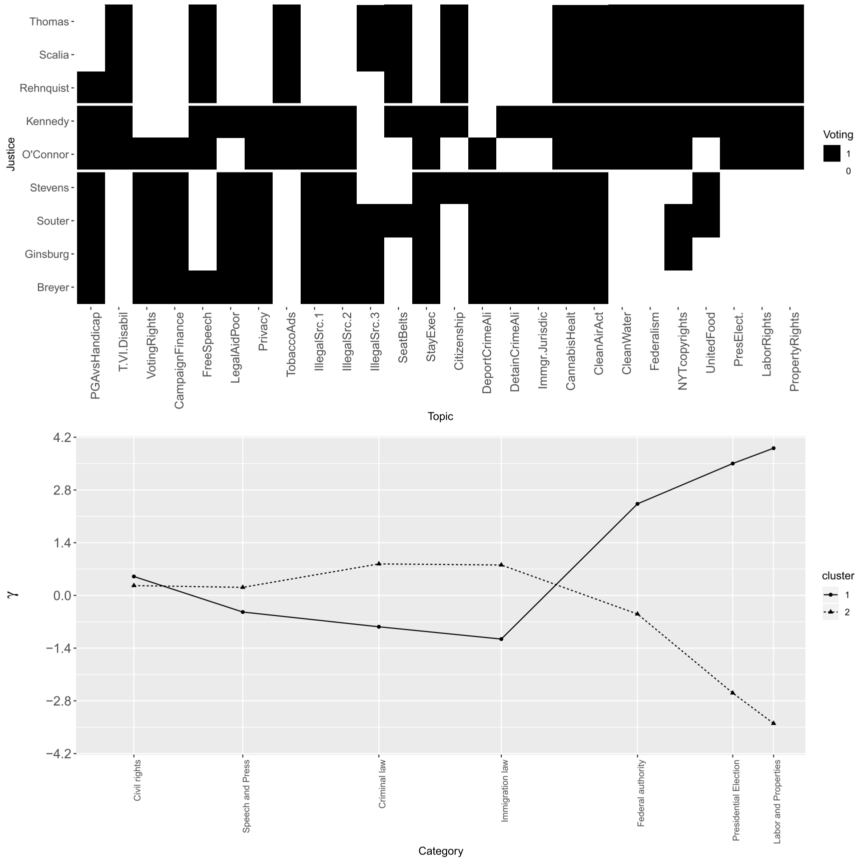


Figure 3: Observed data and results for `miro` with $K = 2$; (*top plot*) black tiles represent observations equal to 1, while empty tiles code for 0; row names are the justices, while column names indicate decisions: tiles are horizontally separated by white lines according to the three clusters; (*bottom plot*) regression coefficients associated to each category for the covariate "type of decision", with solid black line for primary cluster $k = 1$ and the dotted black line for the other primary cluster $k = 2$.

with the EU; (6) "Revocation to avoid no deal": Scottish National Party's proposal to revoke Article 50; (7) "Confirmatory public vote" the proposal for a public vote on any withdrawal bill; (8) "Managed no deal": the proposal to immediately leave the EU seeking a tariff-free trade agreement. The final data matrix consists then of $n = 281$ rows for each of the MPs and $d = 8$ columns representing the eight divisions.

Additional covariates are available for this study, both at the level of units and variables. In particular, we consider:

- (*unit-specific*) "Leave" (x_{1i}): the share of Leave votes at the Brexit referendum in the parliament constituency MP i was elected into;

