SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of

STATISTICS AND QUANTITATIVE METHODS

PhD program in **STATISTICS AND MATHEMATICAL FINANCE** Cycle **XXXI**

Curriculum in **STATISTICS**

New developments in Cluster-Weighted Modeling

Surname: **BARBERIS** Name: **STEFANO**

Registration number: **717748**

Tutor: Prof. MARCO **FATTORE**

Supervisor: Prof. GIORGIO **VITTADINI**

Co-tutor: Prof. MARIO **MEZZANZANICA**

Coordinator: Prof. GIORGIO **VITTADINI**

ACADEMIC YEAR 2017/2018

Contents

Part I
Introduction

Introduction and Context of Cluster Weighted Model

Keywords: Cluster Weighted Model, Mixture models, Model-based clustering.

1. Introduction

The exponential growth of data and its use in supporting business decisions has challenged the processing and storage capabilities of modern information systems. The ability to handle and managing large volumes of data has gradually turned to a strategic one (Ivanov, 2016, ch. 15). A definition of what means the term "Big Data" (Boyd and Crawford, 2012) is actually a subject of debate, however the 3V framework has gained attention since it was introduced by Laney, 2001. In that representation "Big Data" can be defined by three distinctive characteristics:

- Volume: the growing amount of data over time.
- Variety: the multitude of data sources (devices, sensors, media data, internet data).
- Velocity: how fast the data is retrieved, stored and processed.

In last years many techniques have been developed in order to deal with "Big data", and essentially these techniques are based on identifying and describing the variability present in the data following different approach. Some approaches are more "data driven" and come from the world of machine learning and data mining while others approaches are based on a model or more generally on a mathematical relation that is tested on the observed data. Obviously, depending on the specific application of interest, it is necessary to identify the approach that is right for us and in particular the methods that allow to achieve the objectives of the analysis in the fastest and most efficient way.
In this work the interest is to study and to develop new methods to manage complex relationships hidden in the data by generalizing more traditional methods, but at the same time without giving up on a model-based approach. Then, from a statistical point of view, there are at least two important themes that should be considered:

- Heterogeneity
- Significance

Heterogeneity is very general term and it is used in contrast to homogeneity: something that is homogeneous is uniform in composition and characteristics and could be described with a unique and global model. In the presence of a large amounts of data and, for a wide range of phenomena, it is natural to expect that one global model does not apply to all data at the same time and, therefore, it is not sufficient to adequately capture all the complexity within the data. Therefore, the class of mixture models and generally the mixture analysis can be useful to estimate local models that can combined together to describe the entire dataset. Regarding the theme of statistical *significance*, it is well known that when we have large data set, the significance tests associated with linear models refuses the null hypothesis in all cases, and beyond certain threshold it is only the determining factor of the test so that every explicative variable seems to be significant for explaining the outcomes (Royall, 1986; Battle and Rakov, 1993). In particular, statistical inference based on very large sample containing many heterogeneous groups, leads to irrelevance of statistical testing because of the exceeding power. In this sense the idea of fitting local models instead of a global model coming from mixture analysis represents an advantage not only to deal with heterogeneity but also to limit the problems related to the significance.
An extensive literature was developed in last years on the topic of finite mixture models. These models provide a flexible framework to analyze and describe a wide variety of random phenomena characterized by unobserved heterogeneity. A simple picture in Fig. 1 shows that there are two groups of individuals, each one with a different relation between a covariate x and a response variable y. Assuming that no other covariates are available at individual level, it is evident that there are probably one or more latent factors that characterize the population, so that we observe two different trends and we need to capture them considering a suitable class of models.
In mixtures of distributions, it is generally assumed that observed data $\boldsymbol{X}$ are independent and identically distributed from an unknown population density $p(x;\boldsymbol{\theta})$ with G latent groups; each group is represented by a mixture component $p(x;\boldsymbol{\theta}_g)$. The marginal density of $\boldsymbol{X}$ is be written as

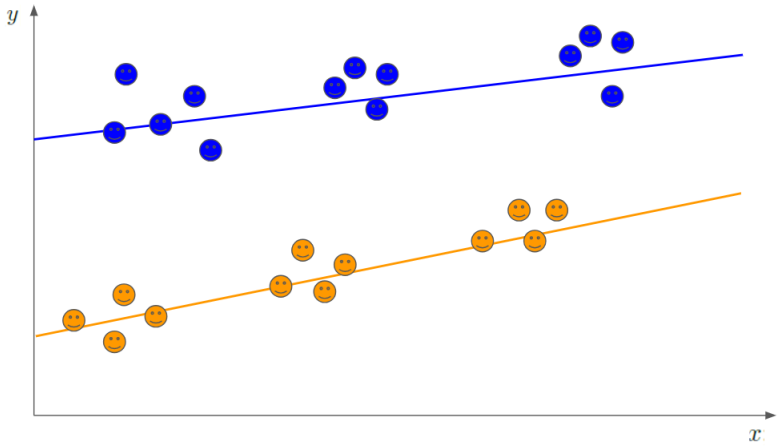
$$p(x;\boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g p(x;\boldsymbol{\theta}_g) \tag{1}$$

where $\pi_g > 0$ are the mixture weights such that $\sum_{g=1}^{G} \pi_g = 1$ and $p(x;\boldsymbol{\theta}_g)$ is a probability density function to describe the process that originates the observed data which depends on a set of unknown parameters $\boldsymbol{\theta}_g$. In mixtures of regressions, a dependence between Y (response variable) and X (a covariate) is introduced and a general mixture regression model can be defined as

$$p(y|x;\boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g p(y|x;\boldsymbol{\theta}_g) \tag{2}$$

In this models, the observations are assumed to arise from unobserved groups in the population, and the two the main objectives of mixture analysis are to identify the groups in the unknown population and to estimate the parameters of the conditional-group regression function to understand the behaviors of the process that generates observed data.

Figure 1: Example of mixture of regression where a latent variable split the population into groups.



A wide range of mixture models have been developed in literature and can be considered to analyze data in which the presence of latent factors is suspected. The main developments related to mixture models or more generally to model-based clustering can be summarized in the following points:

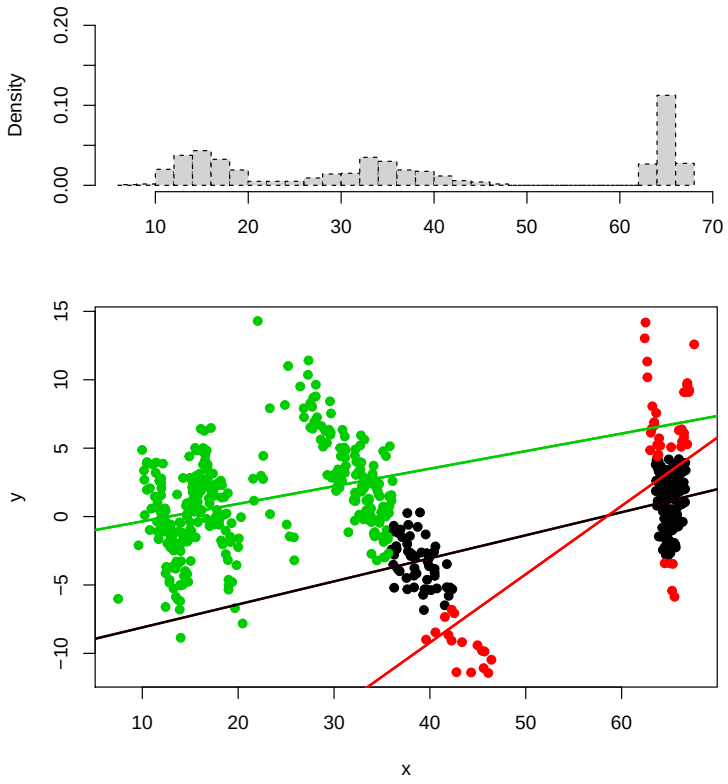
- Finite mixtures of distributions (Everitt and Hand, 1981; McLachlan and Peel, 2000; Bagnato *et al.*, 2013)
- Finite mixtures of linear regressions (Wedel and DeSarbo, 1995; McLachlan, 2008): a set of covariates or explicative variables explain the means and/or the variances of each component. See for example model (2) where the conditional distribution $p(y|\boldsymbol{x};\boldsymbol{\theta}_g)$ depends on x.
- Finite mixtures of generalized linear models (Wedel and Kamakura, 2000; Ng and McLachlan, 2008): to deal with different types of response variables,

if Y belongs to the exponential family these models are an extension of mixtures of regressions including the generalized linear model within each mixture component.

- Finite mixtures of regressions with concomitant variables (Dayton *et al.*, 1998; Wedel, 2002): a set of covariates or explicative variables explains the means and/or the variances of each component and the mixture's weights π_g depend on a set (or subset) of covariates so that $\pi_g = f_g(\boldsymbol{x})$.

All the models listed above, consider the conditional distribution $p(y|\boldsymbol{x})$ as weighted sum of local conditional distributions $p(y|\boldsymbol{x}, \boldsymbol{\theta}_g)$. However, in many cases, the marginal distribution of the covariates takes an important role to split data into clusters, therefore it would be advisable to incorporate this information directly into the model. In Fig. 2 it is possible to see how the fitted model (a three components linear mixture of regressions with concomitant variables) is not able to identify the clusters although they are well separated from the point of view of the marginal distribution of X.

Figure 2: A simulated data set where a three components linear mixture model with concomitant variables is fitted. The marginal distribution of each group is not considered in detecting groups showing, at least in this case, an incorrect detection and classification of the clusters.



In this work we investigate a different approach with respect to those proposed in the literature explained above.
The *Cluster Weighted Modeling* (CWM), is a framework that lets to take a step forward with respect to the previous models, in particular it considers the unconditional distribution $p(y, \boldsymbol{x})$ that is written as a proper weighted sum of local models, and consequently does not model only the conditional distribution but incorporates also the contribution of the marginal.
In the original formulation (Gershenfeld, 1997) this models has been proposed as a machine learning technique applied in the field of analysis and prediction of non-linear time series under linear and Gaussian assumptions in the context of social media technology, with the aim to build a digital violin (Schoner, 2000; Schoner and Gershenfeld, 2001). The CWM plans to write the joint distribution $\boldsymbol{X}$ and Y as

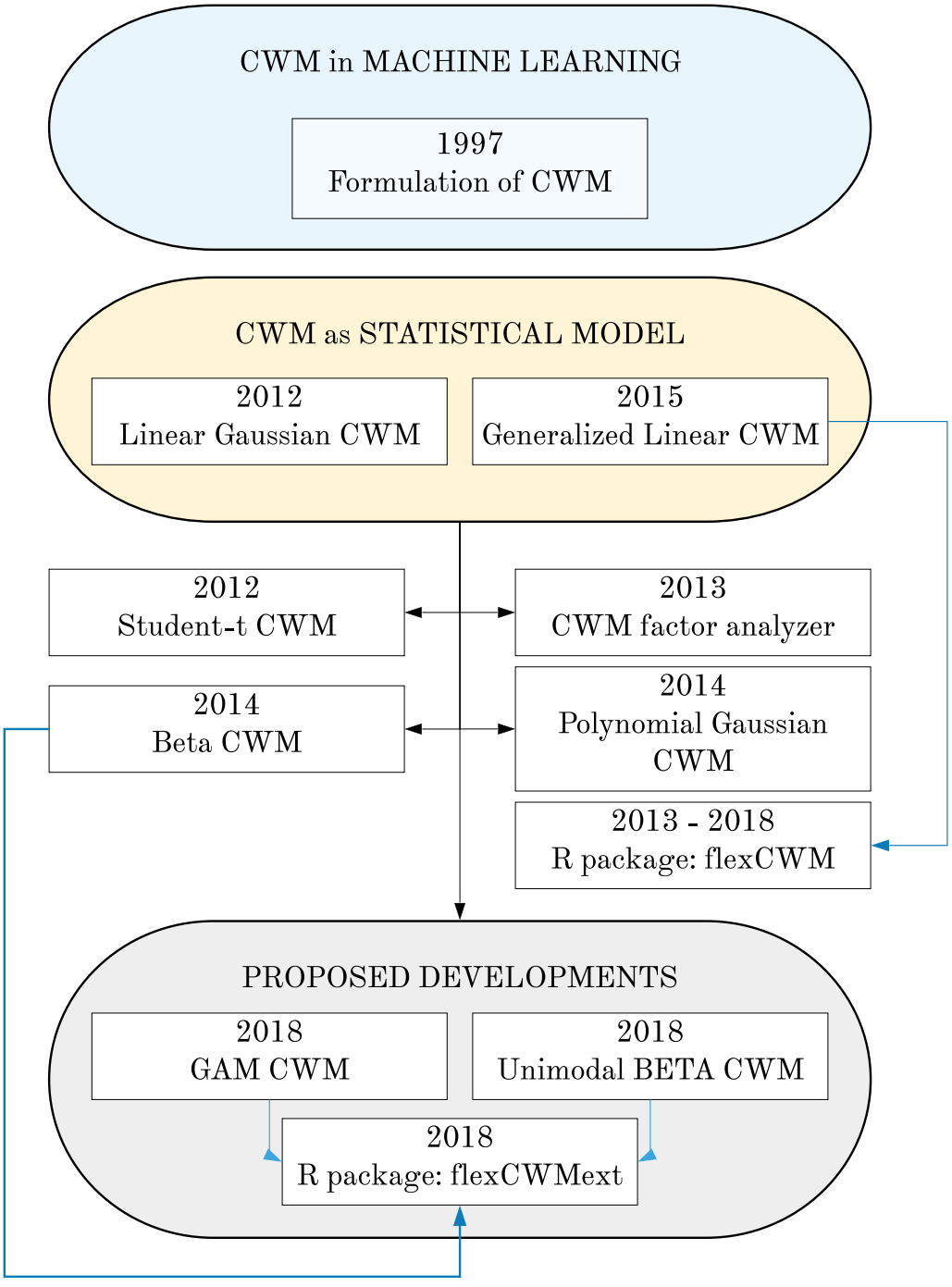
$$p(\boldsymbol{x}, y) = \sum_{g=1}^{G} \pi_g p(y|\boldsymbol{x}, \Omega_g) p(\boldsymbol{x}|\Omega_g) \tag{3}$$

defined on some space Ω that can be partitioned into G groups $G_1, ..., G_G$, where $p(y|\boldsymbol{x}, \Omega_g)$ is the conditional distribution of $Y|\boldsymbol{X} = \boldsymbol{x}$ in the group g and $p(\boldsymbol{x}|\Omega_g)$ is the marginal distribution of $\boldsymbol{X}$ in the group g. Starting from 2012, this framework has been developed from a statistical point of view, with significant contributions by prof. G. Vittadini, S. Ingrassia, A. Punzo and C. Minotti.
In Sect. 2 is summarized the history and the main evolution of CWM with the main references in the published literature, while in Sect. 3 the proposed developments in this work with an overview of the future developments.

2. History and evolution of Cluster Weighted Model

A first exploration of the CWM applied to a statistical problem can be found in Minotti and Vittadini, 2010 and Minotti, 2011 to compare institutional performances in multilevel modeling. Strictly related to this, is the need to control the numerosity of the population that has been highlighted in Vittadini *et al.,* 2006 some years before.
The basic idea of these words, is that fitting a single global model could not be sufficient to explain all the heterogeneity in the data and methods able to capture local behavior are necessary.
In multilevel modeling for example, the observed heterogeneity is expressed in terms of random intercepts and slopes, i.e. continuous latent variables that vary between clusters. However we should take into account also the unobserved heterogeneity that represents qualitatively different relationships that can be captured by latent variables (Muthén and Asparouhov, 2009). In such class of models is possible to obtain a ranking of the first-level random effects units based on confidence intervals (Goldstein, 2011), but the presence of high heterogeneity in first-level units leads to large and overlapped uncertainty intervals. Then, mixture models have an important role to play in multilevel regression analysis allowing the heterogeneity to be investigated more fully and more correctly attributing different portions of the heterogeneity to the different levels.
Coming back to the CWM, its first formulation in a statistical settings can be found in Ingrassia *et al.*, 2012, where has been shown that the CWM represents a very general class of mixture models that includes finite mixtures of distributions, finite mixtures of regressions and finite mixtures of regressions with concomitant variables. In Fig. 3 and in the following subsections are summarized the main developed extensions.

Figure 3: Main history of CWM with new developments.



2.1. Linear Gaussian CWM

The Linear Gaussian CWM (Ingrassia *et al.*, 2012) is the first proposal of CWM in a statistical setting. The model is defined as follows:

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \phi \left(y; \beta_{g,0} + \boldsymbol{\beta}_g^{'} \boldsymbol{x}, \sigma_{\epsilon,g}^2 \right) \phi_d \left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \right) \tag{4}$$

where $\phi(\cdot)$ denotes the probability density of Gaussian distributions. Both marginal and conditional densities are assumed to be Gaussian, where

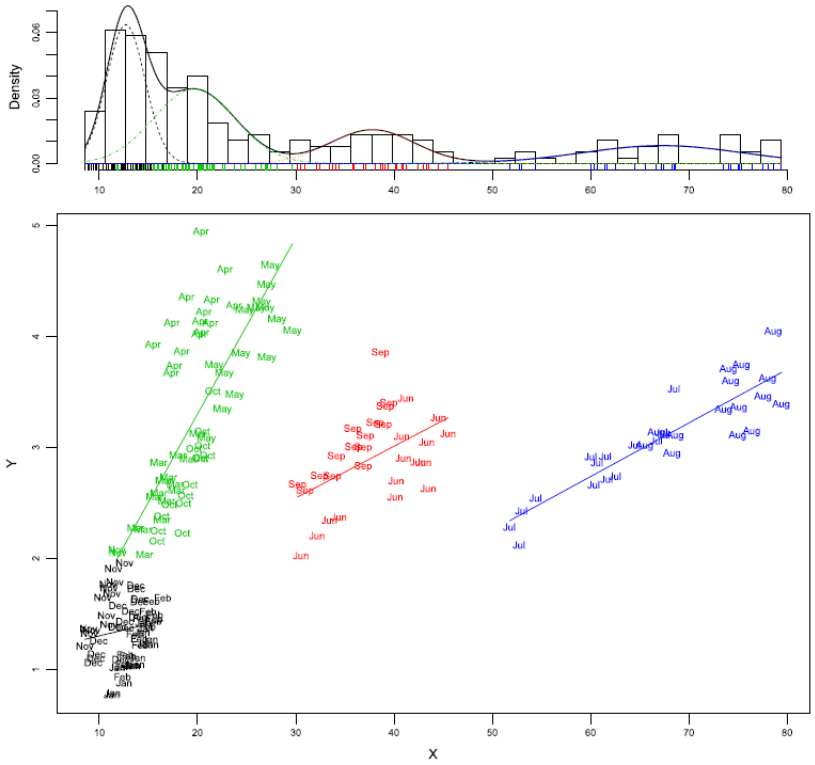
- $\boldsymbol{X}|\Omega_g \sim N_d \left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \right)$
- $Y|\boldsymbol{X} = \boldsymbol{x}, \Omega_g \sim N \left(\beta_{g,0} + \boldsymbol{\beta}_g^{'} \boldsymbol{x}, \sigma_{\epsilon,g}^2 \right)$

In case of classification problems each observation can be assigned to each group according to the maximum posterior probability given by

$$p(\Omega_g|\boldsymbol{x}, y) = \frac{\pi_g \phi \left(y; \beta_{g,0} + \boldsymbol{\beta}_g^{'} \boldsymbol{x}, \sigma_{\epsilon,g}^2 \right) \phi_d \left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \right)}{\sum_{g=1}^{G} \pi_g \phi \left(y; \beta_{g,0} + \boldsymbol{\beta}_g^{'} \boldsymbol{x}, \sigma_{\epsilon,g}^2 \right) \phi_d \left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \right)}$$

Has been proved that the linear Gaussian CWM is strictly related with traditional mixture models, and under suitable assumptions model (4) leads to the same posterior probability of many different mixture models, in particular finite mixtures of Gaussian distributions, finite mixtures of regression models, and finite mixtures of regressions with concomitant variables.
In Fig. 4 is shown an example of linear Gaussian CWM applied on real data; it is possible to see for each cluster the marginal density estimated from the model and the relation between the variables within each group.

Figure 4: Example of linear Gaussian CWM (Ingrassia *et. al.*, 2012) with $G = 4$ components applied on tourism data. The relationship between x and y is estimated within each mixture component thanks to the conditional part $Y|x$ while in the top of the figure is possible to see the marginal distribution of X estimated by the model.



2.2. Student-t CWM

The Student-t CWM (Ingrassia *et al.*, 2012), provide a more robust fitting for observations with heavier tails than normal or noise data. The model is defined as

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g t\left(y; \beta_{g,0} + \boldsymbol{\beta}_g^{'}\boldsymbol{x}, \sigma_{\epsilon,g}^2, \zeta_g\right) t_d\left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, v_g\right) \tag{5}$$

where $t_d(\cdot)$ denotes the d-dimensional probability density of the t-student distribution. Both marginal and conditional densities are assumed to be t distribution, where

- $\boldsymbol{X}|\Omega_g \sim t_d\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, v_g\right)$
- $Y|\boldsymbol{X} = \boldsymbol{x}, \Omega_g \sim t\left(\beta_{g,0} + \boldsymbol{\beta}_g^{'}\boldsymbol{x}, \sigma_{\epsilon,g}^2, \zeta_g\right)$

The posterior probability to belong to the group to the g-th group is given by

$$p(\Omega_g|\boldsymbol{x}, y) = \frac{\pi_g t\left(y; \beta_{g,0} + \boldsymbol{\beta}_g^{'}\boldsymbol{x}, \sigma^2_{\epsilon,g}, \zeta_g\right) t_d\left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, v_g\right)}{\sum_{g=1}^{G} \pi_g t\left(y; \beta_{g,0} + \boldsymbol{\beta}_g^{'}\boldsymbol{x}, \sigma^2_{\epsilon,g}, \zeta_g\right) t_d\left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, v_g\right)}$$

2.3. CWM factor analyzer

The applicability of linear Gaussian CWM in high dimensional matrix of covariates $\boldsymbol{X}$ can be a problem due to the number of parameters involved in the model: $(G-1) + G(p+2) + G\left[p + p(p+1)/2\right]$ where G is the number of latent groups and p the number of covariates. To overcome this problem the CWM factor analyzer (Subedi *et al.*, 2013; Subedi *et al.*, 2015), assumes a latent Gaussian factor structure for $\boldsymbol{X}$ in each mixture component.
The factor regression model lets to write the matrix of covariates $\boldsymbol{X}$ such that

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda U} + \boldsymbol{e}$$

where $\boldsymbol{U} \sim N_q(\boldsymbol{0}, \boldsymbol{I}_q)$ is a q-dimensional vector of latent factors, $\boldsymbol{\Lambda}$ a $p \times q$ matrix of factor loadings and $\boldsymbol{e} \sim N_p(\boldsymbol{0}, \boldsymbol{\Psi})$, with $\boldsymbol{\Psi} = diag(\psi_1^2, ..., \psi_p^2)$ independent of $\boldsymbol{U}$. Then $\boldsymbol{X} \sim N_p(\mu, \boldsymbol{\Lambda\Lambda}^{'} + \boldsymbol{\Psi})$ and the CWM factor analyzer can be written as

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \phi\left(y; \beta_{g,0} + \boldsymbol{\beta}_g^{'}\boldsymbol{x}, \sigma^2_{\epsilon,g}\right) \phi_p\left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Lambda_g\Lambda_g}^{'} + \boldsymbol{\Psi_g}\right) \tag{6}$$

2.4. Polynomial Gaussian CWM

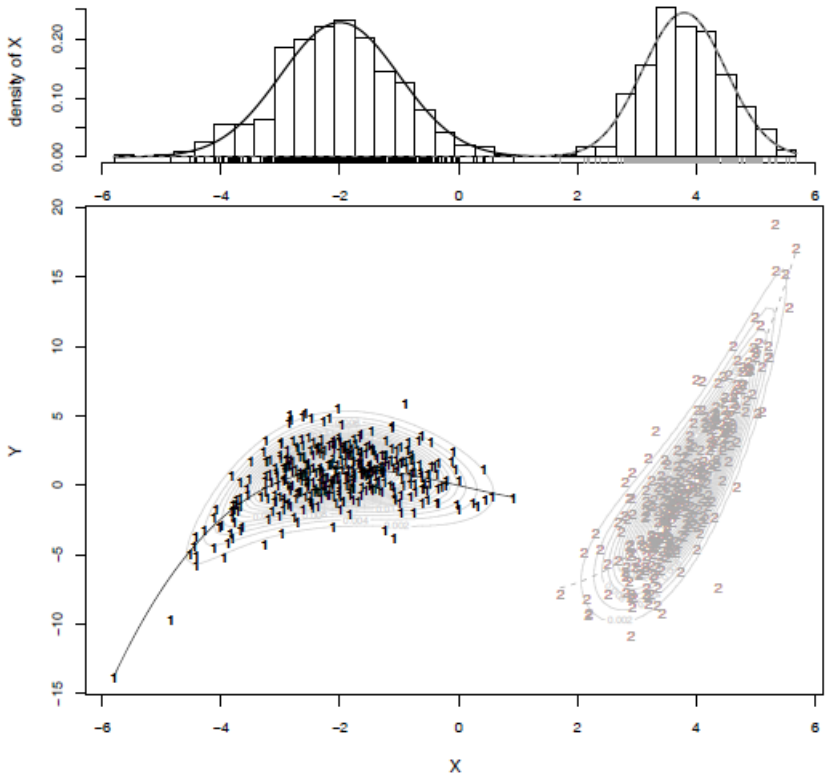
The polynomial Gaussian CWM (Punzo, 2014) allows to describe more complex dependencies in each mixture component by considering a polynomial regression. Actually this model is proposed in the bivariate case. The model is defined as

$$p(x, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \phi\left(y; \mu_r(x; \boldsymbol{\beta}_g), \sigma_g^2\right) \phi\left(x; \mu_g, \sigma^2_{x,g}\right) \tag{7}$$

where $\mu_r(x; \boldsymbol{\beta}_g) = \sum_{l=0}^{r} \beta_{lj} x^l$. As in linear Gaussian CWM both marginal and conditional densities are assumed to be Gaussian, then

- $X|\Omega_g \sim N\left(\mu_g, \sigma^2_{x,g}\right)$
- $Y|X = x, \Omega_g \sim N\left(\mu_r(x; \boldsymbol{\beta}_g), \sigma_g^2\right)$

Figure 5: Example of polynomial CWM (Punzo, 2014, Sect. 7.2) with $G = 2$ components. A flexible relationship between x and y is estimated within each mixture component. In the top of the figure is possible to see the marginal distribution of X estimated by the model.



2.5. Beta CWM

The beta CWM (Nieddu and Vitiello, 2014) brings mixtures of beta regression into the CWM framework. Lets consider the beta distribution parameterized in terms of the expected value μ and a dispersion parameter ϕ

$$f_{beta}(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\left((1-\mu)\phi\right)} y^{(\mu\phi-1)}(1-y)^{((1-\mu)\phi-1)} \tag{8}$$

The beta regression model is defined in each mixture component as

$$g(\mu_g) = \boldsymbol{x}^{'}\boldsymbol{\beta}_g \tag{9}$$

where $g(\cdot) : (0,1) \to \mathbb{R}$. After choosing the link function $g(\cdot)$ the CWM with beta components can be defined as

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g f_{beta}(y|\boldsymbol{x}, \mu_g, \phi_g)\phi_d\left(\boldsymbol{x}; \boldsymbol{\mu}_{x,g}, \boldsymbol{\Sigma}_{x,g}\right) \tag{10}$$

Note that the μ_g in $f_{beta}(y|\boldsymbol{x}, \mu_g, \phi_g)$ represents the expected value of the conditional beta distribution while $\boldsymbol{\mu}_{x,g}$ in $\phi_d\left(\boldsymbol{x}; \boldsymbol{\mu}_{x,g}, \boldsymbol{\Sigma}_{x,g}\right)$ is the mean vector of the d-dimensional Gaussian distribution for the marginal density of $\boldsymbol{X}$.

2.6. Generalized Linear CWM

The generalized linear cluster weighted model (Ingrassia *et al.*, 2015), introduces a new extension where the conditional distributions in each component belong to the exponential family; important sufficient condition for the identifiability are also prooved. The model is defined as

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g q\left(y|\boldsymbol{x}, \boldsymbol{\xi}_g\right) p\left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right) \tag{11}$$

where $q(\cdot)$ denotes the probability density of a distribution belonging to the exponential family.

3. Proposed extensions and future developments

Despite the numerous extensions proposed, some problems related to the CWM are still opened and there is a lot of work to do. In this work two main extensions are proposed to deepen and at the same time to explore the range of applications of this framework. The first considers the concept of non-parametric regression involving the theory of Generalized Additive Models (Hastie *et al.,* 1987; Wood, 2017) while the second explores the use of a proper subclass of beta regression applied to mixture models.
The GAM-CWM (Generalized Additive Models within the CWM framework) is a generalization of the linear Gaussian CWM and the polynomial Gaussian CWM, that lets to model in a flexible way the relation between the covariates and the variable of interest introducing smooth functions (i.e. splines).
Another proposed extension is related to the beta CWM. Although the beta CWM has already been proposed (Nieddu and Vitiello, 2014) an useful improvement plans to consider only a subset of the beta densities, in particular the subset of unimodal beta densities. A problem that could arise in such types of mixtures is due to the high flexibility of the beta distribution: when it is embedded in a mixture component it may be too flexible due to the great variety of shapes (including multi-modal shapes) so that it may be difficult to understand easily the real meaning of each component. Starting from this consideration, and supported by the work of Bagnato and Punzo, 2013 we explore the use of reparameterized beta in the context of mixtures of regressions.
Finally, in the following point are summarized those that are considered to be the most important themes to be explored in future works related to the proposed extensions or more generally on the CWM in a whole:

- Diagnostics: it is quite evident the need to identify and develop a set of instruments (including indices, statistics and graphic representations) to improve the understanding of the phenomenon under analysis through the CWM. For example, in case of classification problem, some indices could be useful to understand the behaviors of the defined clusters from different points of view.

- Big data: we need to study the behavior of the CWM with big data stressing not only the sample size factor but also the increasing number of the covariates. At the same time it is necessary to understand the robustness of the model splitting the dataset in training / validation.

