Simultaneous semi-parametric estimation of clustering and regression

Matthieu Marbac[1], Mohammed Sedki[2], Christophe Biernacki[3], and Vincent Vandewalle[4]

[1]Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France
[2]Univ. Paris-Saclay and Inserm U1018, France
[3]Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille
[4]Inria, Univ. Lille, CHU Lille, ULR 2694 - METRICS : Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France

September 28, 2021

Abstract

We investigate the parameter estimation of regression models with fixed group effects, when the group variable is missing while group-related variables are available. This problem involves clustering to infer the missing group variable based on the group-related variables, and regression to build a model on the target variable given the group and eventually some additional variables. Thus, this problem can be formulated as the joint distribution modeling of the target and of the group-related variables. The usual parameter estimation strategy for this joint model is a two-step approach starting by learning the group variable (clustering step) and then plugging in its estimator for fitting the regression model (regression step). However, this approach is suboptimal (providing in particular biased regression estimates) since it does not make use of the target variable for clustering. Thus, we advise the use of a simultaneous estimation approach of both clustering and regression, in a semi-parametric framework. Numerical experiments illustrate the benefits of our proposition by considering wide ranges of distributions and regression models. The relevance of our new method is illustrated on real data dealing with problems associated with high blood pressure prevention. The proposed approach is implemented in the R package `ClusPred` available on `CRAN`. Supplemental materials containing the technical details and the R codes are available online.

Keywords: clustering; finite mixture; regression model; semi-parametric model.

1 Introduction

Regression models allow the relationship between some covariates and a target variable to be investigated. These models are defined by an equation on the conditional moment of the transformation of the noise. This transformation is generally the piecewise derivative of the loss function that defines the type of regression: mean, robust, quantile (Koenker and Bassett, 1978; Horowitz and Lee, 2005; Wei and Carroll, 2009), expectile (Newey and Powell, 1987; Ehm et al., 2016; Daouia et al., 2018).

The regression model with a fixed group effect is central within this generic paradigm. It considers that the intercept of the regression depends on the group from which the subject belongs (the intercept is common for subjects belonging to the same group but different for subjects belonging to different groups). However, in many applications, the group variable is not observed but other variables related to this variable are observed. For instance, suppose we want to investigate high blood pressure by considering the levels of physical activity among the covariates. In many cohorts, the level of physical activity of a subject is generally not directly available (because such a variable is not easily measurable) but many variables on the mean time spent doing different activities are available. Note that the regression model with a fixed group effect and a latent group variable is a specific mixture of regressions (Wang et al., 1996; Hunter and Young, 2012; Wu and Yao, 2016) where only the intercepts of the regressions are different among the components and where the mixture weights depend on some other variables. Moreover, the regression model with a fixed group effect and a latent group variable can be interpreted as a regression model with specific quantization of the variables that we use to estimate the group membership (see for instance Charlier et al. (2015) for the quantization in quantile regression).

The estimation of a regression model with a fixed group effect is generally performed using a *two-step approach* as for instance in epidemiology or in economics (Auray et al., 2015; Ando and Bai, 2016; Zhang et al., 2019). As a first step, a clustering on the individual based on the group-related variables is performed to obtain an estimator of the group. As a second step, the regression model is fitted by using the estimator of the group variable among

the covariates. The second step considers a regression model with measurement errors on the covariates. Indeed, the group variable is estimated in the clustering step with errors. Hence, it is well-known that the resulting estimators of the parameters of regression are biased (see for instance Carroll and Wand (1991); Nakamura (1992); Bertrand et al. (2017)). The bias depends on the accuracy of the clustering step. Note that, despite the fact that the target variable contains information about the group variable (and so is relevant for clustering), this information is not used in the two-step approach, leading to suboptimal procedures.

Some simultaneous approaches have been considered in the framework of latent variable models, such as latent class and latent profile analysis (Guo et al., 2006; Kim et al., 2016). In this framework, the authors introduce latent class and latent factor variables to explain the heterogeneity of observed variables. However, this approach does not focus on the conditional distribution of particular variable given other ones, and it is limited to a parametric framework. Another related reference is the work of Sammel et al. (1997), where the authors introduce a latent variable mixed effect model, which allows for arbitrary covariate effects, as well as direct modeling of covariates on the latent variable. Some other relevant references can be found in the field of concomitant variables (Dayton and Macready, 1988; Grün and Leisch, 2008; Vaňkátová and Fišerová, 2017), where some additional variables are used to locally adjust the weights of the mixture of regressions. These approaches are rather focused however, on the tasks of the mixture of regressions than on clustering data based on concomitant variables.