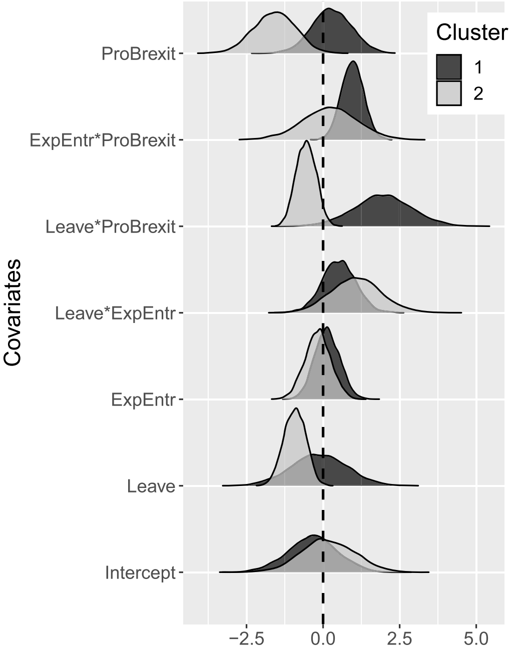


Figure 4: Posterior distributions of the regression coefficients for all covariates and their interaction terms; light-shaded distributions refer to parameters for cluster 1 and dark-shaded distributions refer to parameters for cluster 2.

- (*unit-specific*) “ExpEntr” (x_{2i}): the exponential of the entropy, computed on the shares of votes each party took in the corresponding constituency, as a measure of competitiveness for the seat MP i was given after election;
- (*variable-specific*) “ProBrexit” (w_j): an indicator denoting if voting “Aye” in division j equates to being ‘ProBrexit’ ($w_j = 1$) or ‘Against Brexit’ ($w_j = 0$),

as well as all the interaction terms between these three. We refer the reader to Berrettini et al. (2021) for additional details on both the dataset and the covariate information.

We apply our proposed method on this dataset, in order to discover its underlying structure. The best model selected according to the BIC-MCMC criterion (1335.09) is a mixture with $K = 2$ parent clusters, obtained from 10000 iterations of the MCMC algorithm with an initial 50% burn-in window as in the first example. Figure 4 visualizes posterior distributions of the regression coefficients (and intercepts) for both clusters. In terms of covariates, the two clusters are mainly characterized by having different – in sign and magnitude – effects with respect to the main effect of the variable-specific covariate “ProBrexit” as well as the interaction term between the same “ProBrexit” covariate and the unit-specific covariate “Leave”. In particular, MPs allocated to cluster 1 have a higher probability of voting “Aye” on a motion, when doing so corresponds

	Party				
Cluster	C	GP	Ind.	Lab.	LD
$\mathbf{z} = (1, 0) \longrightarrow 1$	98	0	0	0	0
$\mathbf{z} = (0, 1) \longrightarrow 2$	21	1	7	63	1
$\mathbf{z} = (1, 1) \longrightarrow 3$	83	0	4	3	0

Table 5: Two-way table for cluster allocation according to `miro` and party allegiance of each MP; Parties are coded as: C=Conservative, GP=Green Party, Ind.=Independents, Lab.=Labour, LD=Liberal Democrats.

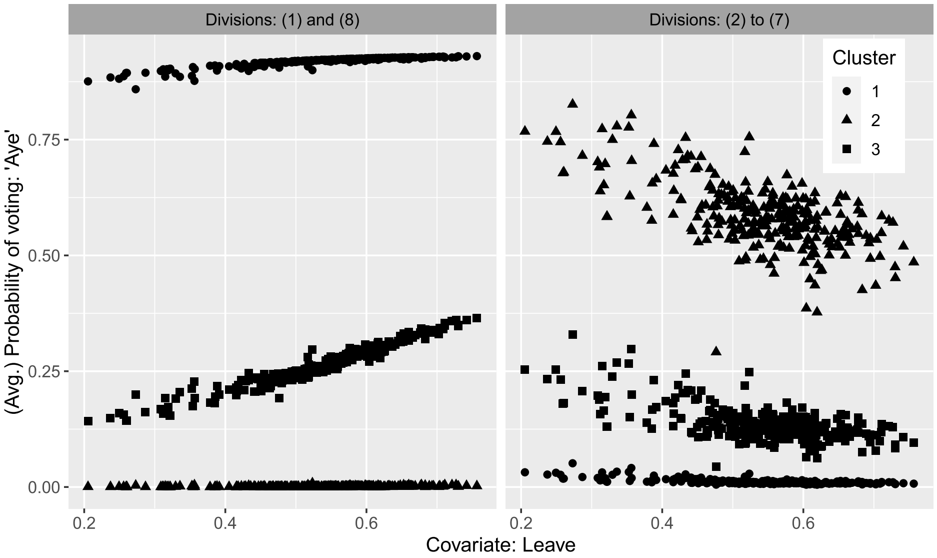


Figure 5: Average probability of voting ‘Aye’ in “Pro-Brexit” *(left)* or non “Pro-Brexit” *(right)* divisions, as a function of the covariate “Leave” and cluster allocation: dot, cluster 1; square, cluster 2; triangle, cluster 3.