- Model selection: although this theme has already been taken into consideration in numerous articles in the literature, we need to consider this topic into a Big data problem. The BIC criterion for example, does not consider

the performance of the model when the dataset is splitted in training and validation. Then, we want to explore this aspect related to the choice of a suitable model that show good performance on the validation data set and at the same time is parsimonious in the number of parameter controlling also for the overfitting.

- Applications: since most statistical tools have been developed to analyze many types of variables, it is necessary to focus on specific applications where the CWM could provide added value compared to existing models and therefore assert itself not only as a good theoretical tool but as a key tool for specific fields of applications.

- Software: CWM is an eminent member of the class of mixtures of regression models with random covariates, but unfortunately there is a lack of support in the most used software packages for statistical computing. Actually only the R package flexCWM (Mazza *et al.*, 2018) is available on CRAN to implement the generalized linear CWM. In this work we include an extension of flexCWM called flexCWMext (that will be published on CRAN in a few months) to provide a support for the new extensions that are proposed, in particular GAM-CWM and beta CWM. A future software implementation could include all the CWM extensions in a unique global and flexible environment and call C++ routines to minimize the computational time of the estimation process.

4. References

[1] Aurore, D., & Peter, H. (2010). Defining probability density for a distribution of random functions. The Annals of Statistics. 38(2), 1171-1193.

[2] Bagnato, L., & Punzo, A. (2013). Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. Computational Statistics, 28(4), 1571-1597.

[3] Battle, M. V., & Rakow, E. A. (1993). Zen and the art of reporting differences in data that are not statistically significant. IEEE Transactions on Professional Communication, 36(2), 75.

[4] Berta, P., Ingrassia, S., Punzo, A., & Vittadini, G. (2016). Multilevel cluster-weighted models for the evaluation of hospitals. Metron, 74(3).

[5] Boyd D. & Crawford K. (2012). Critical Questions For Big Data, Information, Communication & Society, 15(5), 662-679.

[6] Dayton, C. M., & Macready, G. B. (1988). Concomitant-Variable Latent-Class Models. Journal Of The American Statistical Association, 83(401).

[7] Everitt, B. S., & Hand, D. J. (1981). Finite Mixture Distributions. Dordrecht: Springer Netherlands, 1981.

[8] Gershenfeld, N. (1997). Nonlinear Inference and Cluster-Weighted Modeling. Annals Of The New York Academy Of Sciences.

[9] Gershenfeld, N. (1999). The Nature of Mathematical Modelling. Cambridge: Cambridge University Press.

[10] Gershenfeld, N., Schoner, B., & Metois, E. (1999). Cluster-weighted modelling for time-series analysis. Nature, 397(6717), 329.

[11] Goldstein, H. (2011). Multilevel statistical models. Hoboken, N.J.: Wiley, 2011.

[12] Grun, B. (2008). Fitting finite mixtures of linear mixed models with the EM algorithm. In P. Brito (Eds.), Compstat 2008 - International Conference on Computational Statistics (pp. 165-173). Germany: Springer.

[13] Hastie, T., & Tibshirani, R. (1987). Generalized Additive Models: Some Applications. Journal Of The American Statistical Association, 82(398), 371.

[14] Ingrassia, S., Minotti, S., & Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. Journal Of Classification, 29(3), 363-401.

[15] Ingrassia, S., Minotti, S. C., & Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. Computational Statistics and Data Analysis, 71, 159-182.

[16] Ingrassia, S., Punzo, A., Vittadini, G., & Minotti, S. (2015). The Generalized Linear Mixed Cluster-Weighted Model. Journal Of Classification, 32(1), 85-113.

[17] Ivanov, Todor & Izberovic, Sead & Korfiatis, Nikos. (2016). The Heterogeneity Paradigm in Big Data Architectures. (not published).

[18] Laney, Doug. (2001). 3-D Data Management: Controlling Data Volume, Velocity, and Variety. META Group Res Note 6.

[19] Mazza A., Punzo A., & Ingrassia S. (2018). flexCWM: A flexible framework for Cluster-Weighted Models. Journal of Statistical Software, 86(2).

[20] McLachlan, G. J., & Peel, D. (2000). Finite mixture models. New York: Wiley, 2000.

[21] Minotti S.C., Vittadini G. (2010). Local Multilevel Modeling for Comparisons of Institutional Performance. Data Analysis and Classification. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg.

[22] Minotti S.C. (2011). Some Notes on the Applicability of Cluster-Weighted Modeling in Effectiveness Studies. New Perspectives in Statistical Modeling and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg.

[23] Muthén, B., & Asparouhov, T. (2009). Multilevel regression mixture analysis. Journal Of The Royal Statistical Society: Series A (Statistics In Society), 172(3), 639-657.

[24] Ng, S. K., and McLachlan, G.J. (2008). Expert Networks with Mixed Continuous and Categorical Feature Variables: A Location Modeling Approach. Machine Learning Research Progress, eds. H. Peters and M. Vogel, New York: Hauppauge, 355-368.

[25] Nieddu L. & Vitiello C. (2014). Cluster Weighted Beta Regression. Rivista Italiana di Economia Demografia e Statistica.

[26] Punzo, A. (2014). Flexible mixture modeling with the polynomial Gaussian cluster-weighted model. Statistical Modeling: An International Journal, 14(3), 257-291.

[27] Royall, R. (1986). The Effect of Sample Size on the Meaning of Significance Tests. The American Statistician, 40(4), 313-315.

[28] Schoner, B., and Gershenfeld, N. (2001). Cluster Weighted Modeling: Probabilistic Time Series Prediction, Characterization, and Synthesis. Nonlinear Dynamics and Statistics, ed. A.I. Mees, Boston: Birkhauser, 365-385.

[29] Snijders, T. B., & Bosker, R. J. (2012). Multilevel analysis: an introduction to basic and advanced multilevel modeling. Los Angeles, Sage. 2012.

[30] Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2013). Clustering and Classification via Cluster-Weighted Factor Analyzers. Advances in Data Analysis and Classification., 7(1).

[31] Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2015). Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. Statistical Methods and Applications 24(4), 623-649.

[32] Vittadini, G., & Minotti, S. C. (2005). A methodology for measuring the relative effectiveness of healthcare services. IMA Journal Of Management Mathematics, 16(3), 239-254.

[33] Vittadini G., Sanarico M., & Berta P. (2006). Testing Procedures for Multilevel Models with Administrative Data. In Data Analysis, Classification & the Forward Search (p. 329).

[34] Wedel, M. (2002). Concomitant variables in finite mixture models. Statistica Neerlandica, 56(3), 362-375.

[35] Wedel, M. & DeSarbo, W. (1995). A Mixture Likelihood Approach for Generalized Linear Models. Journal of Classification, 12, 21-55.

[36] Wedel, M., & Kamakura, W. A. (2000). Market segmentation: conceptual and methodological foundations. Boston: Kluwer Academic, 2000.

[37] Wood, S. N. (2017). Generalized additive models: an introduction with R. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Part II
Generalized Additive Cluster Weighted Modeling

Generalized Additive Cluster Weighted Modeling

Abstract

An extension of mixture models with random covariates related to the linear Cluster Weighted Model (CWM) is presented for model-based clustering applications. The Generalized Additive Cluster Weighted Model (GAM-CWM) is a very flexible model, able to capture complex relations between a response variable and a set of covariates in each mixture component. Maximum likelihood estimates are provided via EM algorithm and a variant called *adaptive EM* is proposed to control the flexibility of the model during the estimation process. With simulated and real data we investigate performances, limits and benefits comparing this model with other mixture models related to it.

Keywords: Mixture Models, Model-based Clustering, EM Algorithm, Beta Distribution, Beta Regression, Cluster Weighted Model.

1. Introduction

In direct applications of statistical modeling, finite mixture models can be used for model-based clustering assuming that each mixture component represent a group in the original data, while in indirect application can be used as an alternative of non parametric density estimation (Titteringhton *et al.*, 1985, pp. 2-3 and McLachlan and Peel, 2000, p. 8). If no exogenous variables explain the means and variances of each component, we refer to unconditional mixtures called finite mixtures of distributions (Everitt and Hand, 1981; McLachlan and Peel, 2000), otherwise to the class of conditional mixture models considering finite mixtures of regression models and mixtures of generalized linear models (McLachlan and Peel, 2000); an extension of finite mixtures of regressions is called finite mixtures of regression models with concomitant variables (Wedel, 2002), where the weights of the mixture depend on a set of variables called concomitants.
In this article we focus on direct applications where, given a multivariate random vector $(Y, \boldsymbol{X})$, the interest is to study the functional dependency of Y on $\boldsymbol{X}$ within each mixture component.
A new class of mixtures of regression models initially proposed by Gershenfeld (1997) and subsequently taken up in a statistical setting by Ingrassia *et al.*, (2012) called CWM allows to model the joint probability of a response variable and a set of explanatory variables rather then only the conditional probability. The CWM,

modeling the joint probability, shows better performance than mixture models that consider only the conditional distribution, as well discussed and motivated in the developed extensions among which the generalized CWM (Ingrassia *et al.*, 2012, Ingrassia *et al.*, 2015, Mazza *et al.*, 2018), factor CWM (Subedi *et al.,* 2013; Subedi *et al.,* 2015). However, in many regressions mixture models as well as in the developed CWM extensions, is assumed a linear form for the covariates; this assumption in some applications may not be adequate and it would be wise to opt for a more flexible model that can better capture complex relationships between Y and $\boldsymbol{X}$.

Motivated by these considerations, the theory of Generalized Additive Model (GAM, Hastie and Tibshirani, 1987), that extends the generalized linear model precisely with the aim of making it more flexible including a sum of smooth functions of covariates can be grafted in the CWM framework. Some examples of mixture models including the concept of additive models can be found in Conversano *et al.,* 2002 applied in a data-mining context and in modeling time course gene expression data (Grün *et al.*, 2012).

A first attempt to extend the linear Gaussian CWM considering a polynomial model in each mixture component in order to deal with complex and nonlinear relationships can be found in Punzo, 2012 for the bi-variate case (with only one explicative variable). A polynomial model implies the definition of a polynomial basis for the covariate, but unfortunately such type of basis could be a problematic choice because as the dimension of the basis increases then the basis functions become collinear, leading to correlated parameter estimators and numerical problems (Wood and Augustiner, 2002). For these reasons, a natural generalization of CWM is proposed introducing a more flexible specification of the dependence of the response on the covariates considering a smooth function rather then a polynomial relationship, leading to a powerful and general class of models combining the principles of CWM model with the GAM.

An useful statistical tool based on the deviance decomposition is considered to measure the share of the deviance explained by the model with respect to the total deviance. A three terms deviance decomposition of the total sum of squares allows to investigate different aspects of the fitted model, in particular how much the model is able to explain the between-group deviance and how much the introduction of the covariates explains the variation in the dependent variable. We will show how the deviance decomposition represents a key tool that can be directly involved during the estimation of EM algorithm to improve the model's fitting.

This work is organized as follows. In section 2 some examples are presented to introduce this new approach from a qualitative and a graphical point of view; the main concepts of smooth functions and additive models are recalled in section 3 while the CWM is defined in section 4 with the main theoretical properties. The EM algorithm for the estimation of unknown parameter with some computational

details is explained in section 5. The three way deviance decomposition is explained in section 6 and finally in section 7 many applications with simulated and real data are discussed.

Table 1: Summary of the methodological context.

Model	Pros	Cons
GAM (Hastie and Tibshirani, 1986)	- Flexible models that includes smooth function.	- Inadequate in cases of heterogeneous data.
Linear CWM (S. Ingrassia *et al.*, 2012)	- Flexible and powerful mixture models. - Adequate in case of heterogeneous data.	
Polynomial CWM (Punzo, 2012)	- Flexible and powerful mixture models. - Adequate in case of heterogeneous data.	- Only polynomial extension is considered. - Collinearity problems. - Developed only for the bivariate case.
GAM-CWM	- Flexible and powerful mixture models. - Adequate in case of heterogeneous data.	

2. A motivating example

In this section a motivating example based on simulated data is provided before introducing the new model from a theoretical point of view. The aim of this section is to show some limitations of the competing models that have motivated the identification of new solutions.
An artificial data set is generated with $n = 600$ observations with parameters listed in Table 3. We consider the bivariate case where (Y, X) has joint probability distribution $p(x, y)$, X is a $(n \times 1)$ matrix, Y is a response variable with values in $\mathbb{R}$ and suppose that the space Ω can be partitioned into G disjoint groups $(\Omega_1, ..., \Omega_G)$ such that $\mathbf{\Omega} = \Omega_1 \cup ... \cup \Omega_G$.
The GAM is a model where the relation between x_i and y_i is modeled in a flexible way introducing a flexible function, for example a cubic regression splines function

$f(x_i)$ (Wahba, 1991 and Gu, 2013) so that

$$y_i = f(x_i) + \epsilon_i$$

with $\epsilon_i \sim N(0, \sigma^2)$.
However, in many situations where a source of heterogeneity is not directly observed in the data, it is possible to capture latent factors considering a finite mixture of regression models where observations are assumed to arise from unobserved groups in the population. For this purpose it can be considered a GAM-MIX model (a mixture of GAM, Grün *et al.*, 2012). Finally, the LIN-CWM is a mixture model related to the class of CWM, where the joint probability of $p(x_i, y_i; \boldsymbol{\theta})$ is modeled for each mixture component by multiplying the marginal group-density of x, $p(x_i|\Omega_g)$, with the conditional density $p(y_i|x_i, \Omega_g)$. In Table 2 are summarized the models considered in this example.

Table 2: Probabilistic definition of GAM, GAM-MIX and LIN-CWM.

Model definition	$p(y_i\|x_i, \Omega_g)$	$p(x_i\|\Omega_g)$
$p_{GAM}(y_i\|x_i; \boldsymbol{\theta})$	$N(f(x_i), \sigma^2)$	-
$p_{GAM-MIX}(y_i\|x_i; \boldsymbol{\theta}) = \sum_{g=1}^{G} p(y_i\|x_i, \Omega_g)\pi_g$	$N(f_g(x_i), \sigma_g^2)$	-
$p_{LIN-CWM}(x_i, y_i; \boldsymbol{\theta}) = \sum_{g=1}^{G} p(y_i\|x_i, \Omega_g)p(x_i\|\Omega_g)\pi_g$	$N(\beta_{0,g} + \beta_{1,g}x_i, \sigma_g^2)$	$N(\mu_g, \sigma_g^2)$

Table 3: Original parameters for the simulation by cluster.

Parameter	Cluster 1 (red)	Cluster 2 (green)	Cluster 3 (black)
π_g	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$\mu_{x,g}$	15	15	35
$\sigma_{x,g}$	1	3	5
$f_g(x)$	$20 + 2(x-15)^2$	$3sin(x-15)$	$-1-3(x-35)$
$\sigma_{y,g}$	2	2	2

As shown in Fig. 1, the 3 groups are well separated and present different shapes: cluster 1 (red) can be described with a function of degree 2 because has a parabolic shape, cluster 2 (green) has a sinusoidal shape while cluster 3 (black) can be described with a straight line.
To fit these models a standard EM algorithm has been implemented in R Code (R Core Team, 2011).
The GAM (Fig. 2, left), is able to describe the cluster 1 (black) while cluster 2 and 3, as they are overlapped, cannot be described in a proper way and it is quite evident the need to consider a mixture models. The GAM-MIX (Fig. 2, right), on the other hand, turns out to be excessively flexible and unable to describe correctly the shape of each cluster: the excessive flexibility of such class of models can

sometimes be a negative element, because if the fitting is not adequately controlled during the EM algorithm the risks is to converge towards local minimum points. One concrete possibility is to adopt a new strategy to fit the model controlling its flexibility.
The original idea that will be discussed plans to estimate a less flexible model for the first steps of the EM, at least until the clusters have been identified and in a second time it is possible to increase its flexibility to improve the fitting and to better capture local behaviors.
It immediate to notice how the LIN-CWM (Fig. 3) is able to separate the clusters. However, this model well describes the cluster 1 while seems not be able to capture local behaviors of cluster 2 and 3. Although the clusters have been properly outlined, these components cannot be described with a linear model; for this purpose we propose the GAM-CWM that should be enough flexible to identify and describe accurately the shape of each cluster.

Figure 1: Scatter plot of generated data. Cluster 1 in black, cluster 2 in red, cluster 3 in green.

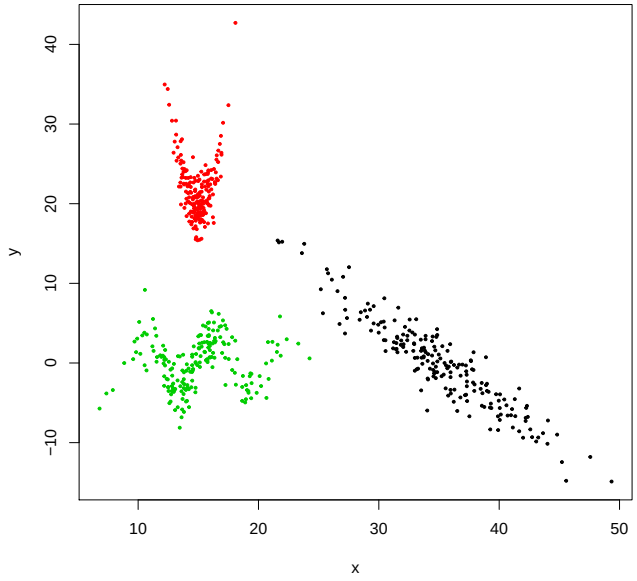


Figure 2: Estimated relationship for GAM and GAM-MIX.

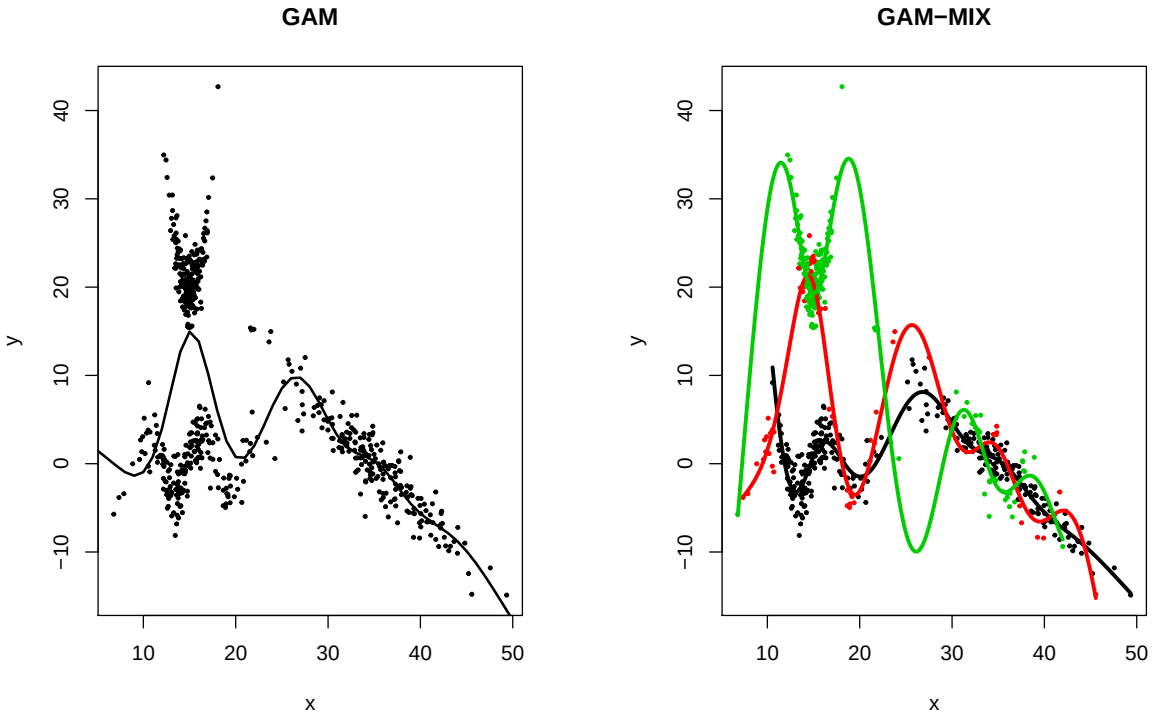
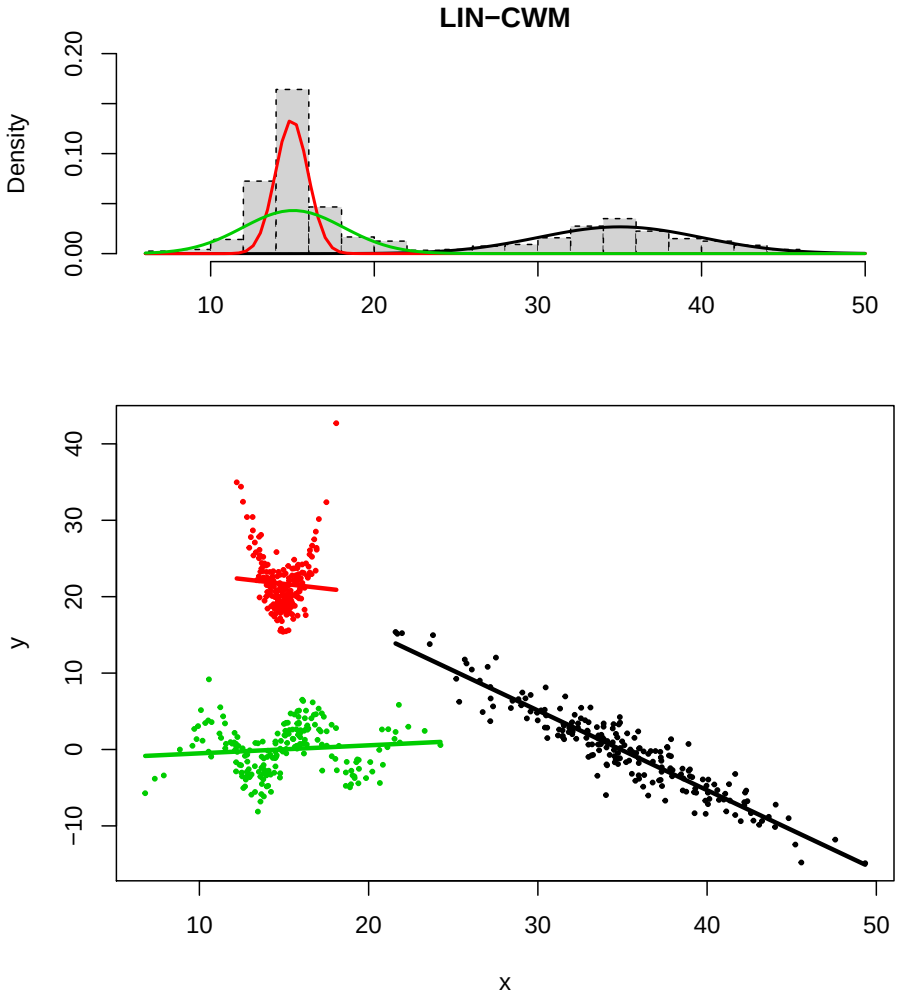


Figure 3: Estimated relationship for LIN-CWM.



3. Smooth functions and Generalized Additive Models

Lets consider the following model for the i-th observation ($i = 1, ..., n$):

$$y_i = f(x_i) + \epsilon_i \tag{1}$$

where y_i is the response variable, x_i a covariate, $\epsilon_i \sim N(0, \sigma^2)$ and $f(\cdot)$ a function with support $S \in \mathbb{R}$ to be defined in order to obtain a linear model. A common approach to define $f(\cdot)$ is to define linear combinations of a set of suitable functions called *basis*. Let $b_j(x)$ the j-th basis function. The function $f(\cdot)$ can be defined as

$$f(x) = \sum_{j=0}^{H} b_j(x)\beta_j \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, ..., \beta_H)$ are parameters to be estimated, $b_0(x) = 1$ (in order to consider the intercept β_0) and H is the basis dimension. For example if $f(x)$ is believed to be a 3rd order polynomial, a polynomial basis for the space can be defined setting $b_0(x) = 1$, $b_1(x) = x$, $b_2(x) = x^2$, $b_3(x) = x^3$ then $f(x) = \sum_{j=0}^{3} b_j(x)\beta_j = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3$ and the model (1) become $y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3 + \epsilon_i$. Cubic splines are an important class of splines composed by several sections of cubic polynomials joined together with continuous the first and the second derivatives. The points at which sections join are called *knots of the spline*. Polynomial basis are generally useful in the neighborhood of a specified point of the domain of the covariate but if the interest is in the whole domain these basis may have some problem (Wood, 2006, ch. 3, p. 120-126; Wahba, G. 1990). Once the basis has been defined the model design matrix (a $n \times H$ matrix) become

$$\boldsymbol{X}_H = \begin{bmatrix} 1 & b_1(x_1) & ... & b_H(x_1) \\ 1 & \vdots & \vdots & \vdots \\ 1 & b_1(x_n) & ... & b_H(x_n) \end{bmatrix}$$

The model (1) can be estimated minimizing $\sum_{i=1}^{n} (f(x_i) - y_i)^2$ using the least squares approach but in this way is not possible to control the degree of model's smoothing. Usually, to control the degree of smoothing, a penalized regression splines can be used to fit the model minimizing

$$||y - \boldsymbol{X}_H\boldsymbol{\beta}||^2 + \lambda \int_s f''(x)^2 dx \tag{3}$$

where s is the support of $f(\cdot)$ and λ controls the trade off between model fit and model smoothness (Fig. 4) that can be estimated with cross validation techniques.

The extension of model (1) with p explanatory variables $\boldsymbol{X} = [\mathbf{1}, x_1, ..., x_p]$ leads to an additive model that can be defined as

$$\begin{aligned} y_i &= f_1(x_{1,i}) + f_2(x_{2,i}) + ... + f_p(x_{p,i}) + \epsilon_i \\ &= \sum_{j=1}^{p} f_j(x_{j,i}) + \epsilon_i \end{aligned} \tag{4}$$

where $f_j(\cdot)$ is the smooth function referred to x_j and $\epsilon_i \sim N(0, \sigma^2)$. Supposing that the basis dimension H is fixed for all the p covariates then the vector of parameters become $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_p)$ with dimension $(1 \times Hp)$.
As explained before, the estimation can be performed extending (3) in the case with more than one explicative variable

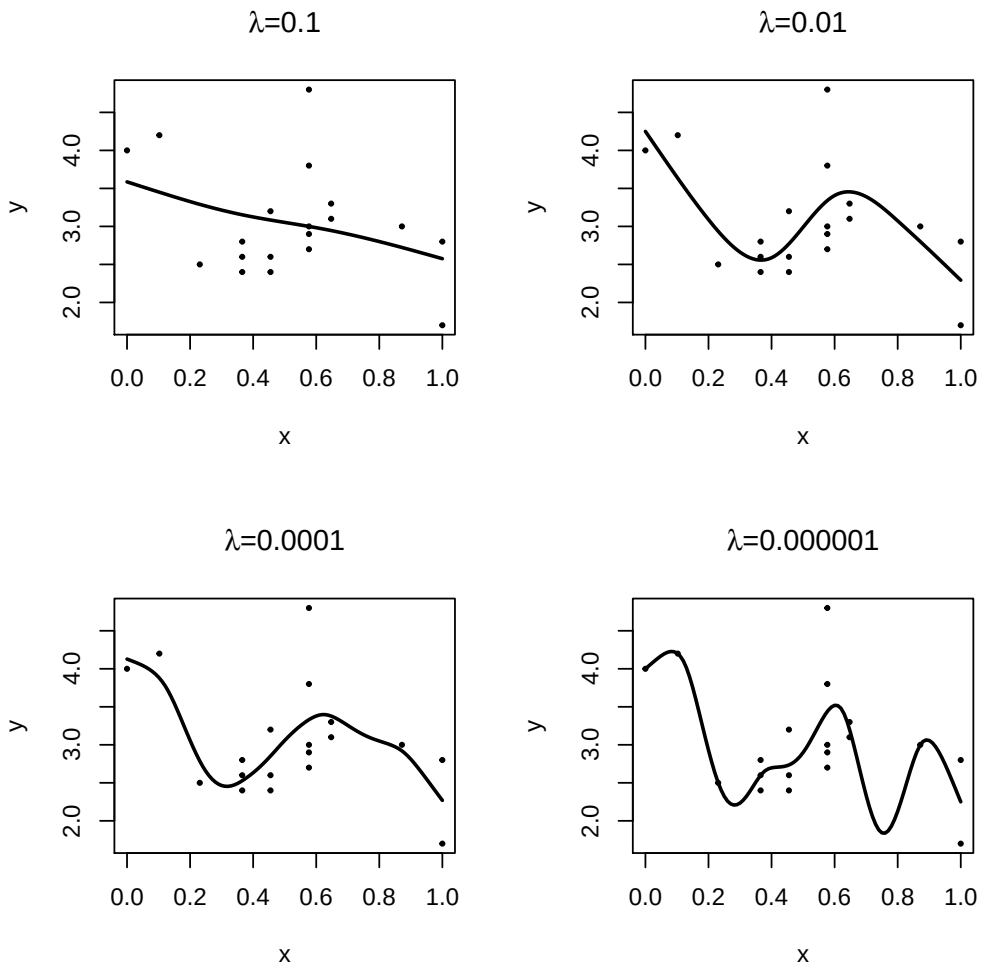
$$||y - \boldsymbol{X}_H\boldsymbol{\beta}||^2 + \sum_{j=1}^{p} \lambda_j \int_{s_j} f_j''(x_j)^2 dx \tag{5}$$

where the design matrix now become a $(n \times Hp)$ matrix defined as

$$\boldsymbol{X}' = \left[\begin{array}{cccccc} \boldsymbol{X}_{1,H} & ... & \boldsymbol{X}_{j,H} & ... & \boldsymbol{X}_{p,H} \end{array} \right]$$

with $\boldsymbol{X}_{j,H}$ the j-th model matrix referred to x_j given a basis.
Generally, model (4) is not an identifiable model, unless each smooth is subject to a centering constraint $\mathbf{1}'\boldsymbol{X}_{j,H}\boldsymbol{\beta}_j = 0$.
Finally, if the response variable comes from an exponential family distribution the theory of generalized additive models follows from additive models as generalized linear models follow from linear models, then the linear predictor $\sum_{j=1}^{p} f_j(x_{j,i})$ predicts some known smooth monotonic function of the expected value of the response.

Figure 4: Examples of penalized regression spline with four different values of the smoothing parameter λ.



4. Generalized Additive Cluster Weighted Model

In this section we define the GAM-CWM model. Let $\boldsymbol{X}$ is a $n \times p$ matrix of fixed covariates and Y is a response variable belonging to exponential family, with joint probability distribution $p(\boldsymbol{x}, y)$. Suppose that Ω can be partitioned into G disjoint groups $(\Omega_1, ..., \Omega_G)$ such that $\boldsymbol{\Omega} = \Omega_1 \cup ... \cup \Omega_G$. CWM models the joint probability $p(\boldsymbol{x}, y)$ as follows:

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} p(y|\boldsymbol{x}, \Omega_g) p(\boldsymbol{x}|\Omega_g) \pi_g \tag{6}$$

where $p(y|\boldsymbol{x}, \Omega_g)$ is the conditional density of the response Y given the predictors $\boldsymbol{X}$ and Ω_g, $p(\boldsymbol{x}|\Omega_g)$ is the probability density of $\boldsymbol{X} = \boldsymbol{x}$ given Ω_g and $\pi_g = p(\Omega_g)$ are the mixing weights of Ω_g so that $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$ and finally $\boldsymbol{\theta}$ is the set of all parameters in the model.
Lets focus now the attention on the conditional part of the model $p(y|\boldsymbol{x}, \Omega_g)$ where the GAM models are involved. In order to deal with various response types we assume that $p(y|\boldsymbol{x}, \Omega_g)$ belongs to the exponential family. A monotone and differentiable link function $h(\cdot)$ is introduced to relate $\mu_g = E(y|\boldsymbol{x}, \Omega_g)$ to the covariates through the relation

$$\begin{aligned} h(\mu_g) &= f_{1,g}(x_1) + ... + f_{p,g}(x_p) \\ &= \sum_{j=1}^{p} f_{j,g}(x_j) \end{aligned}$$

We recall that $f_{j,g}(x_j)$ depends on g because in the group g

$$f_{j,g}(x_j) = \sum_{h=1}^{H} b_h(x_j) \beta_{h,g}$$

The interest is in the parameters $\boldsymbol{\beta}$ so that the distribution of $Y|\boldsymbol{X} = \boldsymbol{x}, \Omega_g$ will be denoted with $q(y|\boldsymbol{x}, \boldsymbol{\beta}_g, \zeta_g)$ where the parameter ζ_g is an additional parameter to take into account when a distribution from a two-parameter exponential family is considered. Then, the generalized additive CWM can be defined as

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} q\left(y|\boldsymbol{x}, \boldsymbol{\beta}_g, \zeta_g\right) p(\boldsymbol{x}|\boldsymbol{\psi}_g) \pi_g$$

Note that the marginal distribution of $\boldsymbol{X}$, that depends on the type of the covariates involved in the model, is generally indicated with $p(\boldsymbol{x}|\boldsymbol{\psi}_g)$, where $\boldsymbol{\psi}_g$ include all the unknown parameters that have to be estimated.

In case of classification problem the posterior probability belonging to the g-th group can be calculated with the maximum posterior probability as

$$p(\Omega_g|\boldsymbol{x}, y) = \frac{p(\boldsymbol{x}, y, \Omega_g)}{p(\boldsymbol{x}, y)} = \frac{q\left(y|\boldsymbol{x}, \boldsymbol{\beta}_g, \zeta_g\right) p(\boldsymbol{x}|\boldsymbol{\psi}_g)\pi_g}{\sum_{g=1}^{G} q\left(y|\boldsymbol{x}, \boldsymbol{\beta}_g, \zeta_g\right) p(\boldsymbol{x}|\boldsymbol{\psi}_g)\pi_g}$$

5. Model Estimation

Given a sample of size n the model can be estimated with EM algorithm (Dempster *et al.,* 1977) to maximize the global log-likelihood function in order to obtain maximum likelihood estimates for the unknown parameters. By controlling for the basis dimension during the EM algorithm, it is possible to avoid the problem that arises in Fig. 2 (right), and as highlighted before, the concept of deviance decomposition is particularly useful for this purpose. The structure of the EM is not different from the standard implementation, because the term *adaptive* is referred to the value H (2) which determines only the size of the basis. Then we describe in this section the EM algorithm on the iteration $(k+1)$ and in section 6 how control the modification of this value during EM.
Let $\boldsymbol{z}_i$ a G-dimensional component label vector where the j-th element z_{ig} is one or zero if respectively the mixture component of $(x_i, y_i)'$ is equal to j or not. The log-likelihood for the defined model fixed G (the number of groups) is:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln(\pi_g) + \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln\left(q(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}_g, \zeta_g)\right) + \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln\left(p(\boldsymbol{x}_i|\boldsymbol{\psi}_g)\right) \end{aligned} \tag{7}$$

On the iteration $(k+1)$, the E-step requires the calculation of the conditional expectation of the random variable Z_{ig} related to each z_{ig} given the augmented data sample $\{(\boldsymbol{x}_1, y_1, \boldsymbol{z}_1), ..., (\boldsymbol{x}_n, y_n, \boldsymbol{z}_n)\}$. In particular, for $i = 1, ..., n$ and $g = 1, ..., G$ it follows that

$$\begin{aligned} E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ig}|(\boldsymbol{x}_i, y_i)\right] &= \frac{\pi_g^{(k)} q(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g^{(k)}, \zeta_g^{(k)}) p(\boldsymbol{x}_i|\boldsymbol{\psi}_g)}{\sum_{g=1}^{G} \pi_g^{(k)} q(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g^{(k)}, \zeta_g^{(k)}) p(\boldsymbol{x}_i|\boldsymbol{\psi}_g)} \\ &= \tau_{ig} \end{aligned} \tag{8}$$

which corresponds to the posterior probability that the observation $(\boldsymbol{x}_i, y_i)$ belongs to the g-th component of the mixture given the current $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)}, \zeta^{(k)}, \boldsymbol{\psi}^{(k)})$. During the M-step, on the iteration $(k+1)$, the conditional expectation of $l(\boldsymbol{\theta}^{(k)})$ given the observed data is maximized with respect to $\boldsymbol{\theta}$.

$$
\begin{aligned}
Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) \;=\; & \sum_{i=1}^{n}\sum_{g=1}^{G} \tau_{ig}^{(k)} \ln(\pi_g) + \\
& \sum_{i=1}^{n}\sum_{g=1}^{G} \tau_{ig}^{(k)} \ln\big(q(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g, \zeta_g)\big) + \\
& \sum_{i=1}^{n}\sum_{g=1}^{G} \tau_{ig}^{(k)} \ln\big(p(\boldsymbol{x}_i|\boldsymbol{\psi}_g)\big) +
\end{aligned} \tag{9}
$$

The maximization of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ with respect to the mixture weights is standard and can be obtained with Lagrangian multipliers as well as for the parameters $\boldsymbol{\psi}_g$ (Ingrassia *et al.*, 2015). Finally, the maximization of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\beta}_g$ and ζ_g is equivalent to the maximization problem of a generalized additive model, with the only difference that each observation contributed to the log-likelihood with a known weight $\tau_{ig}^{(k)}$ (Wood, 2017, ch. 3).

5.1. Other computational details

Code for the EM has written in R (R Development Core Team, 2011) while the GAM within each mixture component the function *gam()* of R package *mgcv* has been used. Main computational issues involve the initialization of parameters for the first step and the definition of a convergence criterion to stop the procedure.

- EM initialization: a standard initialization for EM is to defining starting values for the unknown vector of parameters $\boldsymbol{\psi}$. However, another approach (McLachlan and Peel, 2000; Punzo 2012) is to specify the values of $\boldsymbol{z}_{ig}$ for all the observation. A random initialization can be repeated many times from different random position selecting at the end the estimates that maximize the log-likelihood (Leisch, 2004).

- Convergence criterion: Aitken acceleration (Aitken, 1927) can be involved to take a decision about the convergence of the algorithm. It estimates the asymptotic maximum of the log-likelihood at each iteration:

$$
a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}}
$$

where $l^{(k)}$ is the log-likelihood value at iteration k. The asymptotic estimate of the likelihood (Böhning *et al.*, 1994) at iteration $k+1$ is given by

$$l_{\infty}^{(k+1)} = l^{(k)} + \frac{l^{(k+1)} - l^{(k)}}{1 - a^{(k)}}$$

In following simulations we stop the EM if $l_{\infty}^{(k+1)} - l^{(k)} < \epsilon = 0.05$.

- Model selection: among different model selection criteria we consider the Bayesian Information Criteria (BIC, Shwarz, 1978) and the Integrated Complete Likelihood (ICL, Biernacki *et al.,* 2000). In mixture models the BIC as a model selection criterion has been proposed by Disgupta and Riftery, 1998 and is defined as:

$$BIC = 2l(\hat{\boldsymbol{\psi}}) - \eta \ln(n)$$

where η is the number of free parameters included into the model. The ICL can be approximated by

$$ICL \approx BIC + \sum_{i=1}^{n}\sum_{g=1}^{G} MAP(\hat{z}_{ig}) \ln(\hat{z}_{ig})$$

where

$$MAP = \begin{cases} 1 & max\{z_{ig}\}\, occurs\, in\, component\, g \\ 0 & otherwise \end{cases}$$

6. Three way deviance decomposition and C-index

In this section we discuss a method that let to control the flexibility of the GAM-CWM during EM to avoid some situations such as those described in section 2 and at the same time we investigate how some indices (in particular the C-index) can be useful to evaluate the results from a clustering point of view.

6.1. Three way deviance decomposition

Although in penalized spline regression it is possible to control the degree of smoothing of the model with the parameter λ (see (5)) this is not enough when a GAM is included into a mixture component because it is necessary to control another important quantity that is the basis dimension (see the value H in (2)). The idea is that the value H can gradually (or directed) be increased during the EM algorithm described before, allowing the model to adapt to the clusters in the data. Controlling the variance decomposition, it is possible to take a decision if it is the time to increase the size of the basis.

Let $\hat{z}_{ig}^{(k)}$ the value of z_{ig} at the step k of EM algorithm; the total sum of squares of y at step k, say $TSS^{(k)}$, can be decomposed in the sum of two component: $WSS^{(k)}$ represents the within-groups deviance, while the $BSS^{(k)}$ is the between-groups deviance:

$$\begin{aligned} TSS^{(k)} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} (y_{ij} - \bar{y}_g)^2 + \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} (\bar{y}_g - \bar{y})^2 \\ &= WSS^{(k)} + BSS^{(k)} \end{aligned}$$

where

$$\bar{y}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig}^{(k)} y_i}{\sum_{i=1}^{n} \hat{z}_{ig}^{(k)}}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Denoting with $\hat{\boldsymbol{\beta}}_g^{(k)}$ the vector of estimates at step k and $h(\mu_g)$ the link function, the $WSS^{(k)}$ term can be decomposed again as

$$\begin{aligned} WSS^{(k)} &= \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} \left[y_i - h(\mu_{i,g}^{(k)}) + h(\mu_{i,g}^{(k)}) \right] - \bar{y}_g \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} \left[y_i - h(\mu_{i,g}^{(k)}) \right]^2 + \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} \left[h(\mu_{i,g}^{(k)}) - \bar{y}_g \right]^2 \\ &= WSS_f^{(k)} + WSS_e^{(k)} \end{aligned}$$

Summarizing, the total variability of Y can be explained by the latent group variable in BSS and the withing-group sum of squares WSS. In turn, the WSS can be decomposed into WSS_f predictable from the covariates and WSS_e not predictable from the covariates:

$$
\begin{aligned}
TSS^{(k)} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}^{(k)} (\bar{y}_g - \bar{y})^2 + \\
&+ \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}^{(k)} \left[y_i - h(\mu_{i,g}^{(k)}) \right]^2 \\
&+ \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}^{(k)} \left[h(\mu_{i,g}^{(k)}) - \bar{y}_g \right]^2
\end{aligned}
$$

The EM algorithm can be initialized with a low size of the basis dimension say H_0 and let k^* the step of the EM algorithm where $|BSS^{(k^*)} - BSS^{(k^*-1)}| < \epsilon$ (ϵ fixed sufficiently small). If this condition is verified, it means that the BSS is stabilized and the clusters have been identified by the model, because the between-group deviance does not change yet over the iterations. At this point, it is possible to increase the size of the basis to a new value $H_1 > H_0$ so that the model can specialize on the identified clusters and describe better local behaviors. An idea of this process is explained in Fig. 5 where it is possible to see that when BSS is stabilized around the 18th iteration (Fig. 5, column 3) is possible to increase the flexibility of the model on the detected clusters.

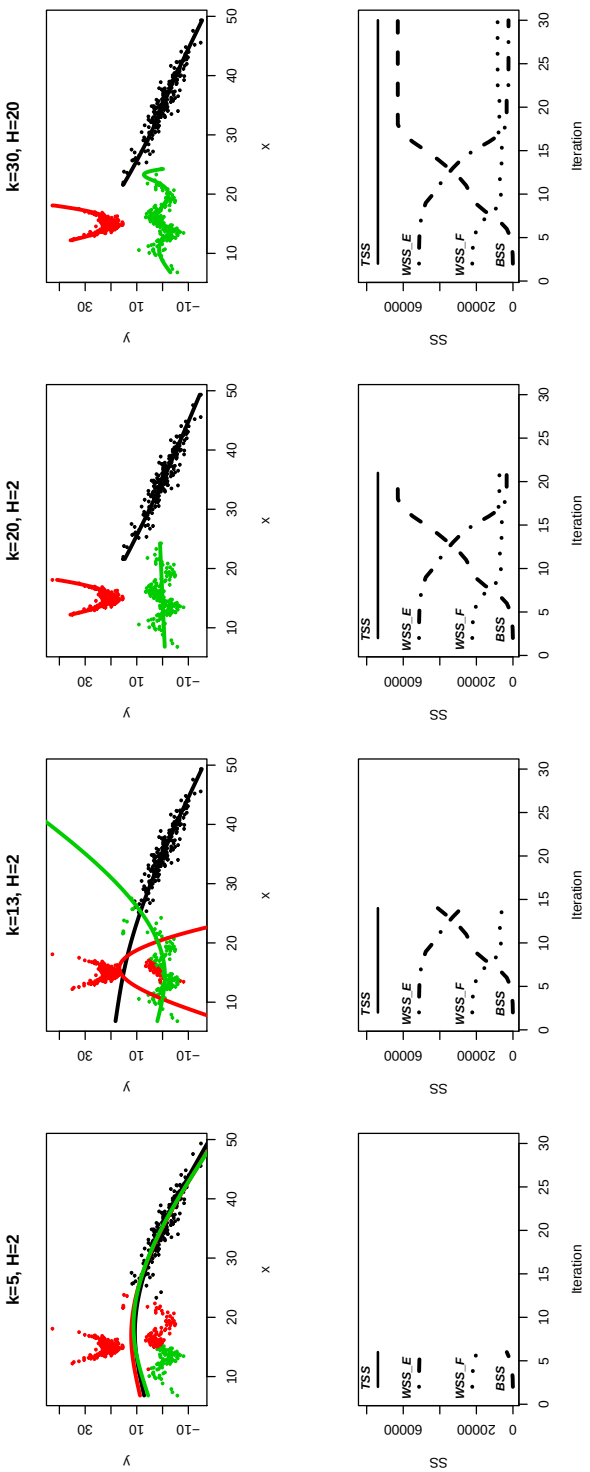


Figure 5: Graphical representation of EM algorithm at fixed steps (5, 13, 20, 30). In the part above the estimated model, while below the corresponding three way deviance decomposition. In particular, considering the BSS we see that it stabilizes around the 18th iteration, then the new value of $H = 20$ is set to increase the flexibility of the model.

6.2. C-Index to measure clusters compactness

Many indices to measure different cluster's behaviors have been developed; we investigate now, how a global index that describes the quality of clustering can be used operatively in the presented framework. Among these indices we selected the C-index (Hubert and Schultz, 1976), that is based on the Euclidean distances between the pairs of points inside each cluster. Let n_g the number of observation classified in the cluster g, in which there are $n_g(n_g-1)/2$ pairs of distinct points. Let N_W the number of such pairs ($N_W = \sum_{g=1}^{G} \frac{n_g(n_g-1)}{2}$) and let $N_T = n(n-1)/2$ the total number of pairs of distinct points in the whole data set. The C-index is defined as

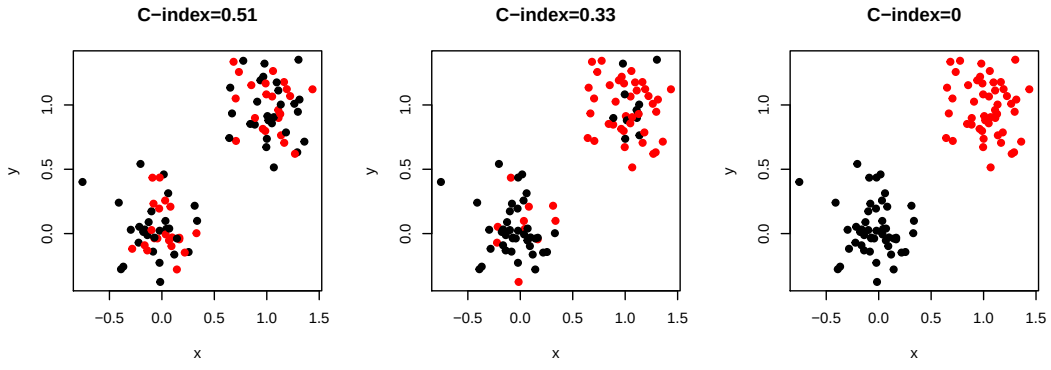
$$C = \frac{S - S_{min}}{S_{max} - S_{min}} \tag{10}$$

where S is the sum of the N_W distances between all pairs of points inside each cluster, S_{min} is the sum of the N_W smallest distances between all the N_T pairs of points and finally S_{max} is the sum of the N_W largest distances between all the N_T pairs of points.

The index is limited to the interval [0, 1] and should be minimized in order to obtain compact clusters. In Fig. 6 are calculated the values of C-index varying the assignment of the observations to the clusters.

Obviously, this index can be calculated considering the entire vector $(Y, \boldsymbol{X})$ of the covariates and the outcome (say $C_{\boldsymbol{X},Y}$) but we can obtain also two C-indices: one can describe the clusters from the point of view of the explicative covariates $\boldsymbol{X}$ (say $C_{\boldsymbol{X}}$) and the second from the point of view of the outcome variable Y (say C_Y). During the EM algorithm we can calculate $C_{\boldsymbol{X}}^{(k)}$ and $C_Y^{(k)}$ to evaluate how evolves clusters compactness. Finally, this method can also be applied to compare different models from a clustering point of view (Fig. 10).

Figure 6: C-index ($C_{\boldsymbol{X},Y}$) varying the assignment if the cluster.



7. Illustrative examples

In this section we illustrate, through simulations and real applications, the behaviors of the proposed model.

7.1. Recalling the motivating example

Recalling the motivating example explained in section 2 it is immediately to observe (Fig. 7, right) as the GAM-CWM is able to capture local behaviors, in particular the parabolic shape of cluster 2 (in red) and for the sinusoidal shape of cluster 3 (in green). In Table 4 the BIC criterion shows that this model is preferable compared to the LIN-CWM. However, at the same time, the three-way deviance decomposition provides important information on how the models are able to describe the clusters. In particular, the BSS is the same because the three groups are well separated and in both cases clusters have been well identified. On the contrary, the within-group sum of squares is different, and the difference is related to the better capacity of the GAM-CWM to capture local behaviors, bringing an improvement of the 60% in terms of WSS_f with respect to the LIN-CWM.

In Fig. 8 are plotted the smooth functions $f_g(x)$ estimated for each cluster and it is easy to recognize in the shape of these functions the functional relationship between x and y.

In Fig. 9 it is visualized the variation of the quantities that make up the TSS during the estimation algorithm. For LIN-CWM the EM converges around iteration 20. For GAM-CWM (Fig. 9, below) the estimation process starts with an upper bound of basis dimension fixed to $H_0 = 3$. Once the BSS stabilizes the upper bound is increased to $H_1 = 10$, allowing the WSS_f to increase.

Some empirical tests suggest starting the algorithm with a low upper bound ($H_0 = 3$ or $H_0 = 4$) and then to increase it to a higher value. Finally, in Fig. 10 the evolution of C-index is plotted during the EM algorithm for the three models. Clearly the CWM reaches the minimum value possible of $C_{\boldsymbol{X}}$ and C_Y while the mixture of GAM does not detect the clusters.

Table 4: Main descriptive statistics.

Measure	LIN-CWM	GAM-CWM
$LogLik$	-3 529.28	-3 325.49
BIC	-7 148.13	-6 894.07
TSS	73 855.62	73 855.62
BSS	85.28%	85.27%
WSS	14.72%	14.73%
WSS_f	7.32%	11.38%
WSS_e	7.40%	3.31%

Table 5: Coefficients of GAM-CWM by cluster.

Coefficient	Cluster 1	Cluster 2	Cluster 3
β_0	32.39	-163.42	4 946.79
$\beta_{1,g}$	-0.75	14.67	-302.34
$\beta_{2,g}$	-0.4	-107.49	683.08
$\beta_{3,g}$	3.6	-24.91	-637.81
$\beta_{4,g}$	-0.35	-114.81	810.77
$\beta_{5,g}$	1.63	-69.87	163.24
$\beta_{6,g}$	-0.75	-120.8	1 074.72
$\beta_{7,g}$	-2.93	-63.62	2 317.13
$\beta_{8,g}$	-8.43	-15.31	3 348.9
$\beta_{9,g}$	-9.92	-6.95	1 282.37
$\beta_{10,g}$	-11.1	129.21	-7 052.58

Figure 7: Comparison between LIN-CWM and GAM-CWM.

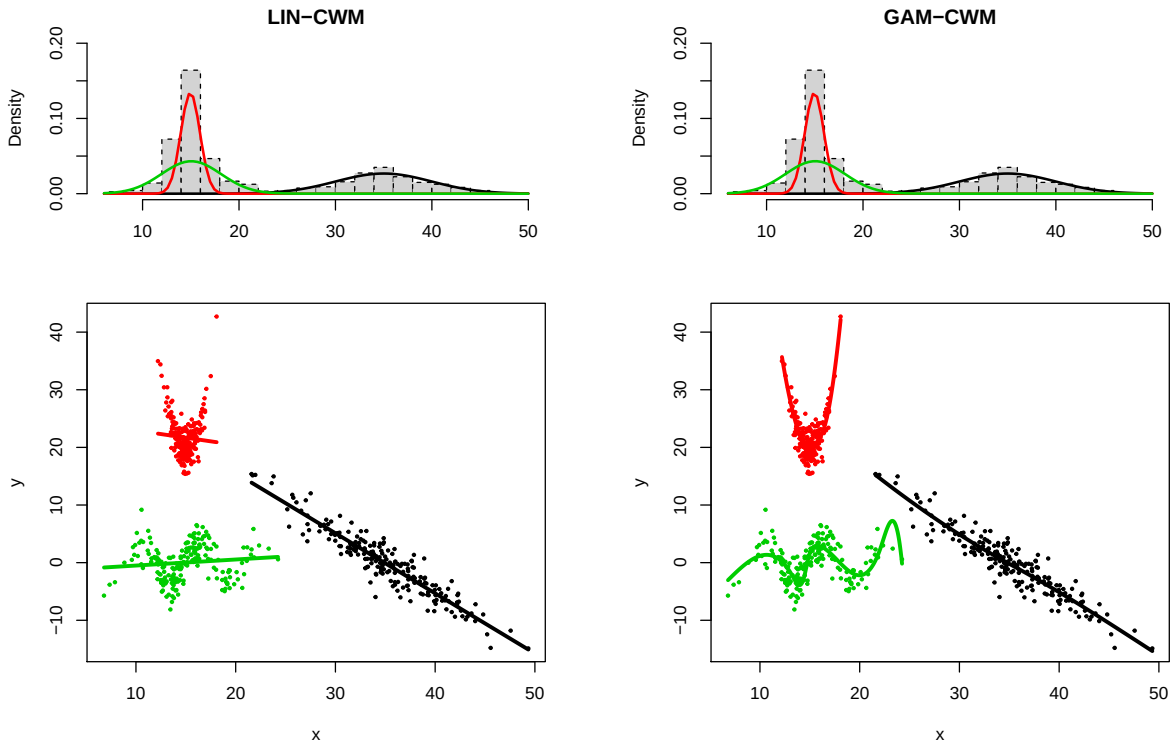


Figure 8: Smooth functions $f_g(x)$ estimated by the GAM-CWM for each cluster.

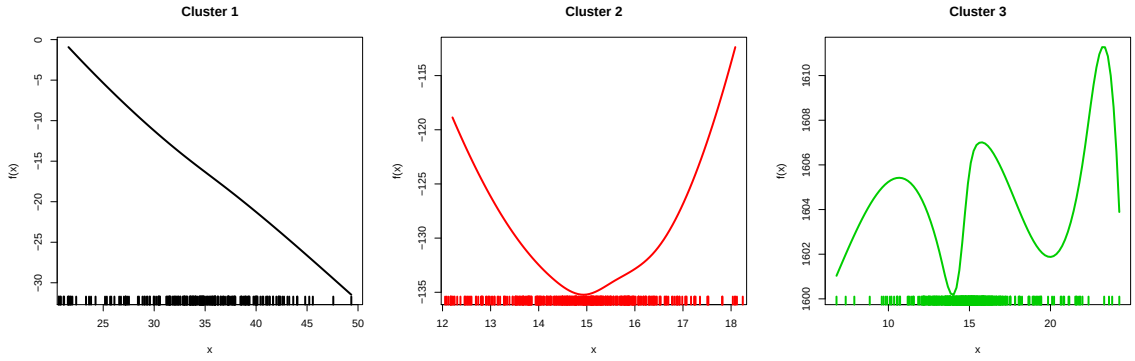


Figure 9: Three-way deviance decomposition for LIN-CWM and GAM-CWM. The BSS is the same for both models, while WSS_f of GAM-CWM is major than LIN-CWM, because GAM components are able to better describe local behaviors with respect to LIN-CWM. The vertical line in GAM-CWM it is in correspondence of the iteration of EM algorithm where the upper bound of the basis dimension increases (from 3 to 10).

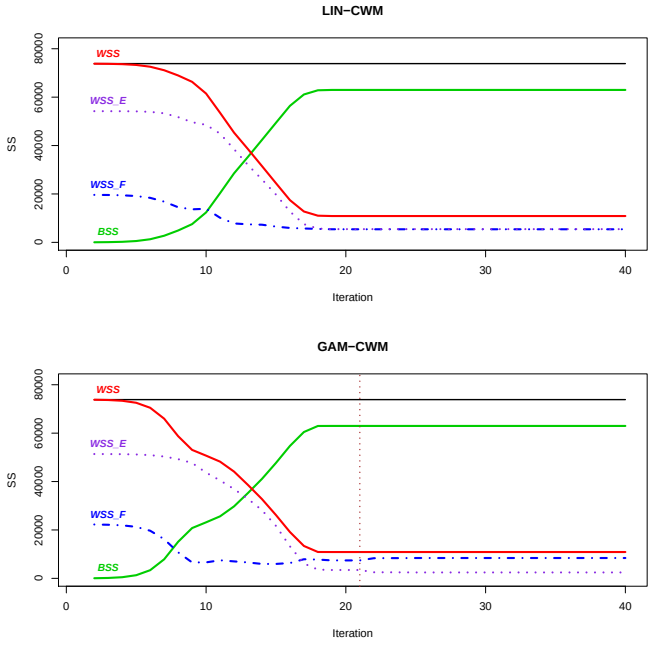
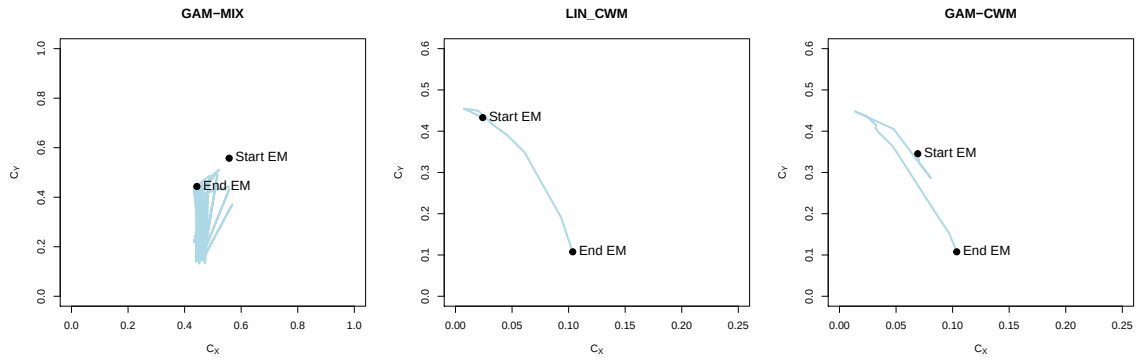


Figure 10: The evolution of C-index's values during the EM algorithm comparing the GAM-MIX, LIN-CWM and GAM-CWM.



7.2. Artificial data simulation to compare GAM mixtures

An artificial data simulation has been considered to compare three different mixture models: a GAM-CWM with the basis dimension fixed to three ($H = 3$), an adaptive GAM-CWM (with $H_0 = 3$ and $H_1 = 10$) and finally a LIN-CWM.
The process to generate data is described in Table 6. Cluster 1 (red) and cluster 3 (green) have a parabolic shape, while cluster 2 (black) has a sinusoidal shape. In Table 7 are synthesized the main measures to compare different models while in Table 8 are explained the true classification rates. In Fig. 11 are plotted the scatter plots and the estimated models varying the number of groups.
As can be seen in Fig. 11 (3rd column), the clusters are well identified in all the three cases. We observe how the GAM-CWM is better suited to the data than the model LIN-CWM. The three way deviance decomposition (Fig. 12, 3rd column) shows a very similar profile for the GAM with the basis dimension fixed to 3 and for the adaptive GAM where the basis dimension, after the stabilization of BSS if fixed to 10. The WSS_f is not different between these two models, meaning that the part of the deviance explained by the fitted models is similar (62.12% and 62.22%). On the contrary the LIN-CWM, as expected from the Fig. 11 (right, below), explains a lower quote of the WSS_f (58.54%).
It is also interesting to note the behaviors of the models estimated with a smaller number of clusters, in particular those with 2 clusters. In this case both GAMs are able to explain a greater share of variability within data; the WSS_f of GAM-CWM explains respectively 43.70% and 47.66% of TSS while the LIN-CWM only 8.48%.
It is straightforward to observe that according to the BIC criterion (Table 7) the best performing model is the adaptive GAM-CWM.
Finally is interesting to see how the true classification rates are very high for all the three estimated models.