to being ProBrexit (first and last division). This effect is further exacerbated if the proportion of “Leave” in the Brexit referendum is higher in the constituency those MPs have been elected into.

As a further characterization of the discovered clusters, Table 5 shows the cross-tabulation of the cluster allocations with the MPs’ parties affiliations; in Figure 5, we visualize the averages – on the retained MCMC iterations – of the estimated probabilities of voting “Aye”, conditional on the potential cluster membership and main effects of the covariates (Figure 5). The results show how the two main clusters identified by the model are predominantly comprised of MPs affiliated to, respectively, the Conservative Party (group $k = 1$) and the Labour Party (group $k = 2$).

Indeed, cluster 1 is exclusively composed of MPs of the Conservative Party, whereas in the other main cluster we can find mostly MPs from the Labour Party, but also one MP belonging to the Liberal Democrats and one MP from the Green party. The overlapping group (cluster 3) contains MPs from the Conservative party who, however, exhibited a voting pattern which is a mixture of the two originating groups.

This characterization of the clusters is enriched by looking at the estimated average probabilities of voting "Aye" conditional on the cluster allocation of the MPs and as a function of selected covariates. In particular, the plot on the left of Figure 5 depicts the effect that the proportion of "Leave" in the Brexit referendum, for each MPs' constituency, has on the probability of them voting "Aye" in "ProBrexit" divisions. Those belonging to cluster 1 (dot point) are associated with a high probability, almost close to 1, of voting "Aye", regardless of the opinion electors of the constituency might have had on the Brexit referendum. Conversely, for cluster 2, this probability is close to zero. On the other hand, units in the overlapping cluster (3) exhibit a probability of voting "Aye" influenced by the "Leave" proportion, with large values of the covariate pushing this probability to be almost a fifty-fifty chance. For divisions where voting "Aye" means being against a hard Brexit option (right plot of Figure 5), the description of the clusters is reversed. Here, there is a clearer relationship between the probability of "Aye" and the covariate for cluster 1 and 2, while the behavior of cluster 1 remains the most extreme.

7 Conclusions

In this manuscript we have presented an approach to perform model-based clustering on multivariate binary data, via a mixture of probit regression models that allows for units to be allocated to more than one cluster, while incorporating additional information in the form of covariates. The proposed method has the advantage of allowing the user to define, from a modeling perspective, how the multiple allocation clusters are related to the main parameters of the model, and, in particular, how to combine the regression coefficients in order to have a high degree of interpretability as well as an accurate identification of the clusters.

A simulation study provided encouraging results with respect to the performance of the method, in terms of the correct identification of the underlying clusters, even in situations where the simulated environment is not the same as the fitted model. The comparison was favorable also against some close competitors, such as mixtures of regression models without overlapping clusters or mixtures with overlapping clusters but without covariates. Finally, we show how the proposed methodology is useful in describing voting behavior both in an example of voting records of the US Supreme Court and on data regarding UK MPs' decisions on Brexit related motions. The data and R code for the analysis can be found in https://github.com/savranciati/miro.

The mixture of probit regressions investigated in this paper can be extended into a more general framework of mixtures of generalized linear regression models, in order to account for data of different types. For example, a multinomial probit formulation could be used to model Brexit data by incorporating all three categories of voting ("Aye", "No", "Absent"). The way overlapping clusters are handled in `miro` poses no direct limitation to the type of data that can be modeled, with the only practical restriction being the computational advantage in each situation, i.e., what is the "best" overlapping function to select in such a way that it would lead to efficiency both in terms of mathematical derivations and computational complexity. The other distribution-specific issue is how to account for additional nuisance and/or dispersion parameters that describe some of the distributions belonging to the exponential family class: in particular, how to combine them for the overlapping clusters and whether they should be set cluster-specific or not.