Table 6: Parameter's definition for the artificial data simulation.

Parameter	Cluster 1 (red)	Cluster 2 (black)	Cluster 3 (green)
n	500	500	500
π_j	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
X	$Unif(15, 30)$	$N(15, \sigma = 3)$	$Unif(15, 30)$
$f_j(x)$	$20 + 2(x - 15)^2 + \epsilon$	$200 + 10\, sin(x - 25) + \epsilon$	$400 - (x - 15)^2 + \epsilon$
ϵ	$N(0, \sigma = 10)$	$N(0, \sigma = 10)$	$N(0, \sigma = 10)$

Table 7: Main statistics about the simulation.

Model	Measure	$g = 1$	$g = 2$	$g = 3$
Fixed GAM-CWM	$LogLik$	-13 716.9	-12 798.48	-11 455.47
	BIC	-27 470.37	-25 677.4	**-23 035.27**
	TSS	1 791 5271.64	17 915 271.64	17 915 271.64
	BSS	0%	35.35%	36.83%
	WSS	100%	64.65%	63.17%
	WSS_f	10.11%	43.70%	62.12%
	WSS_e	89.73%	20.91%	1.04%
Adaptive GAM-CWM	$LogLik$	-13 716.32	-12 668.14	-11 331.49
	BIC	-27 520.41	-25 519.12	**-22 940.89**
	TSS	17 915 271.64	17 915 271.64	17 915 271.64
	BSS	0%	35.85%	36.88%
	WSS	100%	64.15%	63.12%
	WSS_f	9.87%	47.66%	62.22%
	WSS_e	89.66%	16.12%	0.87%
LIN-CWM	$LogLik$	-13 729.5	-13 507.56	-12 312.5
	BIC	-27 488.25	-27 080.94	**-24 727.39**
	TSS	17 915 271.64	17 915 271.64	17 915 271.64
	BSS	0%	54.54%	36.32%
	WSS	100%	45.46%	63.68%
	WSS_f	8.75%	8.48%	58.54%
	WSS_e	91.25%	36.98%	5.14%

Table 8: True classification rates.

Model	TCR
GAM-CWM	0.980
Flexible GAM-CWM	0.986
LIN-CWM	0.969

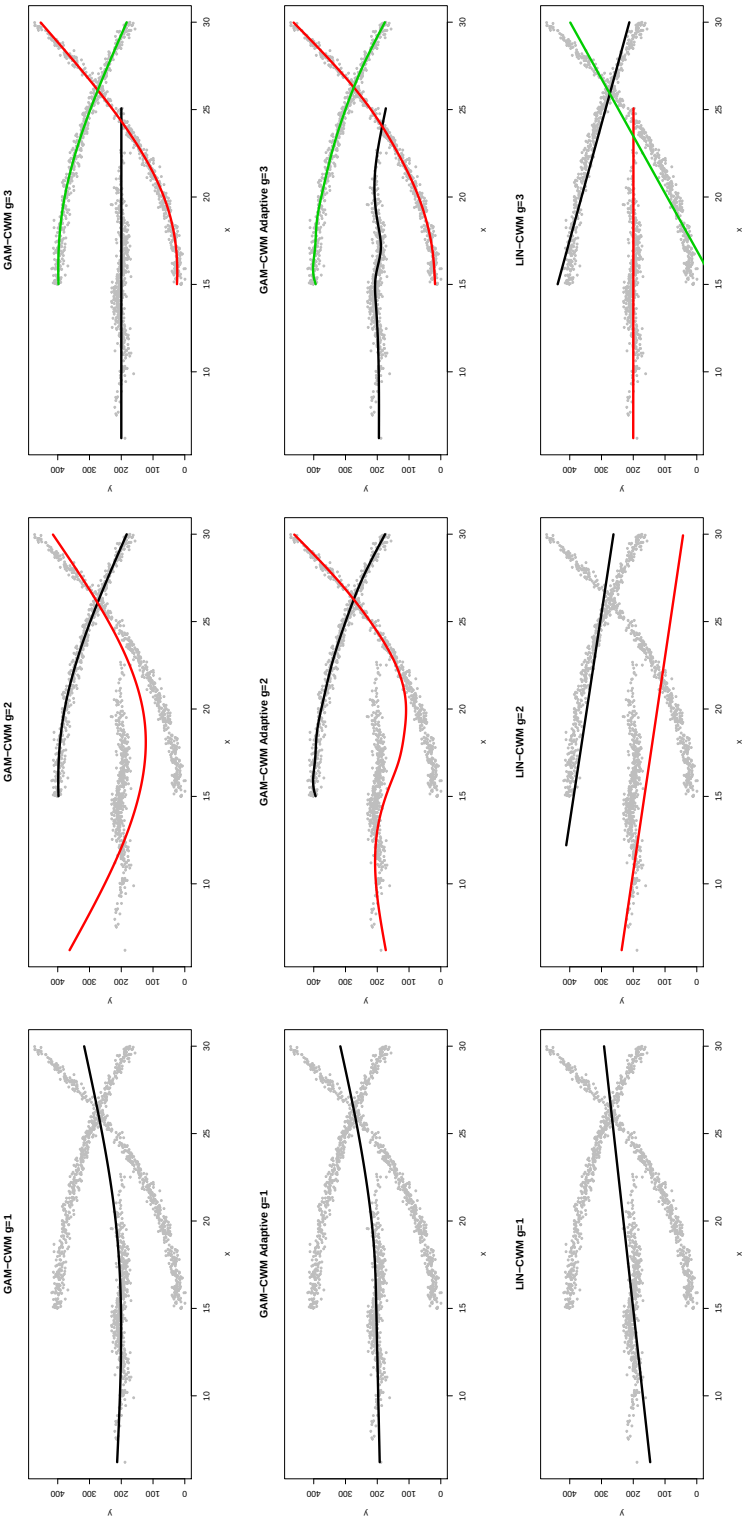


Figure 11: Fitted models for artificial data simulation varying the model (by row) and the number of clusters (by column).

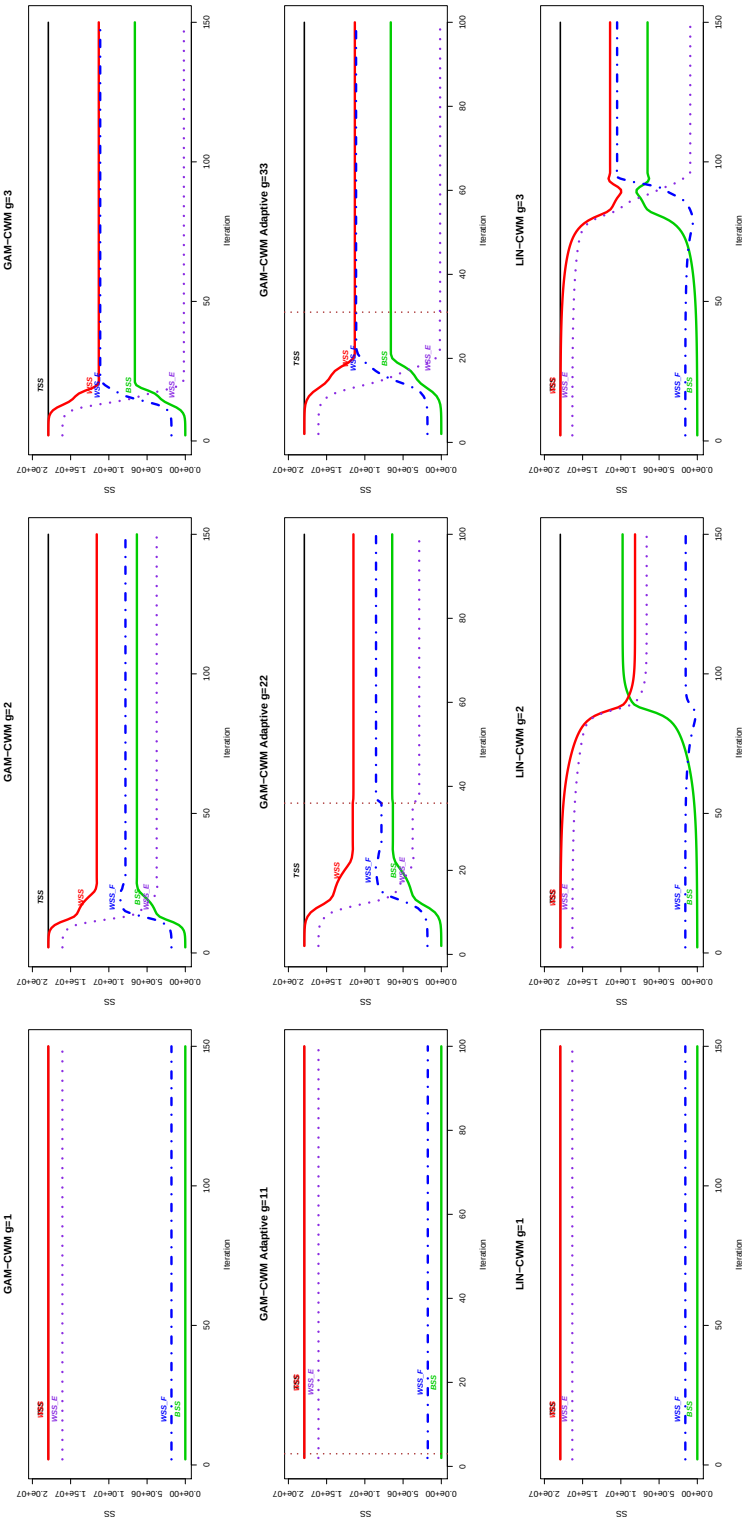
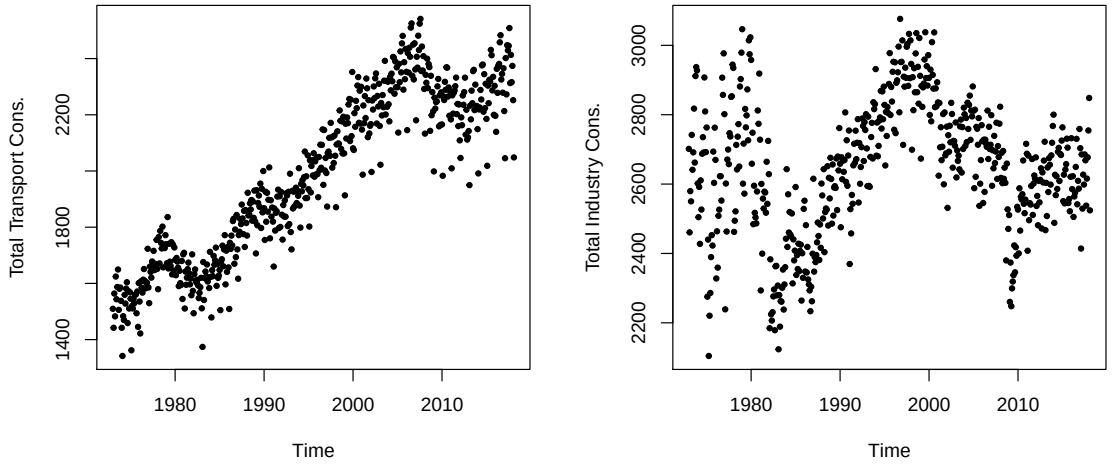


Figure 12: Three way deviance decomposition for each estimated model.

7.3. Monthly Energy Consumption by Sector

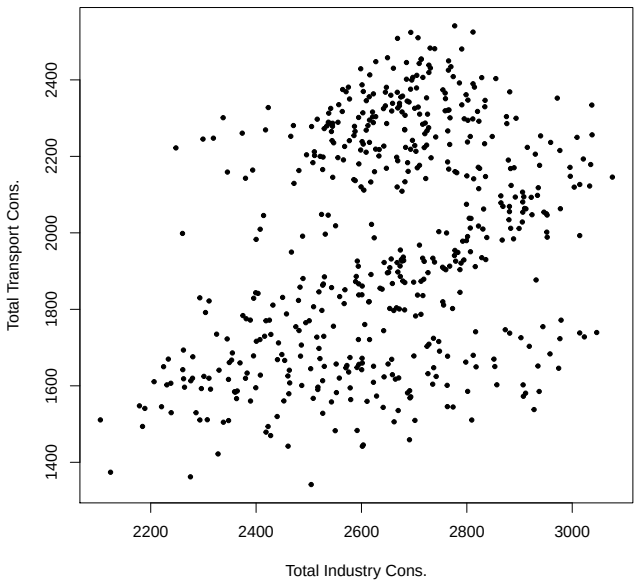
This application focuses on monthly data available from January 1973 to February 2018 on U.S. primary and total energy consumption by sector (residential, commercial, industrial and transportation). The energy consumption is measured in BTU (British Thermal Unit). In this application a GAM-CWM is estimated considering as response variable the total electric consumption in transportation sector (say y) and the industrial consumption is taken as covariate (say x).
Graphical descriptive statistics are available in Fig. 13 and Fig. 14. The growing consumer demand in transportation is quite evident from 1980 to 2008 while the industrial consumption shows a slight deflection starting from the year 2000, with an evident negative peak in correspondence of the global crisis of during 2008-2010. The scatter plot in Fig. 14 shows the data considered to estimate the GAM-CWM with cubic regression splines components.

Figure 13: Monthly energy consumption in transportation and industry from 1973 to 2018.



We consider both the case with two clusters and three clusters, as three clusters are selected according to the BIC criterion while the ICL suggests choosing 2 groups (Table 9). The three way deviance decomposition (Fig. 15) and the evolution of C-index during the EM algorithm (Fig. 16) show that the clusters are separated along the y-axis and consequently the BSS is very high compared to the WSS. The BSS of the model with three groups is about 10% higher than the model with two groups (Table 10).

Figure 14: Scatter plot of monthly observations of industry and transportation consumption.



The C-index of the model with 3 groups has a lower value ($C_X = 0.4$) with respect to the model with 2 groups ($C_X = 0.45$), therefore we can conclude that the clusters identified by the model with 3 groups are more compact from the point of view of x-axis.
This index, together with BIC, allows to justify the choice to consider three groups instead of two.
Finally, in Table 11 are summarized the coefficients estimated from the model with $H_1 = 10$ while in Fig. 17 we can see the estimated relation between x and y from January 1973 to February 2018 with 2 and 3 groups.
From an econometric point of view it seems reasonable to consider 3 groups, as at the same levels of industrial consumption, consumption in transports shows 3 patterns at 3 different levels. In Fig. 18 we can appreciate how each cluster is related to a specific historic period: cluster 2 (in red) it is between 1973 and 1990, cluster 1 (in black) between 1990 and 2000 while cluster 3 (in green) between 2000 and 2018. Moreover, it is interesting to note that in the cluster green (which represents the most recent years) in correspondence with high levels of energy consumption in the industrial sector corresponds to a decrease in consumption in the transport sector. We could interpret this decrease as the result of policies aimed to reduce the environmental impact and to increase the use of efficient means of transport.

Figure 15: Three way deviance decomposition evolution during EM convergence of the model with two groups (on the left) and the model with three groups (on the right). The clusters are well separated with respect to the outcome so that the BSS is very high compared to the WSS.

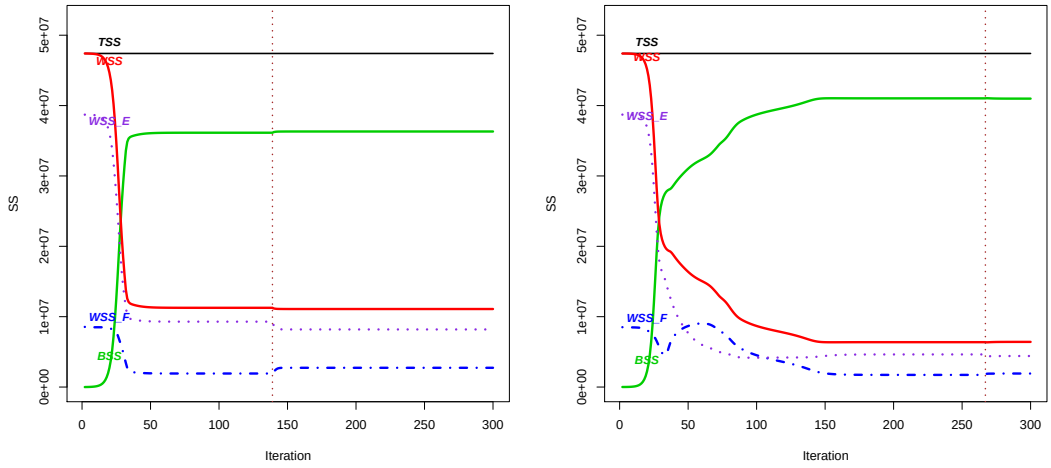


Figure 16: C-index evolution during EM with three different initialization. The paths are different but as we can see, at the convergence of EM the clusters have the same structure. On the left the model with 2 groups, on the right the model with 3. As expected from the scatter plot and from the three way deviance decomposition the clusters are separated along the y-axis and then the evolution of C-index highlights this aspect.

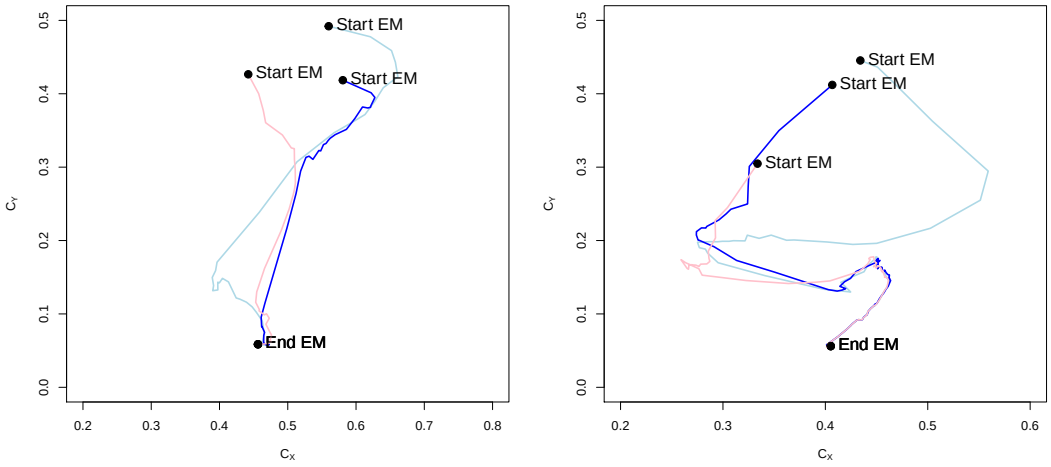


Table 9: BIC and ICL varying the number of groups.

Number of groups	BIC	ICL
1	-14 888.8	-14 888.8
2	-14 723.64	**-14 735.34**
3	**-14 714.73**	-14 750.05
4	-14 790.88	-14 827.61
5	-14 841.49	-14 939.33

Table 10: Three-way deviance decomposition with two and three groups.

Metric	2 groups	3 groups
TSS	47 404 508	47 404 508
BSS	76.61%	86.46%
WSS	23.39%	13.54%
WSS_f	5.77%	4.05%
WSS_e	17.27%	9.31%

Table 11: Coefficients of GAM-CWM with three groups.

Coefficient	Cluster 1	Cluster 2	Cluster 3
β_0	1902.87	1642.79	2245.66
$\beta_{1,g}$	-80.63	57.95	-64.08
$\beta_{2,g}$	-60.43	21.91	-37.34
$\beta_{3,g}$	-24.99	36.33	26.74
$\beta_{4,g}$	-13.9	25.64	49.45
$\beta_{5,g}$	3.61	43.55	79.78
$\beta_{6,g}$	20.72	54.34	94.64
$\beta_{7,g}$	68.1	52.42	63.25
$\beta_{8,g}$	153.88	87.29	-36.13
$\beta_{9,g}$	273.55	105.36	-31.26
$\beta_{10,g}$	-11.1	129.21	-7052.58

Figure 17: Marginal density of x on top and the estimated relation for each cluster with two and three groups (respectively on the left and on the right).

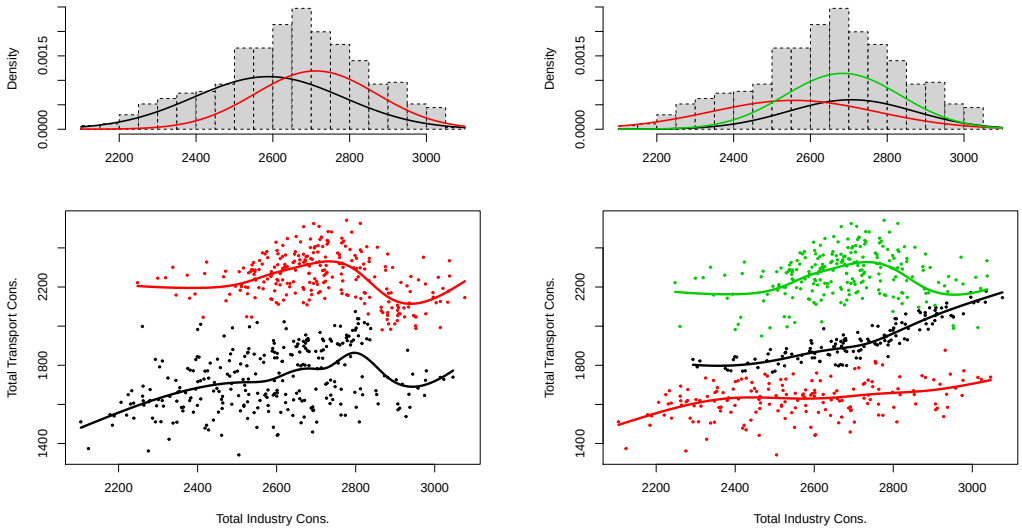
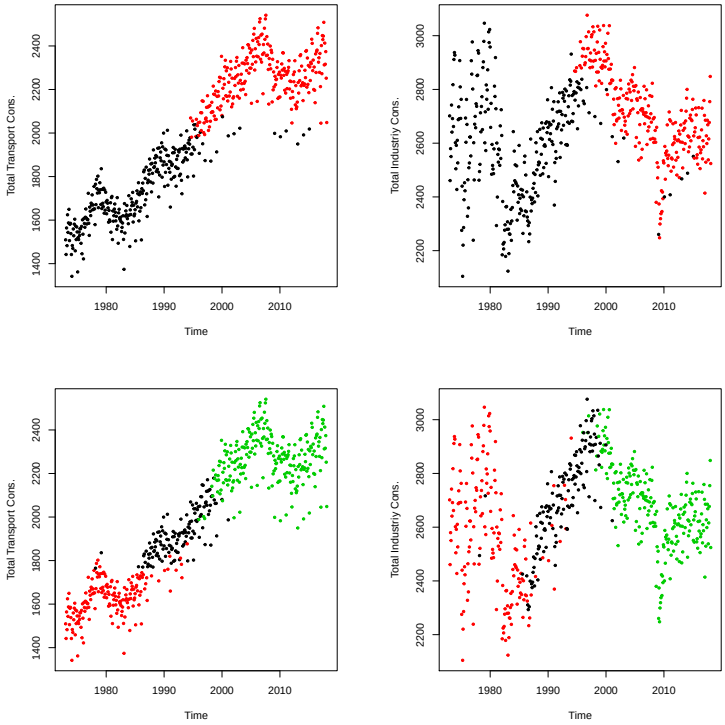


Figure 18: Transportation and industrial consumption during time with the clusters highlighted from the model.



8. Concluding remarks

An extension of the CWM model has been developed in order to deal with complex and non linear relationship between outcome and covariates within each mixture component introducing a GAM. The model has been introduced in terms of density estimation and applied to simulated and real data. Parameters estimation can be achieved with a modification of EM algorithm introducing the three way deviance decomposition as key instrument to avoid the convergence towards local minimum and to choose when the model can specialize on data in order to capture local behaviors. The applications on real and artificial data set show a clear added value in terms of classification and interpretation of the results.
Regarding future research we will focus in some directions. First of all it is important to identify appropriate fields of application where this methodology can be applied for example in the environmental field where the presence of latent factor and the need to develop flexible models are combined together. Moreover, it is opportune to examine the theme of the choice of the sensitivity of the model and at the same time evaluating the applicability of these models in the presence of large number of covariates.

9. References

[1] Aitken, A. C. (1926). On Bernoulli's Numerical Solution of Algebraic Equations. Proceedings of the Royal Society of Edinburgh, 46, 289-305. Royal Society of Edinburgh Scotland Foundation.

[2] Biernacki, C., & Celeux, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. IEEE Transactions on Pattern Analysis & Machine Intelligence, 22(7).

[3] Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., & Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Annals Of The Institute Of Statistical Mathematics, 46(2).

[4] Conversano, C., Siciliano, R., & Mola, F. (2002). Generalized additive multimixture model for data mining. Computational Statistics And Data Analysis, 38 (Nonlinear Methods and Data Mining), 487-500.

[5] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society. Series B (Methodological), 39(1), 1-38.

[6] Dasgupta, A., & Raftery A. E. (1998). Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering. Journal of the American Statistical Association, (441), 294.

[7] Everitt, B. S., & Hand, D. J. (1981). Finite Mixture Distributions. Dordrecht: Springer Netherlands, 1981.

[8] Gershenfeld, N. (1997). Nonlinear Inference and Cluster-Weighted Modeling. Annals Of The New York Academy Of Sciences, 808(1), 18.

[9] Grün, B., Scharl, T., & Leisch, F. (2012). Modelling time course gene expression data with finite mixtures of linear additive models. Bioinformatics (Oxford, England), 28(2), 222-228.

[10] Gu, C. (2013). Smoothing Spline ANOVA Models. New York, Springer.

[11] Hastie, T., & Tibshirani, R. (1987). Generalized Additive Models: Some Applications. Journal Of The American Statistical Association, 82(398).

[12] Hennig, C. (2000). Identifiability of Models for Clusterwise Linear Regression. Journal of Classification, 17(2), 273.

[13] Hubert, L., & Schultz, J. (1976). Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychology, 29, 190-241.

[14] Ingrassia, S., Minotti, S., & Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. Journal of Classification, 29(3), 363-401.

[15] Ingrassia S., Minotti S.C., Punzo A, (2014), Model-based clustering via linear cluster-weighted models. Computational Statistics and Data Analysis, 71(4): 159-182.

[16] Ingrassia, S., Punzo, A., Vittadini, G., & Minotti, S. (2015). The Generalized Linear Mixed Cluster-Weighted Model. Journal Of Classification, 32(1), 85-113.

[17] Leisch, F. (2004). FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. Journal of Statistical Software, 11(8), 1-18.

[18] Mazza A., Punzo A., & Ingrassia S. (2018). flexCWM: A Flexible Framework for Cluster-Weighted Models. Journal of Statistical Software, 86(2). 1-30.

[19] McLachlan, G. J., & Peel, D. (2000). Finite mixture models. New York: Wiley.

[20] R Development Core Team (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria. The R Foundation for Statistical Computing.

[21] Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, (2), 461.

[22] Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2013). Clustering and Classification via Cluster-Weighted Factor Analyzers. Advances in Data Analysis and Classification., 7(1).

[23] Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2015). Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. Statistical Methods and Applications 24(4), 623-649.

[24] Titterington, A. F. M., & Smith, U. E. Makov (1987). Statistical Analysis of Finite Mixture Distributions. Journal Of The American Statistical Association, (398), 694.

[25] Wahba, G. (1991). Spline Models for Observational Data. Mathematics Of Computation, (195), 444.

[26] Wedel, M. (2002). Concomitant variables in finite mixture models. Statistica Neerlandica, 56(3), 362-375.

[27] Wood, S. N., & Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling, 157, 157-177.

[28] Wood, S. N. (2017). Generalized additive models: an introduction with R. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2017.

Part III

On the use of Reparameterized Beta Distribution in Linear Cluster Weighted Model

On the use of Reparameterized Beta Distribution in Linear Cluster Weighted Model

Abstract

An extension of mixture models with random covariates related to the linear Cluster Weighted Model (CWM) is presented for model-based clustering applications. Beta regression is the standard approach to model a dependent variable with the range in the unit interval $[0, 1]$. However, in some situations, a problem that could arise is a direct consequence of the flexibility of the beta distribution. When it is considered as a mixture component in a mixture model, it may be too flexible due to the great variety of shapes (including multi-modal shapes) that can assume so that it may be difficult to understand easily the real meaning of each component. In this paper we developed an extension of the beta mixture models focusing on the subset of unimodal beta distributions, with the aim of improving the interpretation of each mixture component and then identifying better the respective cluster in the population. Estimation is performed via maximum likelihood with EM algorithm. Finally, with simulated and real data we investigate the performances, limits and benefits comparing this model with other models related to it.

Keywords: Mixture Models, Model-based Clustering, EM Algorithm, Beta Distribution, Beta Regression, Cluster Weighted Model.

1. Introduction

Most of work published in the context of mixtures of distributions is related to mixtures of normal densities to approximate a continuous distribution with support $S = \mathbb{R}$ (McLachlan and Peel, 2000). However, this approach could not be adequate if $S \subset \mathbb{R}$ due to the fact that we are allocating a probability mass outside the support S; this problem is commonly called *boundary bias*.
In this work we are interested in situations where the data have support in the standard unit interval $[0, 1]$ and for this reason a reasonable flexible family of distributions that can be considered with support $S \in [0, 1]$ are the *beta* densities. At the same time, we are interested in mixture models, where the aim is to capture the effect of latent factors. In such class of models, the observations are assumed to arise from unobserved groups in the population and the purpose of the analysis

is to provide an estimate of the unknown parameters. However, a possible problem that could arise in this context is a direct consequence of flexibility of the beta distribution. When it is incorporated in a mixture component, if the estimated model includes multi-modal shapes, it may be difficult to understand easily the real meaning of each component, especially if each mixture component represent a cluster in the population that generates observed data. For these reasons mixtures of unimodal beta distributions have been developed (Bagnato and Punzo, 2012) to overcome this problem and some applications are presented in Dean and Nugent, 2013 explaining the practical benefits of this approach.

In a regression application the problem highlighted above in the context of mixtures of distributions is the same. First of all we would like to avoid the usual practice to perform a regression analysis where the dependent variable Y assumes values in the standard unit interval transforming the response in such a way that it takes values in the real line, and then apply a linear regression analysis (for example applying to Y the *logit* transformation where $Y_{logit} = \log(Y/(1-Y))$). This approach, indeed, presents some problems (Ferrari *et al.*, 2004), in particular: (1) distribution of rates and proportions are (generally) asymmetric so that Gaussian approximations for interval estimation are not suitable; (2) regressions coefficients are related to the mean of Y_{logit} and not to the mean of Y, (3) this type of data are typically heteroskedastic. To overcome these problems a new class of a regression models called beta regression has been proposed based on the assumption that the response variable is beta distributed. One of the main advantages of beta regression models is that parameters can be interpreted in terms of the mean of Y and the model is naturally heteroskedastic.

Clearly, the beta regression model can be included in each mixture component (Verkuilen *et al.,* 2012, Grun *et al.,* 2012) to take into account such heterogeneity present in the data.

Motivated by these considerations we define, in the regression context, a new class of mixture of regression models including unimodal beta densities as a reference distributions instead of the more general class of general beta densities with the aim of bringing the benefits of this parameterization also in the context of regression. The developed methodology finds applications every time the variable of interest takes values in the standard unit interval $[0, 1]$, such as concentration indices, proportions and rates in presence of unobserved heterogeneity (Barreto-Souza and Simas, 2017; Huerta *et al.*, 2018).

The reparameterized beta regression model can be applied as an alternative approach with respect to the standard beta regression and can be included in a mixture model with reparameterized beta regression components.

In particular, in this work we are interested in the development of an eminent member of the class of regression models with random covariates called Cluster Weighted Model (CWM, Ingrassia *et al.,* 2012) which recently becomes popular in

statistics and data mining. In CWM the innovative approach consists in modeling the joint probability of data rather than the conditional as in classical mixtures of regressions: taking into account for the joint distribution of the response and the covariates has been shown how the CWM represents an improvement both from the point of view of the interpretation of the parameters and from the point of view of the interpretation of the clusters with respect to the classical mixtures where only the conditional density is modeled (McLachlan and Basford, 1998). Several extensions and examples of CWM have been proposed in literature, including the generalized CWM (Ingrassia *et al.*, 2015), CWM with factor analyzer (Subedi *et al.*, 2013; Subedi *et al.*, 2015) and the CWM beta regression (Nieddu and Vitiello, 2014).

In Table 1 is synthesized the general context where can be contextualized the proposed model with pros and cons of each approach.

The work is organized as follows. In Sect. 2 motivating examples help to contextualize the approach proposed from a qualitative point of view, in Sect. 3 the model is defined. The EM algorithm for estimation of unknown parameter with some computational details is explained in Sect. 4 and 5 while in Sect. 6 applications with simulated and real data are presented.

Table 1: Summary of the methodological context.

Model	Pros	Cons
Beta regression (Ferrari *et al.*, 2004)	- Support of variable of interest in $[0, 1]$.	- Inadequate in cases of heterogeneous data.
Mixtures of beta regressions (Verkuilen *et al.*, 2012; Grun *et al.*, 2012)	- Support of variable of interest in $[0, 1]$. - Adequate in case of heterogeneous data.	- The wide flexibility of beta distribution could results in multi-modal shapes, making difficult the interpretation of the identified clusters.
Linear CWM (Ingrassia *et al.*, 2012)	- Flexible and powerful mixture models. - Adequate in case of heterogeneous data.	- Support of variable of interest in $\mathbb{R}$. - Data transformation is required to treat data in $[0, 1]$.
Beta CWM (Nieddu *et al.*, 2014)	- Flexible and powerful mixture models. - Adequate in case of heterogeneous data.	- The wide flexibility of beta distribution could results in multi-modal shapes, making difficult the interpretation of the identified clusters.
Reparameterized beta CWM	- Flexible and powerful mixture models with support of variable of interest in $[0, 1]$. - Adequate in case of heterogeneous data.	

2. Motivating examples

In this section two motivating examples are provided before introducing the proposed model from a theoretical point of view. We start with a mixture of distributions that shows the improvement brought by the use of a unimodal distribution, while in the second example the same arguments are applied in the regression context.

2.1. Mixtures of reparameterized beta distributions

The first example consists in four simulations from a mixture of distributions. In such type of applications, it is generally assumed that a vector of independent and identically distributed $\boldsymbol{x} = (x_1, ..., x_n)$ is sampled from a random variable $\boldsymbol{X}$ with density $f(x)$. Suppose that we have G latent groups, where each group is represented by a mixture component $f(x; \boldsymbol{\theta}_g)$ depending on a vector of unknown parameter $\boldsymbol{\theta}_g$. The marginal density of $\boldsymbol{X}$ can be generally defined as

$$f(x) = \sum_{g=1}^{G} \pi_g f(x; \boldsymbol{\theta}_g)$$

where $\sum_{g=1}^{G} \pi_g = 1$ and $\pi_g > 0$. If the support of $\boldsymbol{X}$ is compact such that $\boldsymbol{X} \in [p, q]$, $(p > 0,\ q > 0)$ the beta density is a very flexible family of distributions that can be considered in such kind of applications which can assume a lot of different shapes varying the parameter's values. The beta density in each mixture component, in the standard parameterization, is given by

$$f(x|p_g, q_g) = \frac{1}{B(p_g, q_g)} x^{(p_g - 1)} (1 - x)^{(q_g - 1)} \tag{1}$$

where $B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(pq)}$, $\Gamma(.)$ is the gamma function, $p_g > 0$ and $q_g > 0$; the expected value and the mode are respectively

$$E_f(X) = \frac{p_g}{p_g + q_g}$$

$$Mode_f(X) = \frac{p_g - 1}{p_g + q_g - 2}$$

From the standard theory if $p_g > 1$ and $q_g > 1$ the distribution (1) is unimodal and the contrary if $p_g < 1$ and $q_g < 1$. In Fig. 1 are shown different shapes of beta distribution varying the values of p_g and q_g.

The central point, is that it is possible to reparametrize the distribution (1) in order to obtain a new family of unimodal beta distributions. This new family is a subset of beta density depending on two parameters. The unimodal-reparameterized beta distribution is given by

$$f_{rep}(x|m,v) = \frac{x^{m/v}(1-x)^{(1-m)/v}}{B(\frac{m}{v}+1, \frac{1-m}{v}+1)} \tag{2}$$

$$m = \frac{p-1}{p+q-2}$$

$$v = \frac{1}{p+q-2}$$

where $m \in [0,1]$ represents the mode and v the concentration around the mode (Bagnato and Punzo, 2012). In Fig. 2 are shown different shapes of reparameterized beta distribution.

Generally, in mixture analysis, it is assumed that each mixture component represents a cluster in the population. Therefore, the aim is to estimate the unknown parameters in each mixture component in order to understand the mechanism that generates the observed data. Thus, if the distribution $f(x;\boldsymbol{\theta}_g)$ within a group g is bimodal or (more generally) is not unimodal the interpretation of the component as one-group cluster is not possible, making the interpretation of the single component very difficult. For this reason unimodal component densities are generally preferred in different kinds of applications (Dean, 2013).

Lets consider now the following simulation from a mixture with two components

$$f(x) = \sum_{g=1}^{2} \pi_g f_{rep}(x; m_g, v_g) \tag{3}$$

with $n = 200$ observations with parameters (m_g, v_g, π_g) synthesized in Table 2. Data has been sampled from a unimodal beta mixture and we compare now the parameter recovering with a mixture of reparameterized beta and a mixture of beta showing how a mixture of beta presents some limitations in the identification of the mixture components. In Table 3 and Table 4 are summarized the estimated parameters while see Fig. 3 for graphical results. It is immediate to observe in Fig. 3 how unimodal beta components are able to describe the mixture's features for all the four simulations while the standard parameterization provides a good result only in simulation 2. Furthermore, abnormal behaviors for classical beta mixture are present in simulations 1 and 4 where the model is not able to detect the two clusters while in simulation 3 the first component can not capture the behavior of the first cluster.

Figure 1: Different shapes of beta densities (1) varying p_g and q_g.

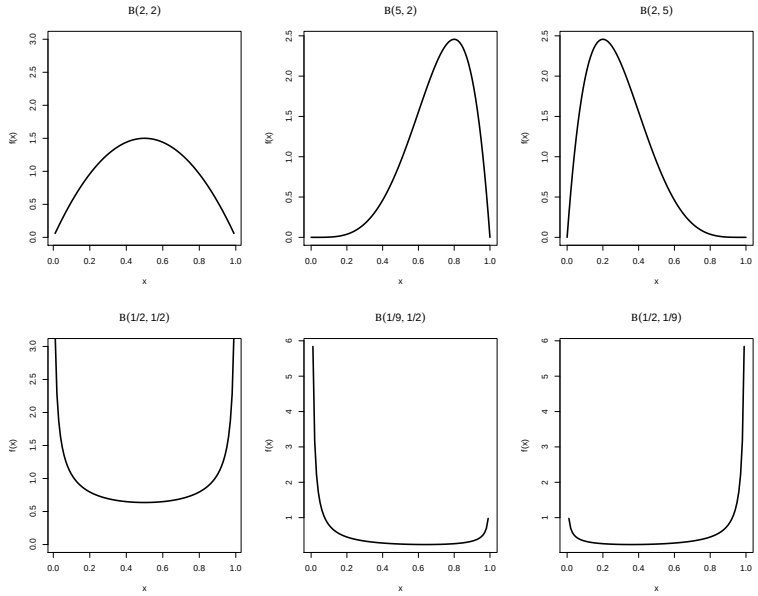


Figure 2: Different shapes of beta densities (2) varying m and v.

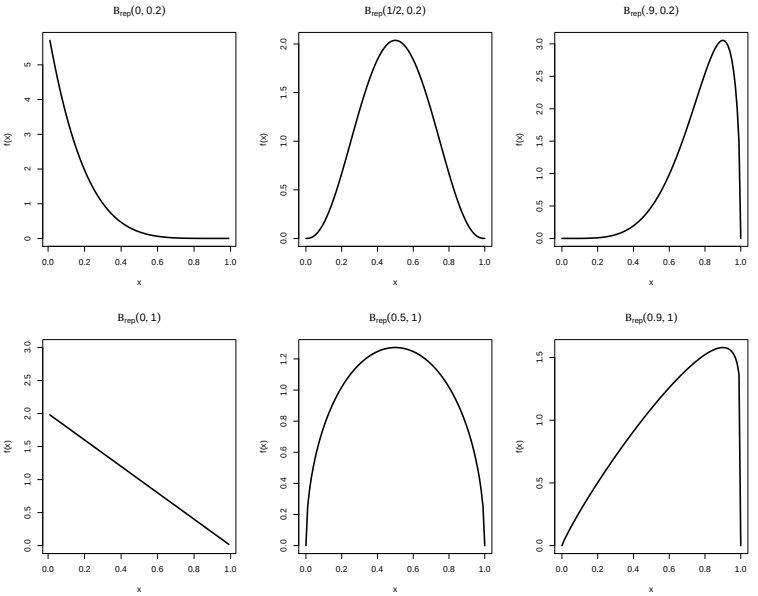


Table 2: Original parameters of a unimodal mixture model (3).

Simulation	Mode (m)	Dispersion (v)	Weight (π)
1	[0.05, 0.95]	[0.2, 0.2]	[0.5, 0.5]
2	[0.30, 0.70]	[0.2, 0.2]	[0.5, 0.5]
3	[0.20, 0.80]	[0.1, 0.1]	[0.2, 0.8]
4	[0.20, 0.80]	[0.1, 0.1]	[0.5, 0.5]

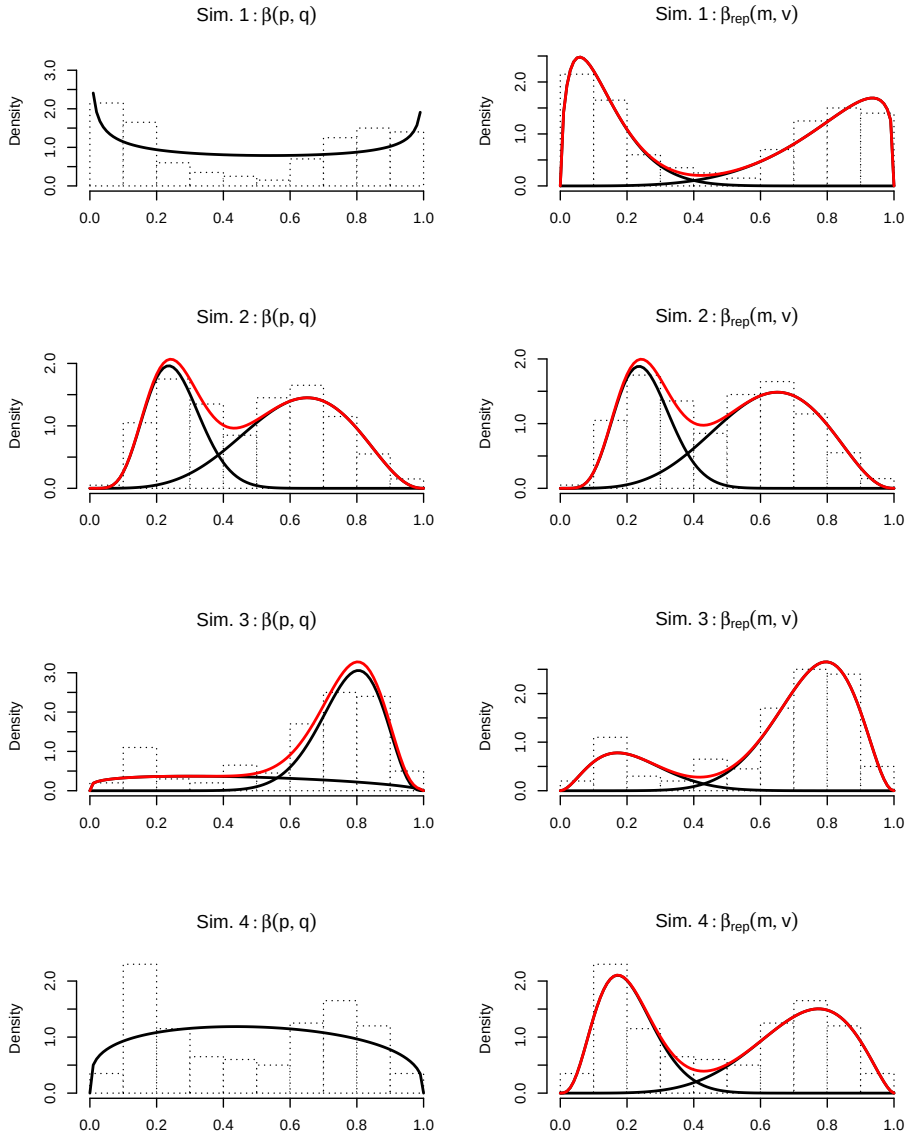
Table 3: Estimated parameters with a reparameterized beta mixture model (3).

Simulation	Mode (m)	Dispersion (v)	Weight (π)
1	[0.058, 0.935]	[0.099, 0.232]	[0.475, 0.524]
2	[0.236, 0.651]	[0.038, 0.145]	[0.383, 0.616]
3	[0.171, 0.796]	[0.089, 0.096]	[0.210, 0.789]
4	[0.172, 0.773]	[0.055, 0.139]	[0.455, 0.545]

Table 4: Estimated parameters with a beta mixture model (1).

Simulation	p	q	Weight (π)
1	[0.664, -]	[0.716, -]	[1, 0]
2	[7.154, 5.492]	[20.904, 3.404]	[0.4, 0.6]
3	[14.668, 1.256]	[4.319, 1.611]	[0.715, 0.285]
4	[1.280, -]	[1.367, -]	[1, 0]

Figure 3: Graphical results of beta mixture (left) and a reparameterized beta mixture (right). In red is traced the global density while in black the group-densities.



2.2. Mixture of standard beta regressions

The same idea can be applied in the context of regression where a dependence between Y (response variable) and X (covariate) is introduced. In this example (without enter into details about the model) we consider the following mixture models

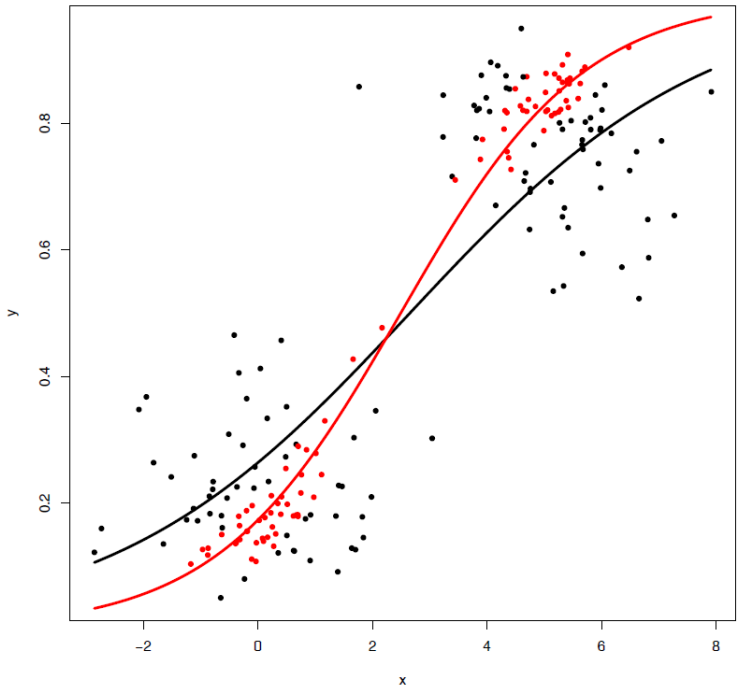
$$f(y|x;\boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g f(y|x;\boldsymbol{\theta}_g)$$

where the conditional distribution $f(y|x;\boldsymbol{\theta}_g)$ is a beta regression model. In Table 4 are synthesized the original parameters for this simulation while in Fig. 5 is easy to see how this model is not able to describe in a proper way the 2 clusters.

Figure 4: Original parameters if beta mixture regression.

Cluster	Mode (m_g)	Dispersion (v_g)	Weight (π_g)	Covariate X
1	0.2	0.05	0.5	$N(0,1)$
2	0.8	0.05	0.5	$N(5,1)$

Figure 5: Graphical results of estimated beta mixture model.



3. CWM Definition

Let $(\boldsymbol{X}, Y)$ be a pair of a covariate vector $\boldsymbol{X}$ and Y a response variable defined on some space Ω, and assume that Ω can be partitioned into G groups $G_1, ..., G_G$. CWM models the joint distribution $p(\boldsymbol{x}, y)$ of $(\boldsymbol{X}, Y)$ as a convex combination, with weights $\pi_1, ..., \pi_G$ such that

$$p(\boldsymbol{x}, y) = \sum_{g=1}^{G} \pi_g p(y|\boldsymbol{x}, \Omega_g) p(\boldsymbol{x}|\Omega_g) \tag{4}$$

where $p(y|\boldsymbol{x}, \Omega_g)$ is the conditional distribution of $Y|\boldsymbol{x}$ in the group g and $p(\boldsymbol{x}|\Omega_g)$ is the marginal distribution of $\boldsymbol{X}$. Model (4) is the general specification of CWM. This model represents a very general family of mixture models and moreover a large class of mixture models can be generalized using (4) (Ingrassia *et al.*, 2012; Ingrassia *et al.*, 2015; Mazza *et al.*, 2018). Regarding the marginal distribution of $\boldsymbol{X}$ if both discrete and continuous covariates are available the vector of covariates can be written as $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{V})$, where $\boldsymbol{U}$ is a p-variate vector of continuous covariates, and $\boldsymbol{V}$ is a q-variate vector of finite discrete covariates. Assuming that $\boldsymbol{U}$ and $\boldsymbol{V}$ are locally independent (that is, they are independent within each mixture component) model (4) can be written as

$$\begin{aligned} p(\boldsymbol{x}, y) &= \sum_{g=1}^{G} \pi_g p(y|\boldsymbol{x}, \Omega_g) p(\boldsymbol{x}|\Omega_g) \\ &= \sum_{g=1}^{G} \pi_g p(y|\boldsymbol{x}, \Omega_g) p(\boldsymbol{u}|\boldsymbol{\psi}_g^1) p(\boldsymbol{v}; |\boldsymbol{\psi}_g^2) \end{aligned}$$

where $p(\boldsymbol{u}|\boldsymbol{\psi}_g^1)$ is the marginal distribution of continuous covariates $\boldsymbol{U}$ depending on a vector of unknown parameters $\boldsymbol{\psi}_g^1$ and $p(\boldsymbol{v}; |\boldsymbol{\psi}_g^2)$ is the marginal distribution of discrete covariates $\boldsymbol{V}$ in the g-th component. In case of classification problem, each observation can be assigned to the group with the maximum posterior probability calculated as

$$p(\Omega_g|\boldsymbol{x}, y) = \frac{p(\boldsymbol{x}, y, \Omega_g)}{p(\boldsymbol{x}, y)} = \frac{\pi_g p(y|\boldsymbol{x}, \Omega_g) p(\boldsymbol{x}|\Omega_g)}{\sum_{g=1}^{G} \pi_g p(y|\boldsymbol{x}, \Omega_g) p(\boldsymbol{x}|\Omega_g)}$$

for $g = 1, ..., G$.

3.1. CWM with standard beta components

A first definition of the CWM with beta components can be found in Nieddu and Vitiello, 2014. The beta density with support $S = [0, 1]$ is defined as

$$f_{beta}(y;p,q) = \frac{y^{p-1}(1-y)^{q-1}}{B(p,q)} \tag{5}$$

where $Y \in [0,1]$, $p > 0$, $q > 0$ and $B(.)$ is the beta function. If Y has density defined in (5) then

$$E(Y) = \frac{p}{p+q}$$

$$Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}$$

A more useful parameterization (Ferrari and Cribari-Neto, 2004) can be obtained by setting $\mu = p/(p+q)$ and $\phi = p+q$ in (5) leading to the following reparameterization:

$$f_{beta}(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\left((1-\mu)\phi\right)} y^{(\mu\phi-1)}(1-y)^{((1-\mu)\phi-1)} \tag{6}$$

where $\mu \in (0,1)$, $\phi > 0$. With this parameterization if $Y \sim Beta(\mu,\phi)$ then

$$E_{f_{beta}}(Y) = \mu$$

$$Var_{f_{beta}}(Y) = \frac{\mu(1-\mu)}{(1+\phi)}$$

where μ is the mean and ϕ is a dispersion parameter around the mean.
Given a set of covariates $\boldsymbol{X}$ and a random sample Y such that $Y_i|\boldsymbol{X} = \boldsymbol{x} \sim Beta(\mu_i,\phi)$, the beta regression is defined as

$$g(\mu_i) = \boldsymbol{x}_i^{'}\boldsymbol{\beta} \tag{7}$$

where $\boldsymbol{\beta} = (\beta_1,...,\beta_k)^{'}$ is a vector of unknown regression coefficients and $\boldsymbol{x}_i = (x_{i1},...,x_{ik})^{'}$ the vector of covariates. The function $g(.) : (0,1) \to \mathbb{R}$ is the link function, strictly increasing and twice differentiable such that $\mu_i = g^{-1}(\boldsymbol{x}_i^{'}\boldsymbol{\beta})$.
As usual, maximizing the log likelihood function we obtain an estimate of the vector of unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta},\phi)$. Including the density (6) in (4), the beta CWM is defined, for the i-th observation, as

$$p_{CWM-beta}(y_i,\boldsymbol{x}_i) = \sum_{g=1}^{G} \pi_g f_{beta}(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_g,\phi_g)p(\boldsymbol{u}_i|\boldsymbol{\psi}_g^1)p(\boldsymbol{v}_i;|\boldsymbol{\psi}_g^2) \tag{8}$$

Also the dispersion parameter ϕ can be linked to the linear predictors defining a function $g_1(\phi_i) = \boldsymbol{x}_i^{'}\boldsymbol{\lambda}$ strictly increasing and twice differentiable.

3.2. CWM with unimodal beta components

A general framework for univariate finite mixtures of densities with support $Y \in [0, \infty)$ parameterized in terms of modes and dispersion (around the mode) can be found in Bagnato and Punzo, 2012. A finite mixture of distributions is defined as

$$f(y; \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{v}) = \sum_{g=1}^{G} \pi_j f_{rep}(y; m_g, v_g)$$

where $f_{rep}(\cdot)$ is the unimodal component density belonging to a parametric family, $\boldsymbol{m} = (m_1, ..., m_G)$ the vector of modes and $\boldsymbol{v} = (v_1, ..., v_G)$ is a vector of positive parameters describing the concentration around the mode. Lets consider the subclass of beta densities with support $S = [0, 1]$ such that

$$f_{rep}(y; m, v) = \frac{x^{\frac{m}{v}}(1-x)^{\frac{1-m}{v}}}{B\left(\frac{m}{v}+1, \frac{1-m}{v}+1\right)} \tag{9}$$

where $B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(pq)}$, $m \in [0, 1]$ and $v > 0$.
Given the standard parameterization of beta distribution (1) it is possible to obtain (9) according to the following transformation system:

$$\begin{cases} p = \frac{m}{v} + 1 \\ q = \frac{1-m}{v} + 1 \end{cases} \quad \rightarrow \quad \begin{cases} m = \frac{p-1}{p+q-2} \\ v = \frac{1}{p+q-2} \end{cases}$$

Since $m \in [0, 1]$ and $v > 0$ the new parameterization coincides with (1) when $(p, q) \in [1, \infty) \times [1, \infty)$ and $(p, q) \neq (1, 1)$. Eq. (5) is unimodal if $p > 1$ and $q > 1$, thus we are focusing on the subclass of unimodal beta densities, omitting some shapes among which unlimited J-shaped, unlimited reverse J-shaped, the U-shaped and the uniform density.
In order to obtain estimates $(\hat{m}, \hat{v})$ for the unknown vector of parameters (m, v) is possible to maximize numerically the following derivatives of log-likelihood function:

$$\begin{aligned} \frac{\delta ln\, f(y; m, v)}{\delta m} &= \frac{1}{v}\left\{\left[\psi\left(\frac{1-m}{v}+1\right) - \psi\left(\frac{m}{v}+1\right)\right] + \right. \\ &+ \left. \ln(y) - \ln(1-y)\right\} \end{aligned} \tag{10}$$

$$\begin{aligned}\frac{\delta ln\, f(y;m,v)}{\delta v} &= \frac{1}{v^2}\left\{\left[-\psi\left(\frac{2v+1}{v}\right)\right]+\right. \\ &+ \left[m\,\psi\left(\frac{m}{v}+1\right)+(1-m)\,\psi\left(\frac{1-m}{v}+1\right)\right]+ \\ &+ \left.-m\,\ln(y)-(1-m)\,\ln(1-y)\right\}\end{aligned} \tag{11}$$

where $\psi(\cdot)$ is the digamma function. If $Y|\boldsymbol{X}=\boldsymbol{x}$ follows a reparameterized beta distribution i.e. $Y|\boldsymbol{X}=\boldsymbol{x}\sim Beta_{rep}(m_i,v)$ than the unimodal beta regression model can be defined, as explained before as

$$g(m_i)=\boldsymbol{x}_i'\boldsymbol{\beta} \tag{12}$$

Then, including the density (9) into (4) the CWM with unimodal beta component is defined for the i-th observation as

$$p_{CWM-rep}(y_i,\boldsymbol{x}_i)=\sum_{g=1}^{G}\pi_g f_{rep}(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_g,v_g)p(\boldsymbol{u}_i|\boldsymbol{\psi}_g^1)p(\boldsymbol{v}_i;|\boldsymbol{\psi}_g^2) \tag{13}$$

4. Maximum likelihood estimation (EM algorithm)

The model (13) can be estimated with EM algorithm (Dempster *et al.*, 1977) to maximize the log-likelihood function in order to obtain maximum likelihood estimates for the unknown parameters. The likelihood once fixed the number of components G is given by

$$L(\boldsymbol{\theta})=\sum_{i=1}^{n}\sum_{g=1}^{G}\pi_g f_{rep}(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_g,v_g)p(\boldsymbol{u}_i|\boldsymbol{\psi}_g^1)p(\boldsymbol{v}_i;|\boldsymbol{\psi}_g^2)$$

The complete log-likelihood for the defined model given the number of groups G is:

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln(\pi_g) + \\
&+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\left(f_{rep}(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_g,v_g)\right) + \\
&+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\left(p(\boldsymbol{u}_i|\boldsymbol{\psi}_g^1)\right) \\
&+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\left(p(\boldsymbol{v}_i|\boldsymbol{\psi}_g^2)\right)
\end{aligned} \tag{14}
$$

where z_{ig} is the indicator variable that describes the individual membership to the latent group, i.e. $z_{ig} = 1$ if the individual i belongs to the latent group g, 0 otherwise. On the iteration $(k+1)$, the E-step requires the calculation of the conditional expectation of the random variable Z_{ig} related to each z_{ig} given the augmented data sample $\{(\boldsymbol{x}_1, y_1, \boldsymbol{z}_1), ..., (\boldsymbol{x}_n, y_n, \boldsymbol{z}_n)\}$. It follows that

$$
\begin{aligned}
E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ig}|(\boldsymbol{x}_i, y_i)\right] &= \tau_{ig} \\
&= \frac{\pi_g^{(k)} f_{rep}(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_g^{(k)},v_g^{(k)})p(\boldsymbol{u}_i|\boldsymbol{\psi}_g^{1,(k)})p(\boldsymbol{v}_i;|\boldsymbol{\psi}_g^{2,(k)})}{\sum_{g=1}^{G}\pi_g^{(k)} f_{rep}(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_i^{(k)},v_g^{(k)})p(\boldsymbol{u}_i|\boldsymbol{\psi}_g^{1,(k)})p(\boldsymbol{v}_i;|\boldsymbol{\psi}_g^{2,(k)})}
\end{aligned} \tag{15}
$$

which corresponds to the posterior probability that the observation $(\boldsymbol{x}_i, y_i)$ belongs to the g-th component of the mixture given the current $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)}, v^{(k)}, \boldsymbol{\psi}^{1,(k)}, \boldsymbol{\psi}^{2,(k)})$. In the M-step, on the iteration $(k+1)$, the conditional expectation of $l(\boldsymbol{\theta}^{(k)})$ given the observed data is maximized with respect to $\boldsymbol{\theta}$. Let

$$
\begin{aligned}
Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) &= \sum_{i=1}^{n}\sum_{g=1}^{G} \tau_{ig}^{(k)}\ln(\pi_g) + \\
&\sum_{i=1}^{n}\sum_{g=1}^{G} \tau_{ig}^{(k)}\ln\left(f_{rep}(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_g,v_g)\right) + \\
&\sum_{i=1}^{n}\sum_{g=1}^{G} \tau_{ig}^{(k)}\ln\left(p(\boldsymbol{u}_i|\boldsymbol{\psi}_g^1)\right) + \\
&\sum_{i=1}^{n}\sum_{g=1}^{G} \tau_{ig}^{(k)}\ln\left(p(\boldsymbol{v}_i|\boldsymbol{\psi}_g^2)\right)
\end{aligned} \tag{16}
$$

where $\tau_{ig}^{(k)}$ are the current expectation of z_{ig} provided by (15). As the terms in (16) have zero cross-derivatives, they can be maximized separately. The maximization of $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ with respect to the mixture weights is standard and can be obtained with Lagrangian multipliers as well as for the parameters related to $\boldsymbol{U}$ and $\mathbf{V}$ (Ingrassia *et al.*, 2015). Finally, the maximization of $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\beta}_g$ and v_g is obtained solving the following equations:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(k)}} = \sum_{i=1}^{n} \tau_{ig}^{(k)} \frac{\partial \ln\left(f(y_i|x_i, \boldsymbol{\beta}_g^{(k)}, v_g)\right)}{\partial \boldsymbol{\beta}^{(k)}} = 0$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial v^{(k)}} = \sum_{i=1}^{n} \tau_{ig}^{(k)} \frac{\partial \ln\left(f(y_i|x_i, \boldsymbol{\beta}_g^{(k)}, v_g)\right)}{\partial v^{(k)}} = 0$$

4.1. Computational details and initialization

Code for the EM has written in R (R Development Core Team, 2011). For the conditional part $Y|\boldsymbol{X} = \boldsymbol{x}$ with classical beta the function *betareg()* of R package betareg (available on CRAN) has been used. The computational issues to discuss involve the initialization of parameters and the definition of a convergence criterion to stop the procedure.