Another potential trajectory for future work involves the definition of the linear predictor, which could be further enriched by adding, for example, random effects for known grouping structures, or regression coefficients that vary for each possible combination of cluster k, unit i, and variable j. However, this may have the drawback of increasing significantly the number of parameters to be inferred. Finally, further extensions could revolve around the idea of introducing dependence between the response variables. In this direction, Nikoloulopoulos and Karlis (2010) adopted copula models in the context of multivariate GLMs, while other authors explored the case of mixtures of bivariate Poisson GLMs (Bermúdez and Karlis, 2012).

References

Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., and Mooney, R. J. (2005). "Model-based overlapping clustering." In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining*, 532–537. ACM. 844

Bermúdez, L. and Karlis, D. (2012). "A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking." *Computational Statistics & Data Analysis*, 56(12): 3988–3999". MR2957848. doi: https://doi.org/10.1016/j.csda.2012.05.016. 864

Berrettini, M., Galimberti, G., Ranciati, S., and Murphy, T. B. (2021). "Flexible Bayesian modelling of concomitant covariate effects in mixture models." *arXiv preprint arXiv:2105.12852*. 859, 861

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet allocation." *Journal of machine Learning research*, 3(Jan): 993–1022. 844

Carmona, C., Nieto-Barajas, L., and Canale, A. (2018). "Model-based approach for household clustering with mixed scale variables." *Advances in Data Analysis and Classification*, 13: 559–583. MR3954522. doi: https://doi.org/10.1007/s11634-018-0313-6. 855

Celeux, G. (1998). "Bayesian inference for mixture: The label switching problem." In Payne, R. and Green, P. (eds.), *COMPSTAT*, 227–232. Physica, Heidelberg. 854

Doreian, P., Batagelj, V., and Ferligoj, A. (2004). "Generalized blockmodeling of two-mode network data." *Social networks*, 26(1): 29–53. 858

Doreian, P. and Fujimoto, K. (2003). "Structures of supreme court voting." *Connections*, 25(3). 858

Fahrmeir, L. and Tutz, G. (2013). *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media. MR1284203. doi: https://doi.org/10.1007/978-1-4899-0010-4. 844, 847

Frühwirth-Schnatter, S. (2011). "Dealing with Label Switching under Model Uncertainty." In Mengersen, K. L., Robert, C. P., and Titterington, D. M. (eds.), *Mixtures*, chapter 10, 213–239. John Wiley & Sons. MR2883354. doi: https://doi.org/10.1002/9781119995678.ch10. 854

Fu, Q. and Banerjee, A. (2008). "Multiplicative mixture models for overlapping clustering." In *2008 Eighth IEEE International Conference on Data Mining*, 791–796. IEEE. 844

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4): 711–732. MR1380810. doi: https://doi.org/10.1093/biomet/82.4.711. 854

Greenhouse, L. (2001). "In year of Florida vote, Supreme Court did much other work." *New York Times*, 2. 858

Grün, B. and Leisch, F. (2008a). "Finite mixtures of generalized linear regression models." In Heumann, C. (ed.), *Recent advances in linear models and related areas*, 205–230. Springer. MR2523852. doi: https://doi.org/10.1007/978-3-7908-2064-5_11. 843, 847

Grün, B. and Leisch, F. (2008b). "FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters." *Journal of Statistical Software*, 28(4): 1–35. URL https://www.jstatsoft.org/v028/i04 844

Heller, K. and Ghahramani, Z. (2007). "A nonparametric Bayesian approach to modeling overlapping clusters." In *Artificial Intelligence and Statistics*, 187–194. 844

Heller, K. A., Williamson, S., and Ghahramani, Z. (2008). "Statistical models for partial membership." In *Proceedings of the 25th International Conference on Machine learning*, 392–399. 844