A standard initialization for EM plans to randomly defining starting values for the unknown vector of parameters $\boldsymbol{\psi}$. However, a possible approach (McLachlan and Peel, 2000; Punzo, 2012) is to specify the values of z_{ig} for all the observations. A random initialization can be repeated many times from different random positions selecting at the end the estimates that maximize the log-likelihood (Leisch, 2004).

4.2. Convergence criterion

Aitken acceleration (Aitken, 1926) can be used to make a decision about the convergence of the algorithm. It estimates the asymptotic maximum of the log-likelihood at each iteration:

$$a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}}$$

where $l^{(k)}$ is the log-likelihood value at iteration k. The asymptotic estimate of the likelihood (Böhning *et al.*, 1994) at iteration $k+1$ is given by

$$l_{\infty}^{(k+1)} = l^{(k)} + \frac{l^{(k+1)} - l^{(k)}}{1 - a^{(k)}}$$

By default, in following simulations we stop the EM if $l_{\infty}^{(k+1)} - l^{(k)} < \epsilon = 0.05$ (McNicholas, 2010; Punzo, 2012).

5. A simulation study

Monte Carlo simulations have been conducted to investigate parameters recovery of the EM algorithm. The aim of this analysis is to check whether the EM algorithm is able to recover the generating parameters, in particular if the mean of the estimates across replications is statistically significantly different from the generating parameters. The variability of the estimates is another important quantity to keep under control.

5.1. Design

Three different scenarios have been implemented varying the number of groups, the number of observations and the relation between x and y in the linear predictor. In order to analyze the impact of the number of observations n each scenario is replicated with tho different sample size $n = (500, 1000)$.
For each scenarios, and for each replication, the group membership z_{ig} were randomly generated from a multinomial distribution, while the values of x and y are generating according to the parameters in Table 5.
The data have a very different configuration between scenarios (see Fig. 6); in particular in scenario 1 data intersects to form a cross, in scenario 2 data are unbalanced on the right tail of the distribution (see cluster black and green) while in scenario 3 data are separated with respect to the covariate.

5.2. Measures for analysis

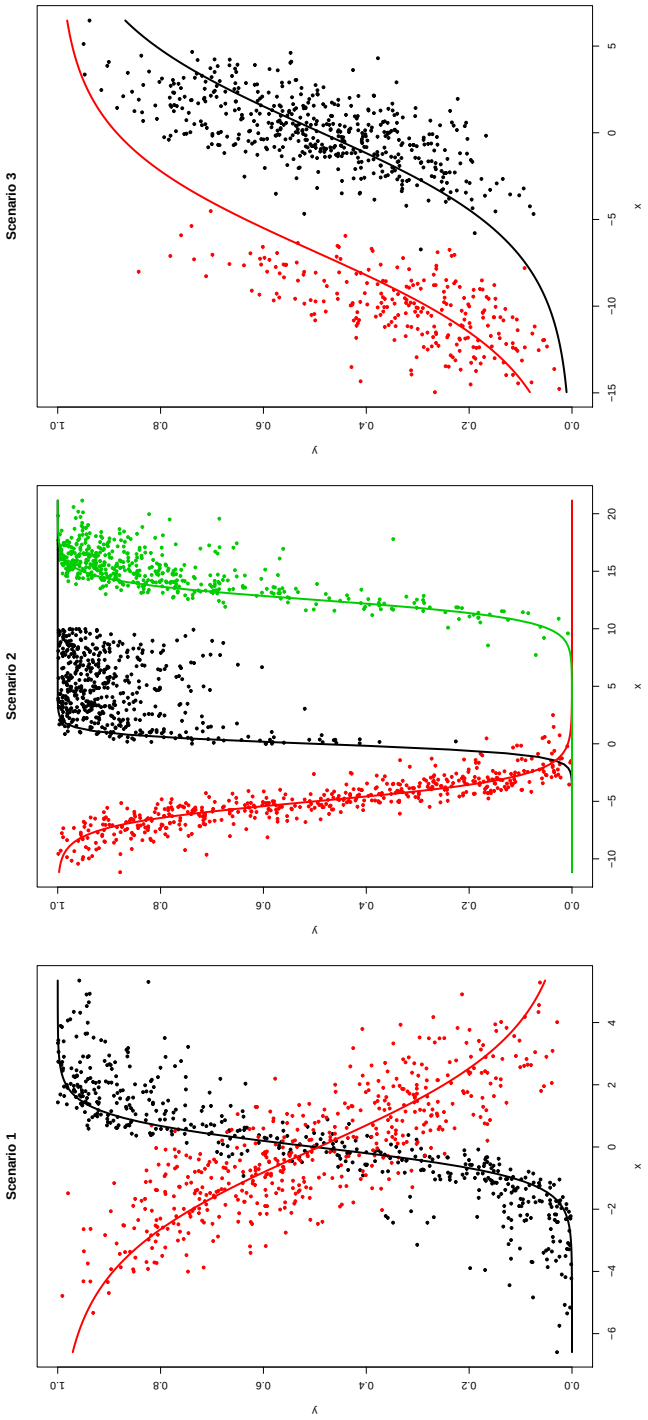
We consider the following quantities to measure the recovering of the originating parameters. The value R denotes the total number of replications, while $\hat{\beta}_r$ represent the estimated parameter for the replication r and β the true parameter.

$$BIAS(\hat{\beta}) = \sum_{r=1}^{R} (\hat{\beta}_r / R) - \beta$$

$$V(\hat{\beta}) = \sum_{r=1}^{R} (\hat{\beta}_r - \bar{\hat{\beta}})^2 / (R-1)$$

$$MSE(\hat{\beta}) = \Big[BIAS(\hat{\beta})\Big]^2 + \frac{R-1}{R} V(\hat{\beta})$$

Figure 6: Example of a generated data with $n = 1000$ samples for each scenario.



5.3. Results

Table 6, Table 7 and Table 8 report the results for scenarios 1, 2 and 3; within each scenario the measures of BIAS, SV and MSE can be compared varying the cluster, the number of observation and the relative parameter while in Fig. 7, Fig. 8 and Fig. 9 the histograms of estimated values by cluster and by parameters explain the variability of the estimates.
In all the considered simulations ($1500 \times 3 = 4500$), convergence at the true model was attained. In scenario 1 the BIAS and MSE are very low (practically negligible) for all the estimated parameters and the histograms in Fig. 7 shows how the estimate's variability is low for all the parameters.
In scenario 2 (see Fig. 8), due to the configuration of the cluster 1 (green) the parameters β_0 and β_1 show a magnitude of BIAS relatively higher than in scenario 1, due to the fact that the observations generated randomly at low values of x make the estimation of the 2 parameters more variable; this fact is confirmed by the histogram in Fig. 8 related to the parameter β_0 in cluster 3 and β_1 in cluster 1. However, considering still the scenario 2, we note how in the conformation of cluster 3, although similar to that of cluster 2, the values of BIAS are far lower than in cluster 2 because at low values of x we have enough observations to estimate more accurately the relationship between x and y.
Finally in scenario 3, as in scenario 1, there is no evidence of convergence problems.
Table 9 reports summary statistics about the true classification rates between the simulations. We cannot compare directly the scenarios among them from the point of view of cluster classification because they are characterized by a different underlying overlap. However, for all the 3 scenarios, the true classification rates are very high; it is interesting the impact of n that did not affect the mean of the rates, and the std. dev. decreases as expected.

Table 5: Simulation design parameters.

Scenario	Cluster	π_g	x_i	$g(m_i)$	v
1	1	1/2	$N(0,2)$	$g(m_i) = 0 + 2\,x_i$	0.1
	2	1/2	$N(0,2)$	$g(m_i) = 0 - 1\,x_i$	0.1
2	1	1/3	$Unif(0,10)$	$g(m_i) = 0 + 2\,x_i$	0.1
	2	1/3	$N(0,2)$	$g(m_i) = -5 - 1\,x_i$	0.1
	3	1/3	$N(15,2)$	$g(m_i) = -15 + 1.2\,x_i$	0.1
3	1	2/3	$N(0,2)$	$g(m_i) = 0 + 0.2\,x_i$	0.1
	2	1/3	$N(10,2)$	$g(m_i) = 5 + 2\,x_i$	0.1

Table 6: Parameters recovery for scenario 1.

		Cluster 1		Cluster 2	
Parameter	Measure	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$
β_0	BIAS	0.00272	-0.00086	-0.00256	-0.00027
	SV	0.00423	0.00225	0.00146	0.00071
	MSE	0.00423	0.00225	0.00146	0.00071
β_1	BIAS	-0.00366	0.00213	-0.00123	0.00072
	SV	0.00915	0.00435	0.00054	0.00024
	MSE	0.00915	0.00435	0.00054	0.00024
v	BIAS	-0.00008	-0.00002	-0.00078	-0.00031
	SV	0.00006	0.00003	0.00007	0.00004
	MSE	0.00006	0.00003	0.00007	0.00004

Table 7: Parameters recovery for scenario 2.

		Cluster 1		Cluster 2		Cluster 3	
Parameter	Measure	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$
β_0	BIAS	-0.15147	-0.16401	0.09576	0.09125	-0.0029	0.01052
	SV	0.03661	0.01626	0.04085	0.02781	0.64143	0.3335
	MSE	0.05948	0.04314	0.04994	0.03611	0.64016	0.33327
β_1	BIAS	0.2213	0.25206	0.01681	0.01613	-0.00096	-0.00212
	SV	0.09229	0.04433	0.00152	0.00107	0.00373	0.00194
	MSE	0.14108	0.10782	0.0018	0.00133	0.00372	0.00194
v	BIAS	0.00129	0.00191	-0.00299	-0.00281	-0.00159	-0.00134
	SV	0.00003	0.00002	0.00005	0.00011	0.00004	0.00002
	MSE	0.00003	0.00002	0.00006	0.00012	0.00004	0.00002

Table 8: Parameters recovery for scenario 3.

		Cluster 1		Cluster 2	
Parameter	Measure	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$
β_0	BIAS	-0.00013	-0.00058	-0.00571	-0.01113
	SV	0.00103	0.00047	0.07328	0.0383
	MSE	0.00103	0.00047	0.07317	0.03839
β_1	BIAS	-0.00092	-0.00069	-0.0003	-0.00131
	SV	0.00031	0.00015	0.00077	0.0004
	MSE	0.00031	0.00015	0.00077	0.0004
v	BIAS	0.00017	-0.00034	-0.00077	-0.00034
	SV	0.00005	0.00003	0.0001	0.00005
	MSE	0.00005	0.00003	0.0001	0.00005

Figure 7: Histograms of estimated values for each parameter (by column) and for each cluster (by row) for scenario 1.

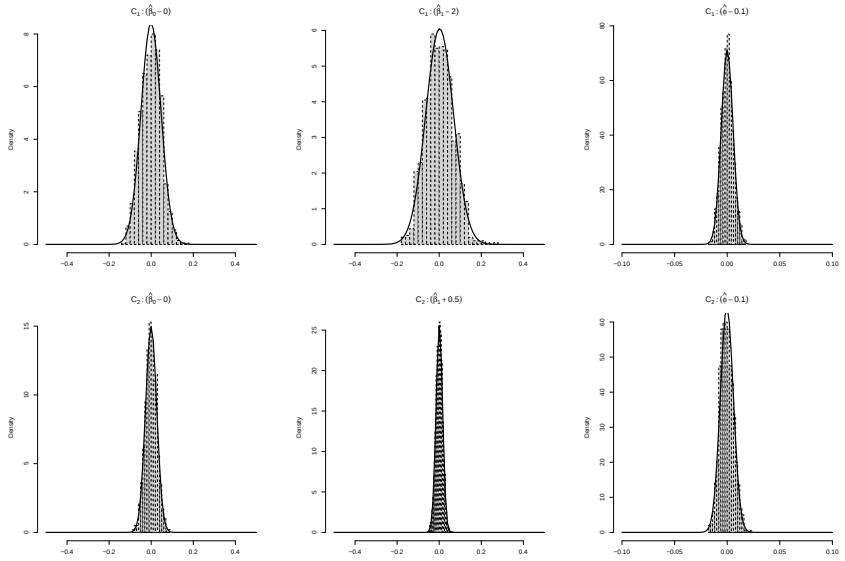


Figure 8: Histograms of estimated values for each parameter (by column) and for each cluster (by row) for scenario 2.

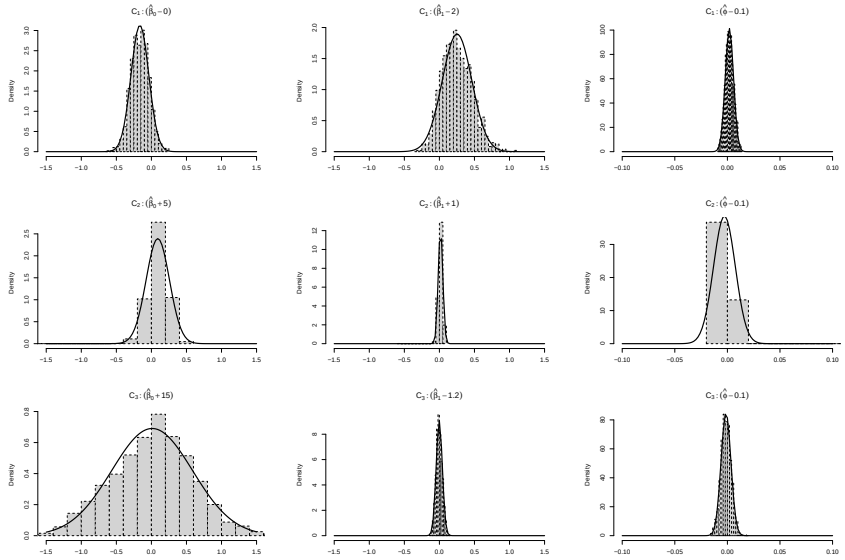


Figure 9: Histograms of estimated values for each parameter (by column) and for each cluster (by row) for scenario 3.

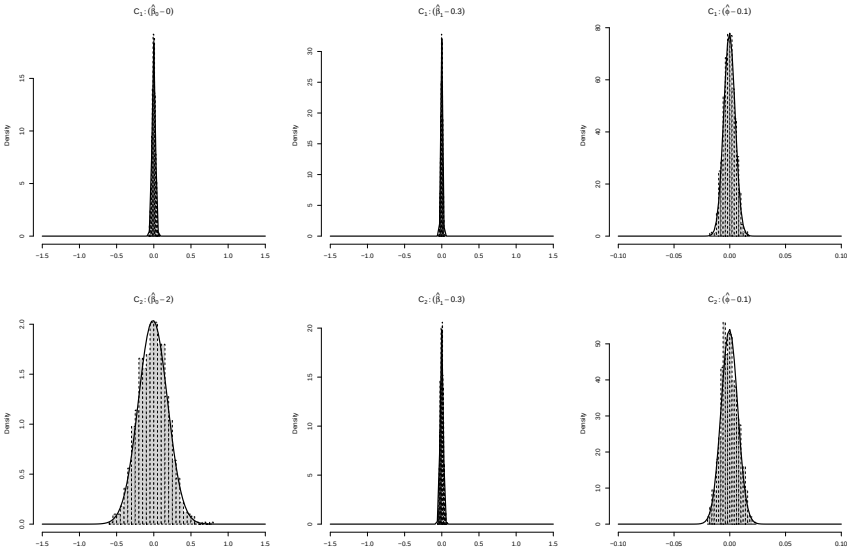


Table 9: Mean and standard deviation of the true classification rates.

Scenario	Statistic	$n = 500$	$n = 1000$
1	min	0.8720	0.8830
	mean	0.9033	0.9045
	max	0.9260	0.9235
	std. dev.	0.0091	0.0066
2	min	0.9820	0.9883
	mean	0.9954	0.9958
	max	1	0.9997
	std. dev.	0.0025	0.0017
3	min	0.9880	0.9927
	mean	0.9978	0.9978
	max	1	1
	std. dev.	0.0017	0.0012

6. Illustrative examples

After recalling the motivating example this section looks at applications of the models on real data.

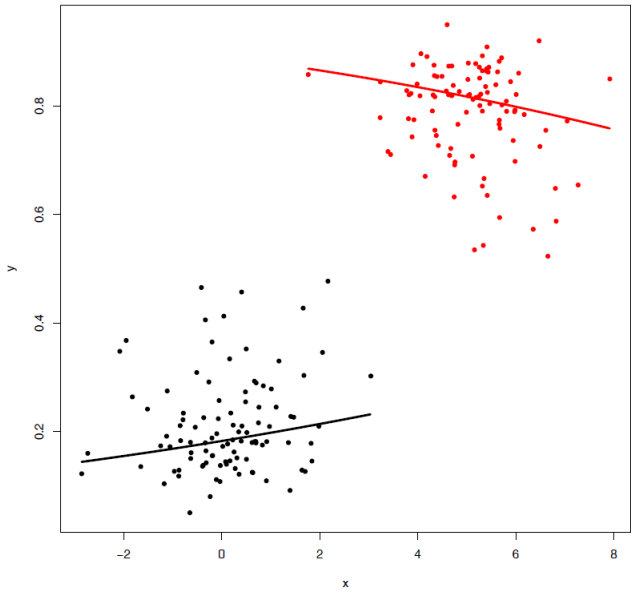
6.1. Recalling the motivating example

The limits shown in section 2 can be overcome considering the proposed model (Fig. 11). It can be clearly seen that the clusters are well separated and well described by the fitted model.

Figure 10: Estimated parameters with unimodal beta CWM.

Parameter	Cluster 1	Cluster 2
π	0.5	0.5
β_0	-1.498	2.100
β_1	0.099	-0.121
v	0.047	0.047

Figure 11: CWM estimated with unimodal beta components.



6.2. USNEWS dataset

The USNEWS dataset (http://lib.stat.cmu.edu/datasets/colleges/) contains information on over 1300 American colleges and universities. This dataset is taken from the 1995 U.S. News & World Report's Guide to America's Best Colleges. In order to test the proposed methods, acceptance rates in American colleges have been chosen as response variable with instate tuition feed and student/faculty ration as covariates. In this example we compare 3 models: CWM with unimodal beta, CWM beta and a mixture of beta regressions. In order to test the models in both bivariate and multivariate case the outcome variable is explained considering intuition, and intuition and study fact.

6.2.1. Univariate case

We start considering the relationship between the acceptance rate and tuition. According to the values of BIC (Table 10) is possible to choose the number of latent groups. In Fig. 11 are listed the estimated parameters while in Fig. 12 is shown a graphical representation of each model.

According to the BIC, the CWM-rep model is more parsimonious about the number of latent clusters, identifying 3 clusters instead of 5 clusters identified by the CWM-beta model while the mixture of betas seems not to be able to separate the clusters.

Given the classification provided by the CWM-rep the average values of acceptance rate by cluster are 0.75 (cluster 3), 0.79 (cluster 2) and 0.54 (cluster 1). The covariate seems to affect the response in cluster 1 and in cluster 3, where increasing the values of x the acceptance rate decreases (Fig. 12). Clusters 1 and 2 include mostly private institutions while cluster 3 contains public institutions.

Figure 12: Marginal density of x and scatter plot with the line showing the estimated relation between x and y with the reparameterized beta mixture CWM (left), CWM with beta regression (center) and a mixture of beta regressions.

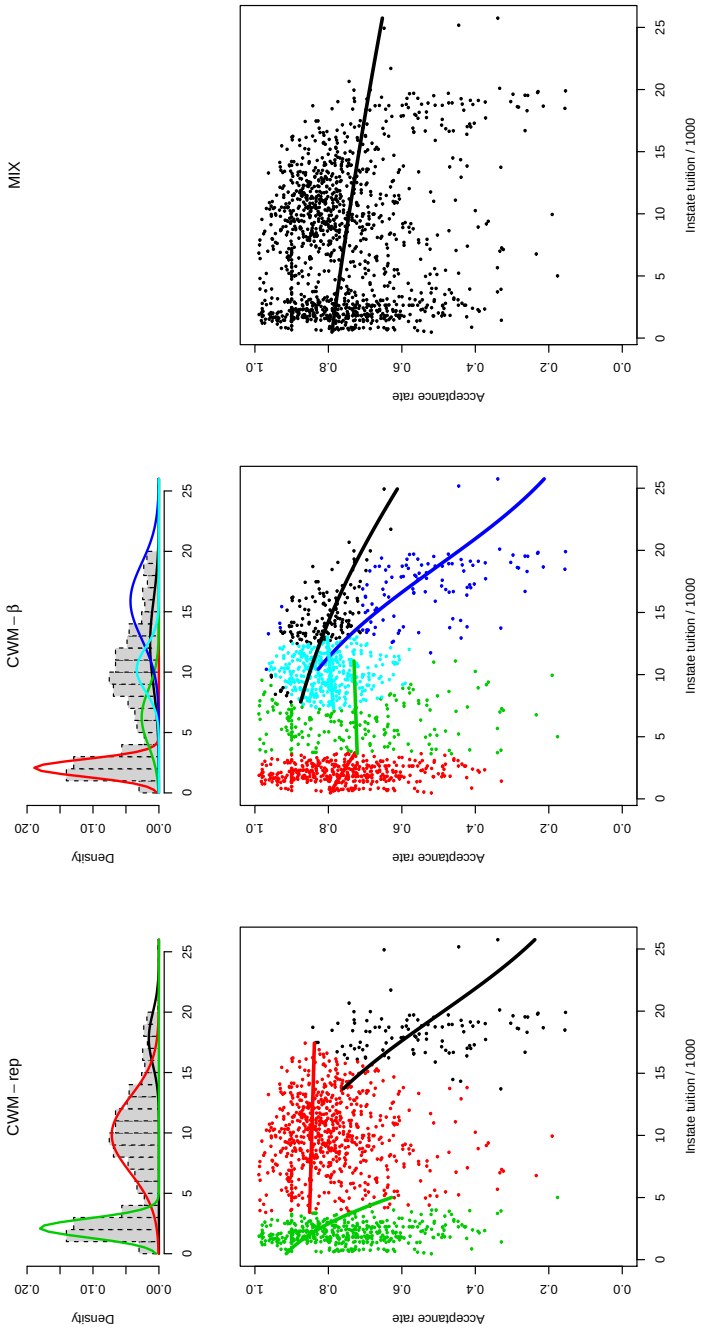


Table 10: BIC values varying the number of cluster.

Model	Cluster	BIC
CWM rep.	1	-6260.215
	2	-5771.226
	3	**-5760.348**
	4	-6203.986
	5	-6581.522
CWM beta	1	-6289.656
	2	-5599.544
	3	-5516.836
	4	-5519.505
	5	**-5507.919**
Mixture of beta	1	**-1538.644**
	2	-1568.956
	3	-1588.126
	4	-1577.100
	5	-1550.593

Table 11: Parameter estimated with reparameterized beta mixture.

Model	Parameter	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
CWM rep.	π	0.082	0.565	0.353	-	-
	β_0	3.801	1.770	2.570	-	-
	β_1	-0.192	-0.007	-0.409	-	-
	v	0.160	0.106	0.171	-	-
CWM beta	π	0.121	0.339	0.152	0.088	0.300
	β_0	2.628	1.409	0.931	3.527	1.172
	β_1	-0.087	-0.116	0.006	-0.188	0.018
	ϕ	3.498	2.124	1.685	2.179	3.136
Mixture of beta	π	1	-	-	-	-
	β_0	1.339	-	-	-	-
	β_1	-0.027	-	-	-	-
	ϕ	0.745	-	-	-	-

6.2.2. Bivariate case

Considering a model with two covariates the number of cluster changes only for the CWM-rep model from 3 to 4 (Table 12). However, we can see (Table 13) how cluster 1 weights relatively little compared to other clusters (0.027) and seems to capture outliers observations. Fig. 13 shows the scatter plots and the marginal distribution of the two covariates by cluster. We can see how the covariate study/-fact rate does not vary much changing the cluster while instate tuition feed seems to discriminate the clusters comparing the medians of the box-plots.

Figure 13: Marginal distribution of $\boldsymbol{x}$ by cluster.

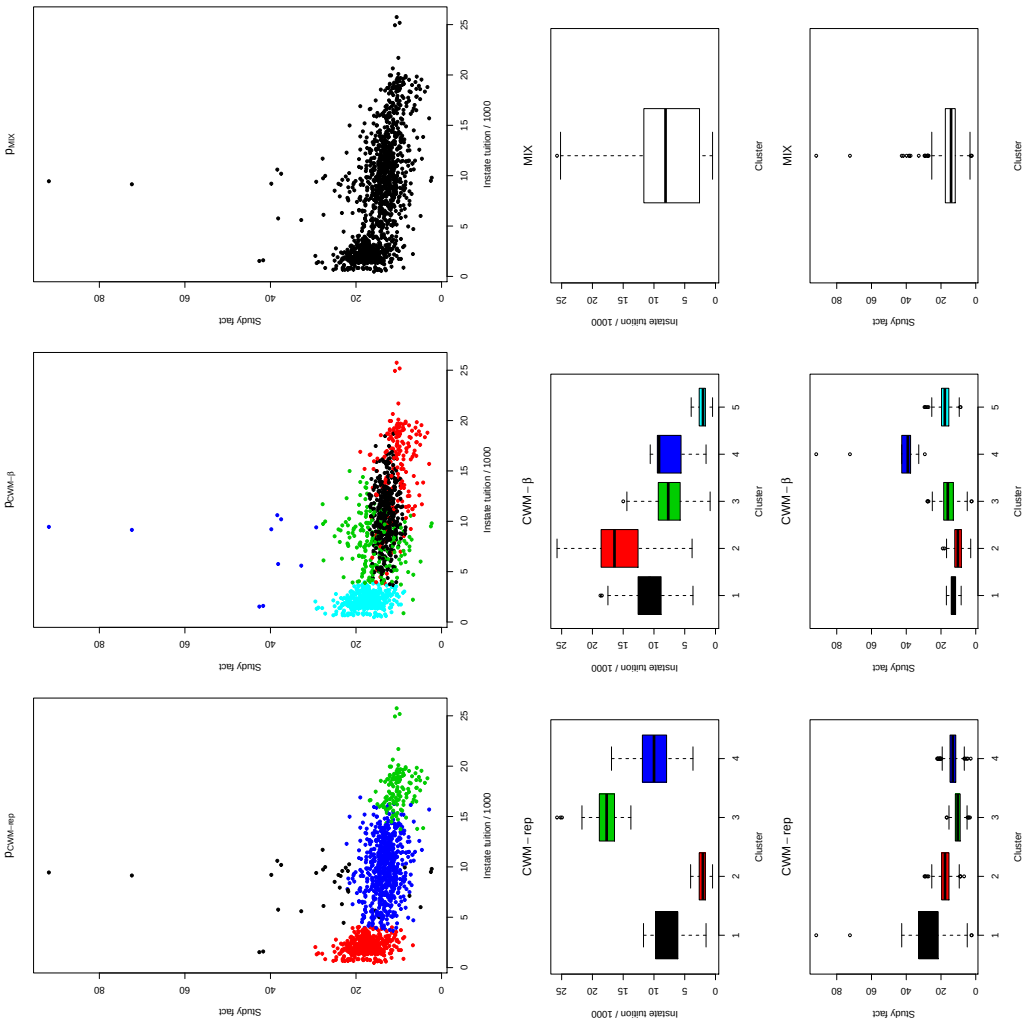


Table 12: BIC values varying the number of clusters.

Model	Cluster	BIC
CWM rep.	1	-13704.43
	2	-13185.65
	3	-12893.64
	4	**-12600.91**
	5	-13208.21
CWM beta	1	-13758.35
	2	-13075.88
	3	-12324.38
	4	-12256.17
	5	**-12241.86**
Mixture of beta	1	**-1535.26**
	2	-1558.24
	3	-1571.88
	4	-1556.29
	5	-1537.24

Table 13: Estimated parameters.

Model	Parameter	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
CWM rep.	π	0.027	0.350	0.104	0.519	-
	β_0	-1.014	1.55	2.264	0.691	-
	β_1	0.112	-0.288	-0.185	0.037	-
	β_2	0.078	0.042	0.147	0.05	-
	v	0.303	0.163	0.123	0.099	-
CWM beta	π	0.351	0.144	0.155	0.008	0.342
	β_0	1.766	2.727	0.15	1.456	0.905
	β_1	-0.037	-0.159	0.055	-0.075	-0.087
	β_2	-4.502	-3.189	-3.904	-4.893	-3.663
	ϕ	32.283	9.412	8.63	7.118	8.474
Mixture of beta	π	1	-	-	-	-
	β_0	1.176	-	-	-	-
	β_1	-0.0234	-	-	-	-
	β_2	0.0089	-	-	-	-
	ϕ	2.111	-	-	-	-

7. Conclusions

In this paper a new extension of CWM has been proposed to model an outcome variable that takes values in the standard unit interval. This model represents an alternative on the beta regression and in particular on the beta mixtures of regressions taking into account for the marginal distribution of the covariates.
We shown how, in some cases, it is useful to opt for a unimodal beta distribution to better identify and describe clusters. Among the possible future developments we identify, from an inferential point of view the importance to deliver an adequate inference for the parameters to assess the significance, while from a descriptive point of view we should test the model capability to include and describe a large number of covariates.