Holmes, C. and Held, L. (2006). "Bayesian auxiliary variable models for binary and multinomial regression." *Bayesian Analysis*, 1(1): 145–168. MR2227368. doi: https://doi.org/10.1214/06-BA105. 853

Hou-Liu, J. and Browne, R. P. (2021). "Chimeral Clustering." *Journal of Classification*, 1–20. MR4394832. doi: https://doi.org/10.1007/s00357-021-09396-3. 844

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). "Model-based clustering based on sparse finite Gaussian mixtures." *Statistics and computing*, 26(1): 303–324. MR3439375. doi: https://doi.org/10.1007/s11222-014-9500-2. 851

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005). "Bayesian modelling and inference on mixtures of distributions." *Handbook of Statistics*, 25: 459–507. MR2490536. doi: https://doi.org/10.1016/S0169-7161(05)25016-2. 854

Marin, J.-M. and Robert, C. (2007). *Bayesian core: a practical approach to computational Bayesian statistics*. Springer Science & Business Media. MR2289769. 854

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall. MR3223057. doi: https://doi.org/10.1007/978-1-4899-3242-6. 847

Murphy, K., Viroli, C., and Gormley, I. C. (2019). "Infinite mixtures of infinite factor analysers." *Bayesian Analysis*. MR4132655. doi: https://doi.org/10.1214/19-BA1179. 854

Nikoloulopoulos, A. K. and Karlis, D. (2010). "Regression in a copula model for bivariate count data." *Journal of Applied Statistics*, 37(9): 1555–1568. MR2758668. doi: https://doi.org/10.1080/02664760903093591. 864

Odell, E. (2017). "hansard: Provides Easy Downloading Capabilities for the UK Parliament API." *R package version 0.8. 0*, 10. 859

Papastamoulis, P. (2016). "label.switching: An R Package for Dealing with the Label Switching Problem in MCMC Outputs." *Journal of Statistical Software, Code Snippets*, 69(1): 1–24. 854

Parliament, U. (2018). "European Union (Withdrawal) Act 2018." Retrieved 2017/10/19. 859

Price, B. S. and Sherwood, B. (2017). "A Cluster Elastic Net for Multivariate Regression." *Journal of Machine Learning Research*, 18: 232–1. MR3845531. 844

Ranciati, S., Vinciotti, V., and Wit, E. C. (2020). "Identifying overlapping terrorist cells from the Noordin Top actor-event network." *Annals of Applied Statistics*, 14(3): 1516–1534. MR4152144. doi: https://doi.org/10.1214/20-AOAS1358. 844, 846, 847, 848, 855

Ranciati, S., Viroli, C., and Wit, E. C. (2017). "Mixture model with multiple allocations for clustering spatially correlated observations in the analysis of ChIP-Seq data." *Biometrical Journal*, 59(6): 1301–1316. MR3731217. doi: https://doi.org/10.1002/bimj.201600131. 844

Rousseau, J. and Mengersen, K. (2011). "Asymptotic behaviour of the posterior distribution in overfitted mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5): 689–710. MR2867454. doi: https://doi.org/10.1111/j.1467-9868.2011.00781.x. 851

Sperrin, M., Jaki, T., and Wit, E. (2010). "Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models." *Statistics and Computing*, 20(3): 357–366. MR2725393. doi: https://doi.org/10.1007/s11222-009-9129-8. 854

Stephens, M. (2000). "Dealing with label switching in mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4): 795–809. MR1796293. doi: https://doi.org/10.1111/1467-9868.00265. 854

Toobin, J. (2008). *The nine: Inside the secret world of the Supreme Court*. Anchor. 859

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge University Press. 846

Wedel, M. and DeSarbo, W. S. (1995). “A mixture likelihood approach for generalized linear models.” *Journal of Classification*, 12(1): 21–55. 844, 847

Wit, E., Heuvel, E. v. d., and Romeijn, J.-W. (2012). “‘All models are wrong...’: an introduction to model uncertainty.” *Statistica Neerlandica*, 66(3): 217–236. MR2955417. doi: https://doi.org/10.1111/j.1467-9574.2012.00530.x. 854