8. References

[1] Aitken, A. C. (1927). On Bernoulli's Numerical Solution of Algebraic Equations. Proceedings of the Royal Society of Edinburgh, 46, 289-305. Royal Society of Edinburgh Scotland Foundation.

[2] Bagnato, L., & Punzo, A. (2013). Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. Computational Statistics, 28(4), 1571-1597.

[3] Barreto-Souza, W., & Simas, A. B. (2017). Improving estimation for beta regression models via EM-algorithm and related diagnostic tools. Journal of Statistical Computation & Simulation, 87(14), 2847–2867.

[4] Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., & Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Annals Of The Institute Of Statistical Mathematics, 46(2), 373.

[5] Dean, N., & Nugent, R. (2013). Clustering student skill set profiles in a unit hypercube using mixtures of multivariate betas. Advances In Data Analysis & Classification, 7(3).

[6] Dempster, AP. Laird NM, & Rubin DB, (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal Of The Royal Statistical Society. Series B (Methodological), (1).

[7] Ferrari, S. L., & Cribari-Neto, F. (2004). Beta Regression for Modeling Rates and Proportions. Journal Of Applied Statistics, 31(7), 799-815.

[8] Gershenfeld, N. (1997). Nonlinear Inference and Cluster-Weighted Modeling. Annals Of The New York Academy Of Sciences, 808(1).

[9] Grün, B., Kosmidis I., Zeileis A. (2012). Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. Journal of Statistical Software, 48(11).

[10] Huerta, M., Leiva, V., Lillo, C., & Rodríguez, M. (2018). A beta partial least squares regression model: Diagnostics and application to mining industry data. Applied Stochastic Models in Business & Industry, 34(3), 305–321.

[11] Ingrassia, S., Minotti, S., & Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. Journal Of Classification, 29(3), 363-401.

[12] Ingrassia, S., Minotti, S. C., & Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. Computational Statistics and Data Analysis, 71, 159–182.

[13] Ingrassia, S., Punzo, A., Vittadini, G., & Minotti, S. (2015). The Generalized Linear Mixed Cluster-Weighted Model. Journal Of Classification, 32(1), 85-113.

[14] Mazza A., Punzo A., & Ingrassia S. (2018). flexCWM: A Flexible Framework for Cluster-Weighted Models. Journal of Statistical Software, 86(2). 1-30.

[15] McLachlan, G., & Basford, K. E. (1988). Mixture models: inference and applications to clustering. Dekker, 1988.

[16] McLachlan, G. J., & Peel, D. (2000). Finite mixture models. New York: Wiley, 2000.

[17] Nieddu L. & Vitiello C. (2014). Cluster Weighted Beta Regression. Rivista Italiana di Economia Demografia e Statistica.

[18] Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2013). Clustering and Classification via Cluster-Weighted Factor Analyzers. Advances in Data Analysis and Classification., 7(1).

[19] Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2015). Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. Statistical Methods and Applications 24(4), 623-649.

[20] Verkuilen, J., & Smithson, M. (2012). Mixed and Mixture Regression Models for Continuous Bounded Responses Using the Beta Distribution. Journal Of Educational And Behavioral Statistics, 37(1), 82-113.

Part IV

flexCWMext: an extension of flexCWM package for CWM Beta and CWM Generalized Additive Models

flexCWMext: an extension of flexCWM package for CWM Beta and CWM Generalized Additive Models

Abstract

Cluster Weighted Models (CWM) are mixtures of regressions models with random covariates. In this paper we introduce a new `R` package called `flexCWMext`, developed to extend the features of `flexCWM` implementing the beta distribution and the family of Generalized Additive Models in the CWM framework. The `flexCWM` package has been recently developed to estimates some mixture models related to the family of CWM. The presented package introduces three main extensions that are not yet implemented: the GAM-CWM (CWM mixtures of generalized additive models) the BETA-CWM (CWM mixtures of beta regressions) and unimodal BETA-CWM (CWM mixtures of unimodal beta regressions). The EM algorithm is used to obtain maximum-likelihood estimates of the parameters and some applications to artificial and real dataset are presented to explain the features of this package.

Keywords: Cluster Weighted Model, EM algorithm, Mixture models, GAM mixture models, Beta regression models.

1. Introduction

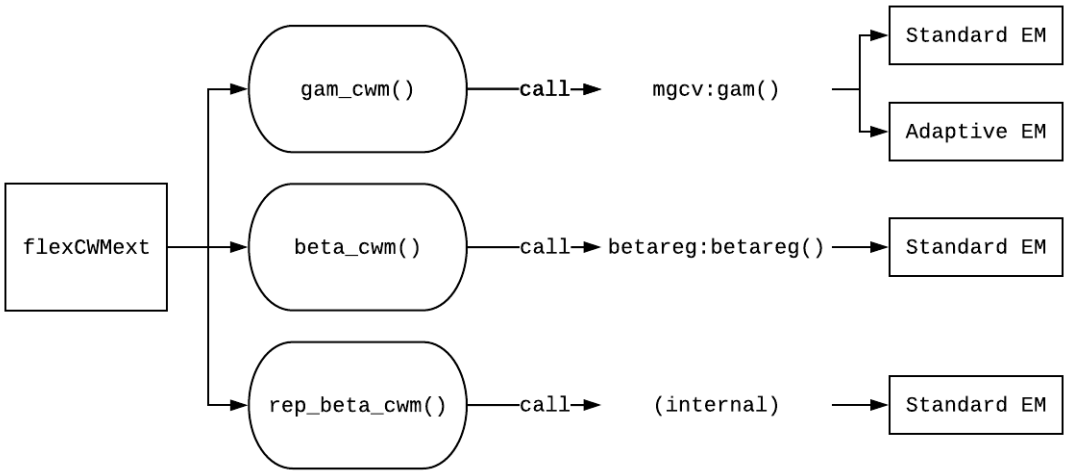
This paper extends the `flexCWM` package (Mazza and Punzo, 2018), a recent R package that implements a novel class of mixture models called Cluster Weighted (CWM). The CWM factorizes the joint distribution $p(\boldsymbol{x}, y)$ into the product of the conditional distribution $Y|\boldsymbol{X} = \boldsymbol{x}$ and the marginal distribution of $\boldsymbol{X}$ with the aim to capture latent sources of heterogeneity that split data into clusters.
In the `R` framework (R Core Team 2011) `flexCWM` is available to fit a large variety of CWMs, in particular the package supports modeling of the conditioned response variable by means of most important distributions of exponential family (Gaussian, Gamma, Poisson, Binomial, t-Student). Covariates can be also of mixed type: multivariate Gaussian, multinomial, binomial and Poisson. Although the `flexCWM` package allows to model a wide variety of phenomena is assumed a linear form that describes the relation between a response variable and the covariates. Clearly this assumption, in some applications, may not be appropriate and it would be wise to opt for a more flexible model able to capture in a proper way complex relationships between the variables.

Firstly, motivated by these considerations `flexCWMext` born with the aim to extend the `flexCWM` package including a sum of smooth functions of covariates, introducing the class of GAM models (Wood, 2017) within the CWM framework. This model is called GAM-CWM. In `R`, mixtures of GAM models (Hastie and Tibshirani, 1987) can be estimated with `flexmix:FLXMRmgcv()` (Grun and Leisch, 2008), however, the excessive flexibility of GAM can sometimes be a negative element, so that if the fitting process based on EM algorithm is not adequately controlled the risks is to confuse the clusters bringing the model to converge in local minimum. One possibility to control the model fitting (introduced in this package) is to control the basis dimension during the EM algorithm. This procedure, that is involved in the EM, is called "adaptive EM". As will be detailed in the follows a three-way deviance decomposition for mixtures of regressions models will be a key decision-making tool to control such flexibility.
Secondly, another important extension lets to model a response variable $Y \in [0, 1]$. The beta regression model (Ferrari and Neto, 2004) turns out to be a suitable model in this context, however a problem that could arise in some cases is a direct consequence of great flexibility of the beta distribution when it is embedded in a mixture component. The beta distribution, in some cases, may be too flexible due to the great variety of shapes (including multi-modal shapes) so that it could be difficult to understand the real meaning of each component. For these reasons, mixtures of reparameterized beta distributions has been developed (Bagnato and Punzo, 2012) to overcome these problems and some interesting applications are presented in Dean and Nugent, 2013. Motivated also by these considerations we implement a new class of mixtures of regressions models involving uni-modal beta densities. In this package three main functions has been implemented that lets to estimate these models (Fig. 1):

- `gam_cwm()`: CWM with generalized additive models in each mixture component.
- `beta_cwm()`: CWM with a beta regression component.
- `rep_beta_cwm()`: CWM with a unimodal beta regression component.

The paper is organized as follows. In Sect. 2 is specified the basic framework with the three extension. Sect. 3 outlines the EM algorithm for estimation while in Sect. 4 are illustrated some technical details related to computational and operational aspects. Finally in Sect. 5 some examples with simulated and real data are provided.

Figure 1: Overview of the package structure.



2. Model specification

In this section the general CWM model (Gershenfeld, 1997; Ingrassia *et al.*, 2012, 2014, 2015, Subedi *et al.*, 2013; Subedi *et al.*, 2015) is defined while the details of the implemented extensions are defined in following subsections. Let $\boldsymbol{X}$ is a $n \times p$ matrix of covariates and Y is a response variable belonging to exponential family, with joint probability distribution $p(\boldsymbol{x}, y)$. Suppose that Ω can be partitioned into G disjoint groups $(\Omega_1, ..., \Omega_G)$ such that $\boldsymbol{\Omega} = \Omega_1 \cup ... \cup \Omega_G$. CWM models the joint probability $p(\boldsymbol{x}, y)$ as follows:

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} p(y|\boldsymbol{x}, \Omega_g) p(\boldsymbol{x}|\Omega_g) \pi_g \tag{1}$$

where $p(y|\boldsymbol{x}, \Omega_g)$ is the conditional density of the response Y given the predictors $\boldsymbol{X}$ and Ω_g, $p(\boldsymbol{x}|\Omega_g)$ is the probability density of $\boldsymbol{X}$ given Ω_g, $\pi_g = p(\Omega_g)$ are the mixing weights of Ω_g so that $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$ and finally $\boldsymbol{\theta}$ is the set of all parameters in the model. Defining in different ways the conditional distribution $p(y|\boldsymbol{x}, \Omega_g)$ is possible to obtain different models, in particular the mixture of GAM and the beta regression.
In case of classification problem, the posterior probability belonging to the g-th group can be calculated as

$$p(\Omega_g|\boldsymbol{x}, y) = \frac{p(\boldsymbol{x}, y, \Omega_g)}{p(\boldsymbol{x}, y)} = \frac{p(y|\boldsymbol{x}, \Omega_g) p(\boldsymbol{x}|\Omega_g) \pi_g}{\sum_{g=1}^{G} p(y|\boldsymbol{x}, \Omega_g) p(\boldsymbol{x}|\Omega_g) \pi_g}$$

for $g = 1, ..., G$.

2.1. CWM mixtures of Generalized Additive Models

Lets focus the attention on the conditional part of the model $p(y|\boldsymbol{x}, \Omega_g)$ where the GAM models are involved. In order to deal with various response types we assume that $p(y|\boldsymbol{x}, \Omega_g)$ belongs to the exponential family. A monotone and differentiable link function $h(\cdot)$ is introduced to relate $\mu_g = E(y|\boldsymbol{x}, \Omega_g)$ to the covariates through the relation

$$h(\mu_g) = f_{1,g}(x_1) + ... + f_{p,g}(x_p) = \sum_{j=1}^{p} f_{j,g}(x_j)$$

where $f_{j,g}(\cdot)$ is the smooth function referred to the covariate x_j within the group g. A common approach to specify $f(\cdot)$ is to define linear combinations of a set of suitable functions called *basis*. Let $b_j(x)$ the j-th basis function. The function $f_{j,g}(x)$ can be defined as the following linear combination

$$f_{j,g}(x_j) = \sum_{j=0}^{H} b_j(x_j)\beta_{j,g} \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, ..., \beta_H)$ are parameters to be estimated, $b_0(x) = 1$ (in order to consider the intercept β_0) and H is the basis dimension. Because the interest is in the parameters $\boldsymbol{\beta}$ the distribution of $Y|\boldsymbol{X} = \boldsymbol{x}, \Omega_g$ will be denoted as $q(y|\boldsymbol{x}, \boldsymbol{\beta}_g, \zeta_g)$ where the parameter ζ_g is an additional parameter to take into account when a distribution from a two-parameter exponential family is considered. Then, the generalized additive CWM is defined as

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} q\left(y|\boldsymbol{x}, \boldsymbol{\beta}_g, \zeta_g\right) p(\boldsymbol{x}|\boldsymbol{\alpha}_g)\pi_g$$

and the log-likelihood is

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln(\pi_g) + \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln\left(q(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}_g, \zeta_g)\right) + \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln\left(p(\boldsymbol{x}_i|\boldsymbol{\alpha}_g)\right) \end{aligned} \tag{3}$$

In a mixture model context the basis dimension H (2) should be controlled, to prevent the model converging towards local minimum points. With a modification

of EM algorithm that we called *adaptive EM algorithm* it is possible to starting the algorithm with a low value of H and then increase it to better capture the relation between $\boldsymbol{X}$ and Y and at the same time to avoid the convergence towards local minimum points. The three way deviance decomposition (Sect. 4.1) is directly involved in the estimation process as a decision tool to choose the step when is useful to increase the flexibility.
The function `gam_cwm()` in `flexCWMext` estimates the model calling the function `mgcv:gam()` within each mixture component during the EM algorithm. Each parameter available in `mgcv:gam()` can be specified in the function `gam_cwm()`. Some additional parameters for this function are detailed in Table 2. The function `mgcv:s()` can be used in the `formula` specification of `flexCWMext` to specify the basis and the smooth functions.

2.2. CWM mixtures of Beta Regression Models

A first definition of the CWM with beta components can be found in Nieddu and Vitiello, 2014. The beta density with support $S = [0, 1]$ is defined as

$$f_{beta}(y; p, q) = \frac{y^{p-1}(1-y)^{q-1}}{B(p, q)} \tag{4}$$

where $Y \in (0, 1)$, $p > 0$, $q > 0$ and $B(.)$ is the beta function. If Y has density defined in (4) then

$$E(Y) = \frac{p}{p+q}$$

$$Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}$$

A more useful parameterization (Ferrari and Cribari-Neto, 2004) can be obtained by setting $\mu = p/(p+q)$ and $\phi = p+q$ in (4) leading to the following reparameterization:

$$f_{beta}(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\left((1-\mu)\phi\right)} y^{(\mu\phi-1)}(1-y)^{((1-\mu)\phi-1)} \tag{5}$$

where $\mu \in (0, 1)$, $\phi > 0$. With this parameterization if $Y \sim Beta(\mu, \phi)$ then

$$E_{f_{beta}}(Y) = \mu$$

$$Var_{f_{beta}}(Y) = \frac{\mu(1-\mu)}{(1+\phi)}$$

where μ is the mean and ϕ is a dispersion parameter around the mean.

Given a set of fixed covariates and a random sample Y_i, $i = 1, ..., n$, such that $Y_i|\boldsymbol{X} = \boldsymbol{x} \sim Beta(\mu_i, \phi)$, the beta regression is defined as

$$g(\mu_i) = \boldsymbol{x}_i^{'}\boldsymbol{\beta} \tag{6}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^{'}$ is a vector of regression coefficients and $\boldsymbol{x}_i = (x_{i1}, ..., x_{ik})$ the vector of covariates. The function $g(.) : (0, 1) \to \mathbb{R}$ is the link function, strictly increasing and twice differentiable such that $\mu_i = g^{-1}(\boldsymbol{x}_i^{'}\boldsymbol{\beta})$. Then, the CWM with beta components is defined as

$$p_{CWM-beta}(y_i, \boldsymbol{x}_i) = \sum_{g=1}^{G} \pi_g f_{beta}(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g, \phi_g) p(\boldsymbol{x}_i|\boldsymbol{\alpha}_g) \tag{7}$$

The complete log-likelihood for the defined model is:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln(\pi_g) + \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln\left(f_{beta}(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g, \phi_g)\right) + \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln\left(p(\boldsymbol{x}_i|\boldsymbol{\alpha}_g)\right) \end{aligned} \tag{8}$$

The function `beta_cwm()` estimates the model calling the function `betareg:betareg()` within each mixture component during the EM algorithm.

2.3. CWM mixtures of Reparameterized Beta

Given the standard parameterization of beta distribution (4) it is possible to obtain (9) according to the following transformation system:

$$\begin{cases} p = \frac{m}{v} + 1 \\ q = \frac{1-m}{v} + 1 \end{cases} \quad \to \quad \begin{cases} m = \frac{p-1}{p+q-2} \\ v = \frac{1}{p+q-2} \end{cases}$$

$$f_{rep}(y; m, v) = \frac{x^{\frac{m}{v}}(1-x)^{\frac{1-m}{v}}}{B\left(\frac{m}{v} + 1, \frac{1-m}{v} + 1\right)} \tag{9}$$

Since $m \in [0, 1]$ and $v > 0$ the new parameterization coincides with (4) when $(p, q) \in [1, \infty) \times [1, \infty)$ and $(p, q) \neq (1, 1)$. Eq. (4) is unimodal if $p > 1$ and $q > 1$, thus we are focusing on the subclass of unimodal beta densities, omitting

some shapes among which unlimited J-shaped, unlimited reverse J-shaped, the U-shaped and the uniform density.
It follow that if $Y|\boldsymbol{X} = \boldsymbol{x} \sim Beta_{rep}(m_i, v)$ than the unimodal beta regression model can be defined as

$$g(m_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \tag{10}$$

Finally, the CWM with unimodal beta components is defined for the i-th observation as

$$p_{CWM-rep}(y_i, \boldsymbol{x}_i) = \sum_{g=1}^{G} \pi_g f_{rep}(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g, v_g) p(\boldsymbol{x}_i|\boldsymbol{\alpha}_g) \tag{11}$$

and the complete log-likelihood for the defined model is

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln(\pi_g) + \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln\left(f_{rep}(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g, v_g)\right) + \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \ln\left(p(\boldsymbol{x}_i|\boldsymbol{\alpha}_g)\right) \end{aligned} \tag{12}$$

3. Maximum likelihood estimation with EM algorithm

The `flexCWMext` package implements (as `flexCWM`) the EM algorithm (Dempster *et al.,* 1977) to maximize the global log-likelihood function in order to obtain maximum likelihood estimates for the unknown parameters. Let $\boldsymbol{z}_i$ a k-dimensional component label vector in which the jth element z_{ij} is one or zero if respectively the mixture component of $(x_i, y_i)'$ is equal to j or not. On the iteration $(k+1)$, the E-step requires the calculation of the conditional expectation of the random variable Z_{ig} related to each z_{ig} given the augmented data sample $\{(\boldsymbol{x}_1, y_1, \boldsymbol{z}_1), ..., (\boldsymbol{x}_n, y_n, \boldsymbol{z}_n)\}$. In particular, for $i = 1, ..., n$ and $g = 1, ..., G$ it follows that

- GAM-CWM: $E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ig}|(\boldsymbol{x}_i, y_i)\right] = \frac{\pi_g^{(k)} f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g^{(k)}, \boldsymbol{\zeta}_g^{(k)}) p(\boldsymbol{x}_i|\boldsymbol{\alpha}_g)}{\sum_{g=1}^{G} \pi_g^{(k)} f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g^{(k)}, \boldsymbol{\zeta}_g^{(k)}) p(\boldsymbol{x}_i|\boldsymbol{\alpha}_g)} = \tau_{ij}$
- BETA-CWM: $E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ig}|(\boldsymbol{x}_i, y_i)\right] = \frac{\pi_g^{(k)} f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g^{(k)}, \phi_g^{(k)}) p(\boldsymbol{x}_i|\boldsymbol{\alpha}_g)}{\sum_{g=1}^{G} \pi_g^{(k)} f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g^{(k)}, \phi_g^{(k)}) p(\boldsymbol{x}_i|\boldsymbol{\alpha}_g)} = \tau_{ij}$
- REP-BETA-CWM: $E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ig}|(\boldsymbol{x}_i, y_i)\right] = \frac{\pi_g^{(k)} f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g^{(k)}, v_g^{(k)}) p(\boldsymbol{x}_i|\boldsymbol{\alpha}_g)}{\sum_{g=1}^{G} \pi_g^{(k)} f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_g^{(k)}, v_g^{(k)}) p(\boldsymbol{x}_i|\boldsymbol{\alpha}_g)} = \tau_{ij}$

In the M-step, on the iteration $(k+1)$, the conditional expectation of $l(\boldsymbol{\theta}^{(k)})$ given the observed data say $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ is maximized with respect to $\boldsymbol{\theta}$.
The maximization of $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ with respect to the mixture weights is standard and can be obtained with Lagrangian multipliers as well as for the parameters $\boldsymbol{\alpha}_g$ (Ingrassia *et al.*, 2015). In case of generalized additive model the maximization of $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ with respect to $(\boldsymbol{\beta}_g, \zeta_g)$ is equivalent to the maximization problem of a GAM with the only difference that each observation i contributed to the log-likelihood with a known weight $\tau_{ig}^{(k)}$ (Wood, 2017, ch. 3), and the same for the beta regression. Finally, in the case of reparameterized beta regression, the maximization of $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ with respect to $(\boldsymbol{\beta}_g, v_g)$ is obtained solving the equations:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(k)}} = \sum_{i=1}^{n} \tau_{ig}^{(k)} \frac{\partial \ln\left(f_{rep}(y_i|x_i, \boldsymbol{\beta}_g^{(k)}, v_g)\right)}{\partial \boldsymbol{\beta}^{(k)}} = 0$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial v^{(k)}} = \sum_{i=1}^{n} \tau_{ig}^{(k)} \frac{\partial \ln\left(f_{rep}(y_i|x_i, \boldsymbol{\beta}_g^{(k)}, v_g)\right)}{\partial v^{(k)}} = 0$$

4. Some computational and operational aspects

In general, all computational details (EM initialization, convergence and model selection) are almost the same implemented in the package `flexCWM`. We start introducing the concept of three way deviance decomposition involved in the GAM-CWM.

4.1. Three way deviance decomposition

In the GAM-CWM, the value H that represent the basis dimension can be increased during the EM algorithm allowing the model to better adapt to the clusters presents in the data, in particular, controlling for the variance decomposition is possible to take a decision if it is the time to increase the size of the basis. Let $\hat{z}_{ig}^{(k)}$ the value of z_{ig} at the step k of EM algorithm; the total sum of squares of y at step k, say $TSS^{(k)}$, can be decomposed in the sum of two component: $WSS^{(k)}$ explains the within-groups deviance, while the $BSS^{(k)}$ explains the between-groups deviance:

$$\begin{aligned} TSS^{(k)} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} (y_{ij} - \bar{y}_g)^2 + \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} (\bar{y}_g - \bar{y})^2 \\ &= WSS^{(k)} + BSS^{(k)} \end{aligned}$$

where

$$\bar{y}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig}^{(k)} y_i}{\sum_{i=1}^{n} \hat{z}_{ig}^{(k)}}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Denoting with $\hat{\boldsymbol{\beta}}_g^{(k)}$ the vector of estimates at step k and $h(\mu_g)$ the link function (introduced before), the $WSS^{(k)}$ term can be decomposed again as

$$\begin{aligned} WSS^{(k)} &= \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} \left[y_i - h(\mu_{i,g}^{(k)}) + h(\mu_{i,g}^{(k)}) \right] - \bar{y}_g \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} \left[y_i - h(\mu_{i,g}^{(k)}) \right]^2 + \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} \left[h(\mu_{i,g}^{(k)}) - \bar{y}_g \right]^2 \\ &= WSS_f^{(k)} + WSS_e^{(k)} \end{aligned}$$

Summarizing, the total variability of Y can be explained by the latent group variable G in BSS and the withing-group sum of squares WSS. In turn, the WSS can be decomposed into WSS_f predictable from the covariates $\boldsymbol{X}$ and WSS_e not predictable from the covariates; summarizing:

$$\begin{aligned} TSS^{(k)} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} (\bar{y}_g - \bar{y})^2 + \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} \left[y_i - h(\mu_{i,g}^{(k)}) \right]^2 \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}^{(k)} \left[h(\mu_{i,g}^{(k)}) - \bar{y}_g \right]^2 \end{aligned}$$

The adaptive EM algorithm can be initialized with a low size of the basis dimension say H_0 and let k^* the step of the EM algorithm where $|BSS^{(k^*)} - BSS^{(k^*-1)}| < \epsilon$ (ϵ fixed sufficiently small). If this condition is verified it means that the BSS is stabilized over the iterations and the clusters have been identified by the chosen model. At this point it is possible to increase the size of the basis to a new value

$H_1 > H_0$ so that the model can specialize on the previously identified clusters and describe better local behaviors.
The function `plotDeviance` applied to a `gam_cwm()` object return the plot of these three quantities during the EM steps.

4.2. C-Index to evaluate cluster's compactness

A wide range of indices to measure different cluster's behaviors has been developed; in particular we investigate how a global index that describes the quality of clustering can be operatively used in the presented framework. Among different indices we selected the C-index (Hubert and Schultz, 1976), that is based on the Euclidean distances between the pairs of points inside each cluster. Let n_g the number of observation classified in the cluster g, in which there are $n_g(n_g-1)/2$ pairs of distinct points. Let N_W the number of such pairs ($N_W = \sum_{g=1}^{G} \frac{n_g(n_g-1)}{2}$) and let $N_T = n(n-1)/2$ the total number of pairs of distinct points in the whole data set. The C-index is defined as

$$C = \frac{S - S_{min}}{S_{max} - S_{min}} \tag{13}$$

where S is the sum of the N_W distances between all pairs of points inside each cluster, S_{min} is the sum of the N_W smallest distances between all the N_T pairs of points and finally S_{max} is the sum of the N_W largest distances between all the N_T pairs of points.
The index is limited to the interval [0, 1] and should be minimized in order to obtain compact clusters. Obviously, the index can be calculated considering the entire vector $(Y, \boldsymbol{X})$ of the covariates and the outcome (say $C_{\boldsymbol{X},Y}$) but we can obtain also two C-indices: one can describe the clusters from the point of view of the explicative covariates $\boldsymbol{X}$ ($C_{\boldsymbol{X}}$) and the second from the point of view of the outcome variable Y (C_Y). During the EM algorithm we can calculate $C_{\boldsymbol{X}}^{(k)}$ and $C_Y^{(k)}$ to evaluate how the observations are assigned to each cluster.
The function `plotC_Index` lets to visualize the evolution of this index during the EM.

4.3. EM initialization

As concern the EM initialization, the same initialization strategies available in flexCWM have been implemented in this package. The initializations are based on providing initial quantities $\boldsymbol{z}_i^{(0)} = (z_{i1}^{(0)}, ..., z_{iG}^{(0)})$, $i = 1, .., n$ at the first step of the algorithm.

- `"random.soft"`: the k values in $\boldsymbol{z}_i^{(0)}$ are generated from a uniform distribution (see *stat:runif()*) and normalized in order to sum to 1.

- `"random.hard"`: each $\boldsymbol{z}_i^{(0)}$ is extracted from a multinomial distribution with probabilities $(1/G, ..., 1/G)$ (see `vstat:rmultinom()`)
- `"manual"`: the values of $\boldsymbol{z}_i^{(0)}$ are manually provided by the user
- `"kmeans"`: the function `stat:kmeans()`provides hard values of $\boldsymbol{z}_i^{(0)}$
- `"mclust"`: the function `mclust:Mclust()` provides soft values for $\boldsymbol{z}_i^{(0)}$ fitting of an unconstrained mixture of Gaussian distributions.

4.4. Convergence criterion

Aitken acceleration (Aitken, 1926) can be used to take a decision about the convergence of the algorithm and based on this estimate it is possible to take a decision whether or not stopping the algorithm. It estimates the asymptotic maximum of the log-likelihood at each iteration:

$$a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}}$$

where $l^{(k)}$ is the log-likelihood value at iteration k. The asymptotic estimate of the likelihood (Böhning *et al.*, 1994) at iteration $k+1$ is given by

$$l_\infty^{(k+1)} = l^{(k)} + \frac{l^{(k+1)} - l^{(k)}}{1 - a^{(k)}}$$

In following simulations we stop the EM if $l_\infty^{(k+1)} - l^{(k)} < \epsilon$ (McNicholas, 2010; Punzo, 2012).

4.5. Model selection

The number of component G is treated as a fixed *a priori* quantity. In most applications this quantity is unknown, so it is possible to select the appropriate G considering likelihood-based method. Among different model selection criteria we consider the BIC and the Integrated Complete Likelihood. In mixture model the BIC as a model selection criterion is defined as:

$$BIC = 2l(\hat{\boldsymbol{\theta}}) - \eta \ln(n)$$

where η is the number of free parameters included into the model.
The ICL is given by

$$ICL = BIC + \sum_{i=1}^{n}\sum_{g=1}^{k} MAP(\hat{z}_{ig}) \ln(\hat{z}_{ig})$$

where

$$MAP = \begin{cases} 1 & max\{z_{ig}\}\, occurs\, in\, component\, g \\ 0 & otherwise \end{cases}$$

5. Package description

In this section we provide the main features implemented in the package with three applications with simulated and real data. A list of common arguments is available in Table 1 while in Table 2 are listed the specific parameters involved in the function `gam_cwm()`.

Table 1: List of arguments of the functions `gam_cwm()`, `beta_cwm()` and `rep_beta_cwm()`.

Argument	Description
`formula`	An optional object of class `formula`, the symbolic description of the model to be fitted.
`data`	An optional `data.frame` object containing the data to fit the model.
`Y`	An optional vector of outcome variable to describe.
`X`	An optional vector or matrix of covariates.
`groups`	An integer that specifies the number of components (cluster).
`Xnorm`, `Xbin`, `Xpois`, `Xmult`	Optional matrices containing concomitant variables with normal, binomial, Poisson and multinomial distributions.
`initialization`	Initialization strategy for the EM algorithm. It can be: `random.soft`, `random.hard`, `manual`, `kmeans` and `mclust`.
`start.z`	Matrix of dimension $(n \times k)$ of soft or hard initialization. This matrix is considered only if `initialization="manual"`.
`seed`	Seed for the random number generator in case of random initializations.
`iter.max`	Maximum number of EM iterations. Default value is 200.
`AIT_threshold`	Value of ϵ in the Aitken acceleration procedure. Default value is `1.0E04`.

Table 2: List of specific arguments for the functions `gam_cwm()`.

Argument	Description
`gamma`	The same of parameter available in `mgcw:gam()`; increase this beyond 1 to produce smoother models.
`increase_flex`	Logical; if `TRUE` then the basis dimension increases to the `max_basis_dimension` (see the parameter H in (2)) once the `BSS_threshold` has reached.
`max_basis_dimension`	The dimension of the basis used to represent the smooth term. See the parameter k in `mgcw:s()`.
`BSS_threshold`	If $\lvert BSS^{(k)} - BSS^{(k-1)} \rvert < BSSthreshold$ than the the parameter k in `mgcw:s()` is set equal to the `max_basis_dimension` up to convergence of EM.

Table 3: Other auxiliary functions.

Argument	Description
`getMetrics`	Returns main metrics of an estimated model: • `LLK`: log-likelihood. • `BIC`: Bayesian information criterion. • `ICL`: Integrated Completed Likelihood. • `TSS`: total sum of squares. • `BSS`: between-groups sum of squares. • `WSS`: within-groups sum of squares. • `WSS_f`: within-groups sum of squares explained by the chosen model. • `WSS_e`: within-groups sum of squares residual. • `C_index`: the C-index. • `C_index_X`: the C-index calculated on the covariates $\boldsymbol{X}$. • `C_index_Y`: the C-index calculated on Y.

Table 4: Other auxiliary functions.

Function	Description
`rUnimodalBeta`	Generate random data from a mixture of unimodal beta distributions. The parameters are: • `modes`: vector of modes. • `dispersion`: vector of dispersion parameters. • `weights`: mixture weights. • `n`: total sample size.
`plotDeviance`	This function plot the components of the total sum of squares during the estimation process. Actually available only for `gam_cwm()`. • `model`: estimated model object. • `add_plot`: if the resulting plot should be added to an existing plot. • `main`: title of the plot. • `flex_line`: plot a vertical line at the point where the flexibility if the model increases. • `iter`: plot the iterations of the EM from the first up to that specified in this parameter.
`plotC_Index`	This function plots the evolution of C-index during the EM algorithm.

5.1. A simulation with GAM-CWM

The first tutorial uses a generated dataset according to the parameters explained in Table 5 with the aim of explaining how to estimate a GAM-CWM and at the same time showing some limitations of GAM mixtures. This bi-variate dataset consists in 3 groups well separated with different shapes (Fig. 2): cluster 1 (red) can be described with a function of degree 2 because has a parabolic shape, cluster 2 (green) has a sinusoidal shape while cluster 3 (black) can be approximated with a straight line.

Table 5: Parameter's definition for artificial data simulation.

Parameter	Cluster 1 (red)	Cluster 2 (black)	Cluster 3 (green)
n	500	500	500
π_j	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
X	$Unif(15, 30)$	$N(15, \sigma = 3)$	$Unif(15, 30)$
$f_j(x)$	$20 + 2(x - 15)^2$	$200 + 10\, sin(x - 25)$	$400 - (x - 15)^2$
ϵ	$N(0, \sigma = 10)$	$N(0, \sigma = 10)$	$N(0, \sigma = 10)$

```
#-- Data generation from GAM CWM
R> n <- 200

R> x1 <- rnorm(n,15)
R> y1 <- 20 + 2*(x1-15)^2+rnorm(n,0,2)
R> g1 <- rep(1,n)

R> x2 <- rnorm(n,15,sd=3)
R> y2 <- 3*sin(x2-15)+rnorm(n,0,2)
R> g2 <- rep(2,n)

R> x3 <- rnorm(n,35,5)
R> y3 <- -1*(x3-35)+rnorm(n,0,2)
R> g3 <- rep(3,n)

R> d <- data.frame(y=c(y1,y2,y3),x=c(x1,x2,x3),g=c(g1,g2,g3))



#-- Estimation of GAM-CWM with adaptive EM
R> gam_cwm <- gam_cwm(y~s(x,bs="cr",k=3),data=d,
        Xcont="x",groups=3,
        initialization="random.soft",increase_flex=T,
```

```
        max_basis_dimension=10)

#-- Estimation of GAM-CWM with standard EM (not adaptive)
R> gam_cwm_standEM <- gam_cwm(y~s(x,bs="cr",k=20),data=d,
        Xcont="x",groups=3,
        initialization="random.soft",increase_flex=F)

#-- Parameter estimated by cluster
R> gam_cwm$estim
[[1]]

Family: gaussian
Link function: identity

Formula:
y ~ s(x, bs = "cr")

Estimated degrees of freedom:
3.01  total = 4.01

GCV score: 1.278723

[[2]]
...
```

The model can be estimated with the `gam_cwm()` function. Within each mixture component the function `mgcv:gam()` is called and for the specification of the model we need a GAM formula object, which provides the definition of a smooth function calling the function `mgcv:s()` thus, we can easy define the type of smooth function to use specifying the parameter `bs` (see `?s` for details). We estimated a model with three groups, starting the EM with a basis dimension fixed to 3 ($H = 3$).
In Fig. 3 it is straightforward to note that the fit of the adaptive EM is better: the BSS on the right is very high compared to the one on the left showing how the model is able to capture a large part of the variance explained by the latent variable. At the same time the WSS on the right is low as expected. Finally the vertical line in the graph on the right it is in correspondence of the iteration where the basis dimension increases (from $H_0 = 3$ to $H_1 = 10$).
The evolution of the values of C-index is plotted in Fig. 4 during the EM algorithm. Clearly the GAM-CWM with the adaptive EM reaches the minimum value possible of $C_{\boldsymbol{X}}$ and C_Y while the standard EM does not detect the clusters.
Finally, with the following code we can obtain the deviance-decomposition plot and the main metrics related to the estimated model.

```
#-- Plot of the three way deviance decomposition
R> plotDeviance(gam_cwm_standEM,F,"",T,1:40)
R> plotDeviance(gam_cwm,F,"",T,1:40)

R> getMetrics(mod_GAM)
       LLK         BIC         TSS         BSS         WSS       WSS_f
WSS_e
"-3325.49" "-6894.07" "73855.62"   "85.27%"   "14.73%"   "11.38%"
"3.31%"

R> getMetrics(mod_GAM_standEM)
       LLK         BIC         TSS         BSS         WSS       WSS_f
WSS_e
 "-3423.5" "-7281.98" "73855.62"   "37.89%"   "62.11%"    "58.4%"
"3.56%"

R> plotC_Index(gam_cwm_standEM)
R> plotC_Index(gam_cwm)
```

Figure 2: Scatter plot of generated data and estimated models. GAM-CWM with standard EM is on the left while GAM-CWM with adaptive EM on the right.

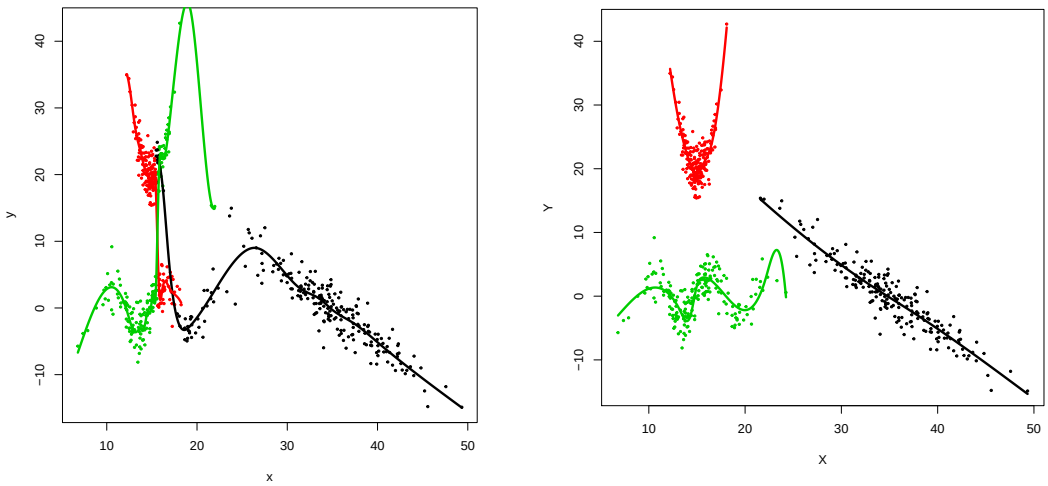


Figure 3: Three way deviance decomposition for non-adaptive EM (left) and adaptive EM (right).

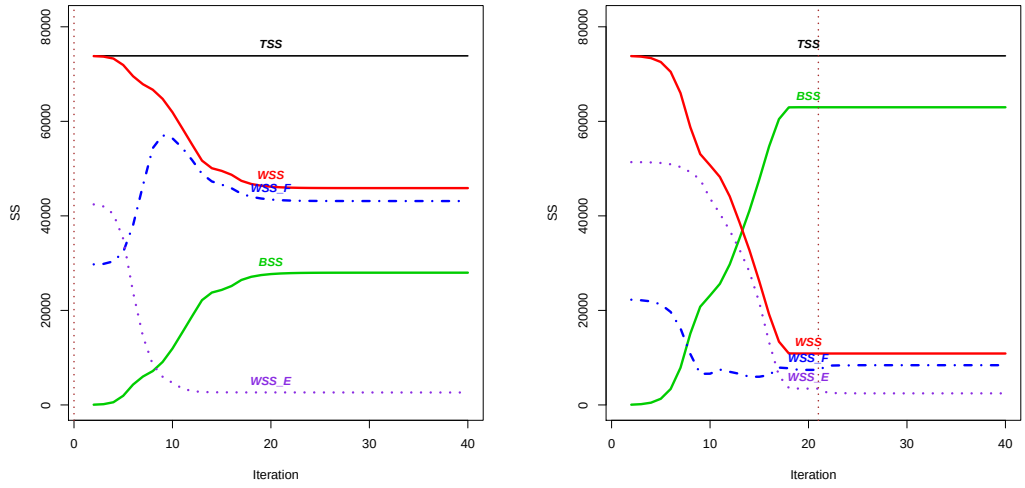
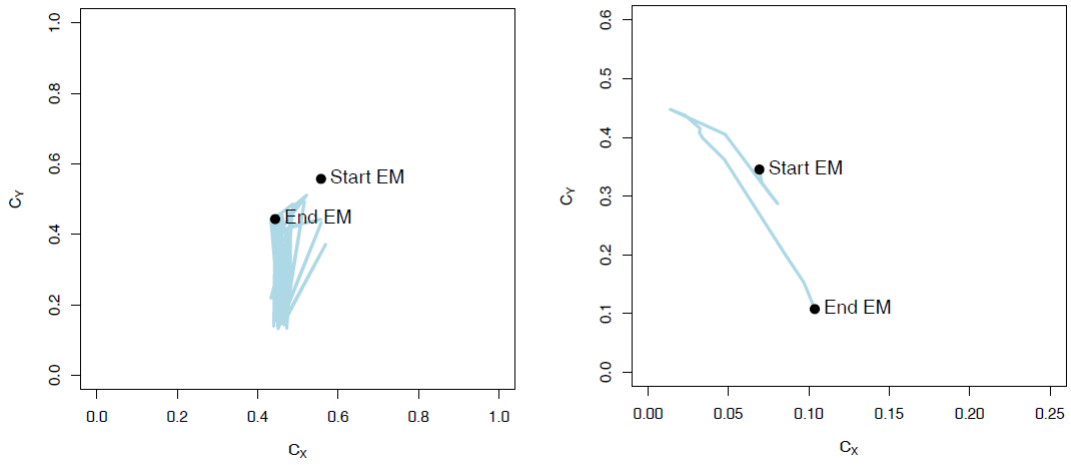


Figure 4: The evolution of C-index's values during the EM algorithm comparing the standard EM (on the left) with the adaptive EM (on the right).



5.2. Dataset AIRPORT

The airport dataset contains information about 22 Italian's airport from 2010 to 2015. Many different KPIs are collected for each airport, among which the total number of passengers, the share between national and international passengers, the different sources of revenues and other performance KPIs related to economics aspects. For this tutorial we choose two quantitative variables with the with the aim of showing how the GAM-CWM can also be used for exploratory analysis; the revenues per passenger is the outcome variable explained by the percentage of international passengers on the total. The aim of the analysis is to detect some clusters in the data and if some differences between airports are present.
First of all we start the analysis choosing the number of groups according to the BIC criterion. The best model according to the BIC criterion provides 4 groups. We start estimating the model defining:

```
#-- Data import
data(Airports)
d <- Airports
d$y <- d$REVENUES.PAX
d$x <- d$INTERNATIONAL.PAX

R> mod_GAM <- gam_cwm(y~s(x,bs="cr",k=2),data=d,Y=d$y,
        Xcont=d$x,groups=4,initialization="random.soft",
        increase_flex=T,max_basis_dimension=5)

R> mod_GAM4$BIC
[1] -1866.122

R> getMetrics(mod_GAM4)
       LLK         BIC         TSS         BSS         WSS       WSS_f
WSS_e
 "-857.73" "-1866.12"  "2871.48"    "24.8%"    "75.2%"   "55.89%"
"18.75%"
```

We can identify (Fig. 5) two compact clusters in black and blue. The labels are composed concatenating the name of the airport preceded by the macro area where the airport is located (North, Center or South of Italy).
The cluster in black is defined by airports only in the south of Italy characterized by low percentage of international passengers and a low revenues per passenger. Clusters in green and red are more heterogeneous. In this two clusters the revenues per passenger directly increases with the percentage of international passengers. Finally the cluster in red contains the most important Italian's airport, including Milano Malpensa and Roma Fiumicino, characterized by high levels of interna-

tional passengers.
The three way deviance decomposition (Fig. 6) show that the latent variable, in this example, is not able to capture a large part of the total sum of squares, but the model GAM-CWM lets to capture in a good way the relation between x and y ($WSS_f = 55.89\%$) within each cluster.
The evolution of C-index is available in Fig. 7, where the same model with different number of latent groups are compared. Models with 4 and 5 latent groups show good clustering performance and considering at the same time the BIC criterion which takes into account also for the number of parameters to be estimated, we can choose the model with 4 latent groups.

Figure 5: Estimated GAM-CWM with 4 clusters for Airport dataset.

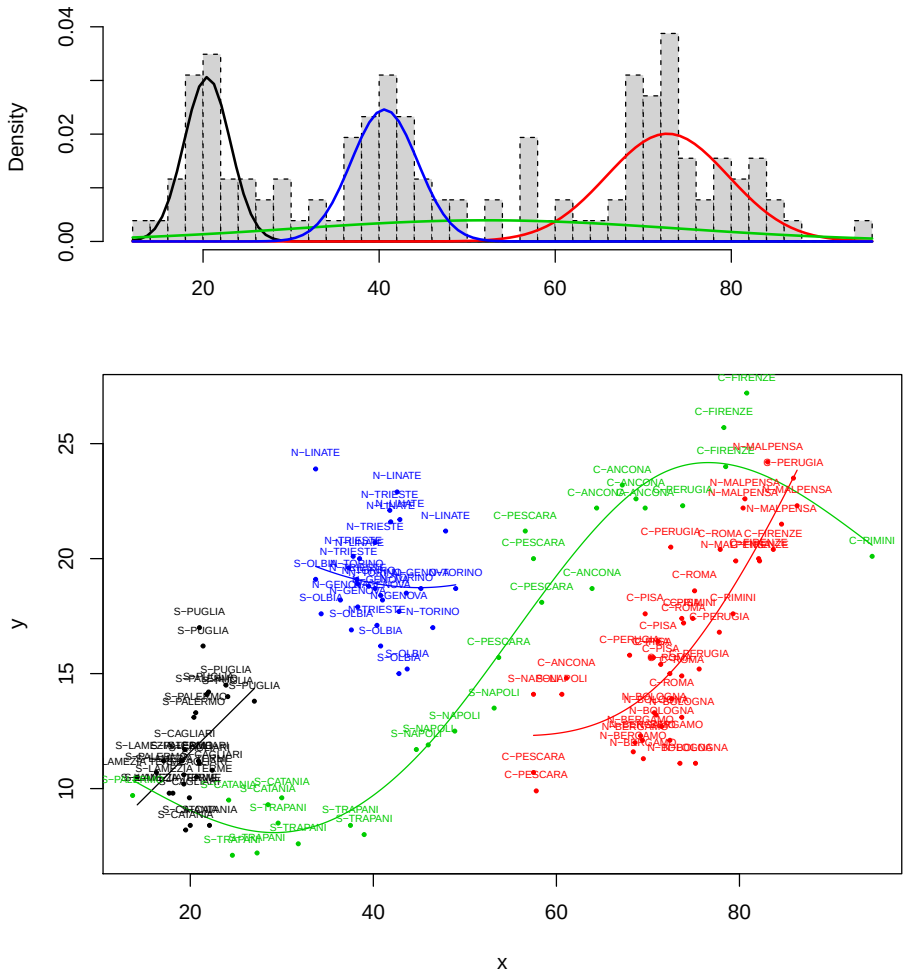


Figure 6: Three way deviance decomposition.

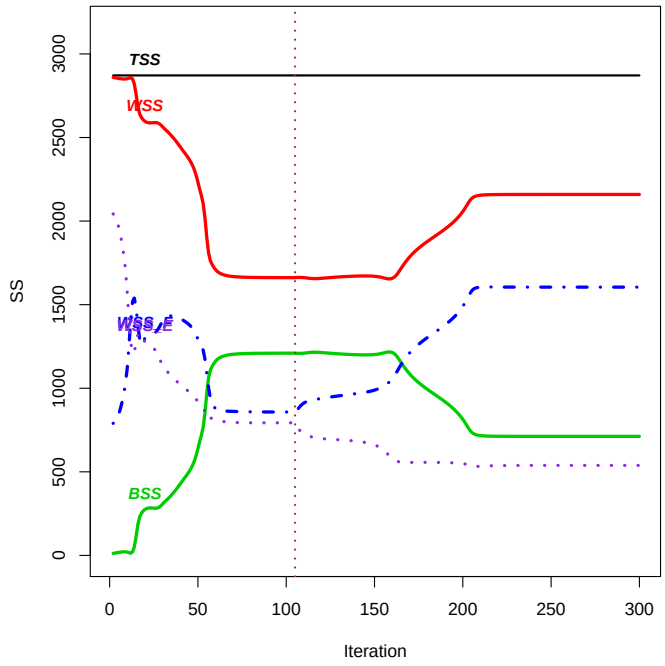
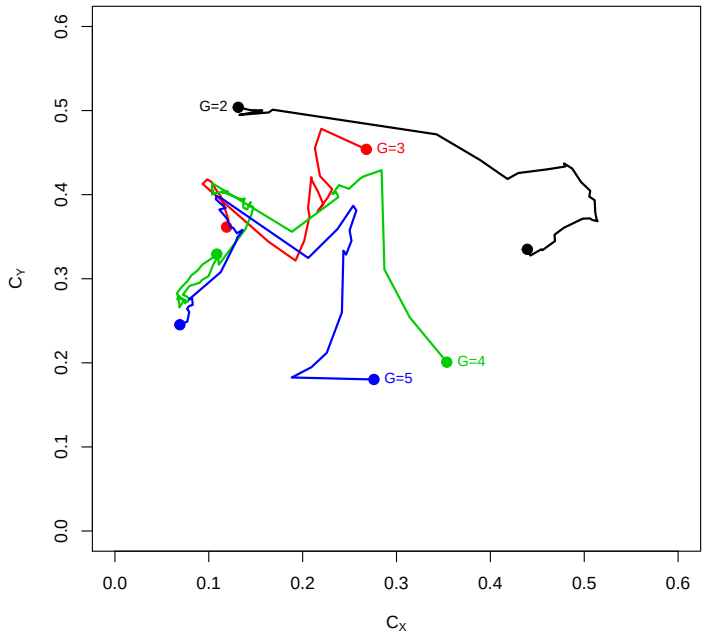


Figure 7: C-index.



5.3. A simulation with beta CWM and reparameterized beta CWM

This tutorial uses generated data (Fig. 8) from a unimodal mixture beta regression that can be easily generated with the function `rUnimodalBeta`. The model can be easily estimated with the following code:

```
#-- Generation of random dataset
R> set.seed(12345)
R> n <- 100
R> y1 <- rUnimodalBeta(modes=.2,dispersion=.05,weights=1,n)
R> y2 <- rUnimodalBeta(.8,.05,1,n)
R> x1 <- rnorm(n)
R> x2 <- rnorm(n,5)
R> d <- data.frame(x=c(x1,x2),y=c(y1,y2))
R> d$group <- c(rep(1,100),rep(2,100))

#-- Plot of the generated data
R> plot(d$x,d$y,pch=19,col=d$group,xlab="X",ylab="Y")


#-- Estimation of the unimodal beta mixture model
R> rep_cwm <- rep_beta_cwm(Y=d$y,X=d$x,Xcont=d$x,
                                groups=2,init="kmeans")

#-- Parameter estimated by cluster
R> rep_cwm$estim
          Cluster_1   Cluster_2
beta_0  -1.49814531  2.10002823
beta_1   0.09878130 -0.12072375
v        0.04768849  0.04726329

R> table(rep_cwm$cluster)
  1   2
100 100
```

The results of the functions `rep_beta_cwm` and `beta_cwm` is a list with the main quantities stored during the EM algorithm:

- `estim`: `data.frame` with the estimated parameters, respectively the coefficients and the dispersion parameters.
- `Xnorm_par`: list containing the estimated parameters for `Xnorm`.
- `Xbin_par`: list containing the estimated parameters for `Xbin`.
- `Xpois_par`: list containing the estimated parameters for `Xpois`.

- `Xmult_par`: list containing the estimated parameters for `Xmult`.
- `z`: matrix of estimated posterior probabilities.
- `cluster`: classification vector of length n.
- `LLK`: log-likelihood values stored during EM.
- `BIC`: Bayesian information criterion.
- `ICL`: integrated complete likelihood.

With the same data is possible to estimate a CWM with beta components. In this case in each component the function `betareg::betareg()` is called.

```
#-- Estimation of the beta CWM
R> beta_cwm <- beta_cwm(Y=d$y,X=d$x,Xcont=d$x,groups=2,
        init="random.soft")

R> mod$estim
[[1]]

Call:
betareg(formula = formula, data = data, weights = w, type = "ML")

Coefficients (mean model with logit link):
(Intercept)            x
    -1.5814       0.6369

Phi coefficients (precision model with identity link):
(phi)
47.36


[[2]]

Call:
betareg(formula = formula, data = data, weights = w, type = "ML")

Coefficients (mean model with logit link):
(Intercept)            x
    -0.9988       0.3802

Phi coefficients (precision model with identity link):
```

```
(phi)
8.437
```

To obtain details about the model estimated within each mixture component is possible to access to each element through the following command. It is important to note that the p-values are computed assuming that the posterior probabilities $\hat{z}_{ig}$ are given, then they should be considered in an exploratory manner.

```
R> summary(mod$estim[[1]])

Call:
betareg(formula = formula, data = data, weights = w, type = "ML")

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-1.3074 -0.5339 -0.0050  0.5361  1.2652

Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.58141    0.06110  -25.88   <2e-16 ***
x            0.63690    0.01894   33.62   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   47.364      7.695   6.155 7.52e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 110.8 on 3 Df
Pseudo R-squared: 0.752
Number of iterations: 29 (BFGS) + 3 (Fisher scoring)
```

5.4. USNEWS dataset

The last tutorial focus on a real application of the Beta CWM considering the USNEWS dataset (http://lib.stat.cmu.edu/datasets/colleges/) containing information on over 1300 American colleges and universities. This dataset is taken from the 1995 U.S. News & World Report's Guide to America's Best Colleges. The acceptance rates in American colleges have been chosen as response variable with instate tuition feed as covariates.
According to the BIC, the reparameterized beta CWM is more parsimonious about the choice of number of latent clusters, identifying 3 clusters instead of the 5 clusters that are identified by the CWM-beta.

```
#-- Estimation of the beta CWM
R> data(colleges)
R> d <- colleges
R> rep_beta_cwm <- rep_beta_cwm(Y=d$y,X=d$instate,
        Xcont=d$instate,groups=3,
        init="random.soft")

#-- Parameter estimated by cluster
R> rep_beta_cwm$estim
         Cluster_1 Cluster_2  Cluster_3
beta_0    3.801          1.770     2.571
beta_1   -0.193          -0.007          -0.409
v         0.160          0.106     0.171

R> table(rep_beta_cwm$cluster)
  1   2   3
104 712 445

R> plotModel(rep_beta_cwm)
```

Given the classification provided by the model the average values of acceptance rate by cluster are 0.75 (cluster 3 in green), 0.79 (cluster 2 in red) and 0.54 (cluster 1 in black). The covariate seems to affect the response in cluster 1 and in cluster 3, where increasing the values of instate tuition the acceptance rate decreases (Fig. 9). Clusters 1 and 2 include mostly private institutions while cluster 3 contains public institutions.

Figure 8: Generated data from a unimodal CWM mixture.

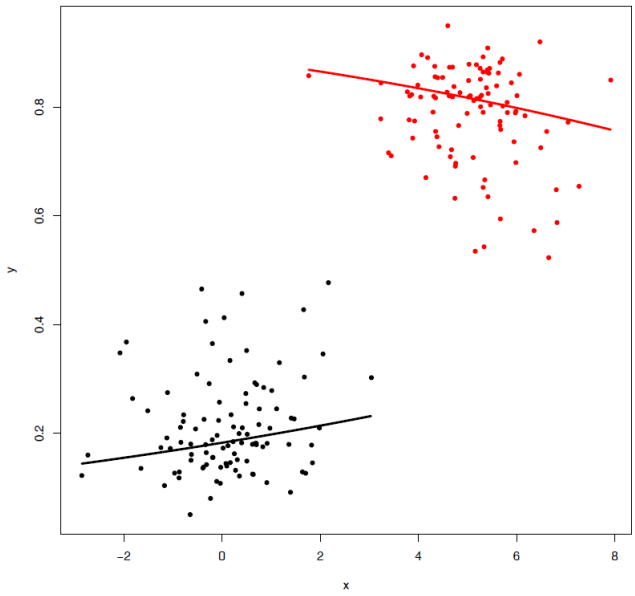
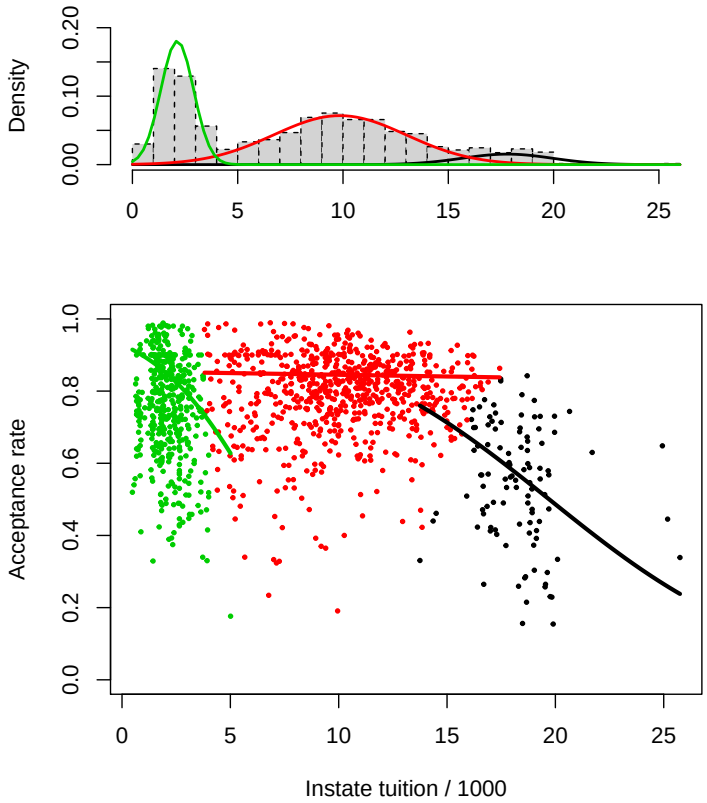


Figure 9: USNEWS model between instate tuition and acceptance rate.



6. Conclusions

Cluster Weighted Model is a new class of mixtures of regressions with random covariates. In this paper we have introduced an extension of `flexCWM` package to increase the chances of applications of the CWM where the response variable is beta-distributed and introducing the GAM models in the CWM framework.
The package offers three main functions to estimate these models and at the same time some additional tools are provided to facilitate the evaluation of the performances from different point of view.
Some future improvements concern the introduction of parallel methods for estimating the models in order to reduce the computational time and to develop new indices and new visualizations to facilitate the evaluation and interpretation of the results.

7. References

[1] Bagnato, L., & Punzo, A. (2013). Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. Computational Statistics, 28(4), 1571-1597.

[2] Dean, N., & Nugent, R. (2013). Clustering student skill set profiles in a unit hypercube using mixtures of multivariate betas. Advances In Data Analysis & Classification, 7(3).

[3] Ferrari, S. L., & Cribari-Neto, F. (2004). Beta Regression for Modeling Rates and Proportions. Journal Of Applied Statistics, 31(7), 799-815.

[4] Gershenfeld, N. (1997). Nonlinear Inference and Cluster-Weighted Modeling. Annals Of The New York Academy Of Sciences, 808(1).

[5] Grun, B., & Leisch, F. (2008). FlexMix Version 2. Finite Mixtures with Concomitant Variables Varying and Constant Parameters. Journal of Statistical Software, 28(4), 1-35.

[6] Hastie, T., & Tibshirani, R. (1987). Generalized Additive Models: Some Applications. Journal Of The American Statistical Association, 82(398).

[7] Ingrassia, S., Minotti, S., & Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. Journal Of Classification, 29(3), 363-401.

[8] Ingrassia, S., Minotti, S. C., & Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. Computational Statistics and Data Analysis, 71, 159-182.

[9] Ingrassia, S., Punzo, A., Vittadini, G., & Minotti, S. (2015). The Generalized Linear Mixed Cluster-Weighted Model. Journal Of Classification, 32(1), 85-113.

[10] R Development Core Team (2011). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria.

[11] Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2013). Clustering and Classification via Cluster-Weighted Factor Analyzers. Advances in Data Analysis and Classification., 7(1).

[12] Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2015). Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. Statistical Methods and Applications 24(4), 623-649.

[13] Wood, S. N. (2017). Generalized additive models: an introduction with R. Boca Raton. CRC Press.