We propose a new procedure (hereafter referred to as the *simultaneous approach*) that simultaneously estimates the clustering and the regression models in a semi-parametric framework (Hunter et al., 2011) thus circumventing the limits of the standard procedure (biased estimators). We demonstrate that this procedure improves both the estimators of the partition and regression parameters. A full parametric setting is also presented, however if one of the clustering or regression models is ill-specified, its bias modeling could contaminate the results of the other. Thus, we focus on a semi-parametric mixture where the component densities are defined as a product of univariate densities (Chauveau et al., 2015; Zhu and Hunter,

2016; Zheng and Wu, 2019), which is identifiable if the univariate densities are linearly independent and if at least three variables are used for clustering (Allman et al., 2009). Note that, mixtures of symmetric distributions (Hunter et al., 2007; Butucea and Vandekerkhove, 2014) could also be considered in a similar way. Semi-parametric inference is achieved by a maximum smoothed likelihood approach (Levine et al., 2011) via a Maximization-Minimization (MM) algorithm (Hunter and Lange, 2004). Note that selecting the number of components in a semi-parametric mixture is not easy (Kasahara and Shimotsu, 2014; Kwon and Mbakop, 2020). However, in our context, the number of components can be selected according to the quality of the prediction of the target variable.

This paper is organized as follows. Section 2 introduces a general context where a statistical analysis requires both methods of clustering and prediction, and it presents the standard approach that estimates the parameters in two steps. Section 3 shows that a procedure that allows for a simultaneous estimation of the clustering and of the regression parameters generally outperforms the two-step approach. This section also briefly presents the simultaneous procedure on a parametric framework, then focuses on the semi-parametric frameworks. Section 4 presents numerical experiments on simulated data showing the benefits of the proposed approach. Section 5 illustrates our proposition for problems associated with high blood pressure prevention. Section 6 provides a conclusion and discussion about extensions. The mathematical details are presented in Appendix A.

2 Embedding clustering and prediction models

2.1 Data presentation

Let $(\mathbf{V}^\top, \mathbf{X}^\top, Y)^\top$ be the set of the random variables where $\mathbf{V} = (\mathbf{U}^\top, \mathbf{Z}^\top)^\top$ is a $d_V = d_U + K$ dimensional vector used as covariate for the prediction of the univariate variable $Y \in \mathbb{R}$, $\mathbf{X}$ is a d_X-dimensional vector and $\mathbf{Z} = (Z_1, \ldots, Z_K)^\top \in \mathcal{Z}$ is a categorical variable with K levels. The variable $\mathbf{Z}$ indicates the group membership such that $Z_k = 1$ if the subject belongs to cluster k and otherwise $Z_k = 0$. The realizations of $(\mathbf{U}^\top, \mathbf{X}^\top, Y)^\top$ are observed

but the realizations of $\mathbf{Z}$ are unobserved. Thus, $\mathbf{X}$ is a set of proxy variables used to estimate the realizations of $\mathbf{Z}$. Considering the high blood pressure example, Y corresponds to the diastolic blood pressure, $\mathbf{U}$ is the set of observed covariates (gender, age, alcohol consumption, obesity and sleep quality), $\mathbf{X}$ is the set of covariates measuring the level of physical activity and $\mathbf{Z}$ indicates the membership of a group of subjects with similar physical activity behaviors. The observed data are n independent copies of $(\mathbf{U}^\top, \mathbf{X}^\top, Y)^\top$ denoted by $\mathbf{u} = (\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_n^\top)^\top$, $\mathbf{x} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top$ respectively. The n unobserved realizations of $\mathbf{Z}$ are denoted by $\mathbf{z} = (\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_n^\top)^\top$.

2.2 Motivating example

We use the following example throughout the paper, which examines the general objective of high blood pressure prevention. Here, we focus on the detection of indicators related to the diastolic blood pressure (Y); see Berney et al. (2018) for the interest of the study. The indicators we wish to consider are the gender, the age, the alcohol consumption, the obesity, the sleep quality and the level of physical activity ($\mathbf{V}$). However, the level of physical activity ($\mathbf{Z}$) of a patient is not directly measured and we only have a set of variables which describes the physical activity ($\mathbf{X}$), such as practice of that recreational activity, hours spent watching TV, hours spent on the computer, *etc.* More details of the data are provided in Section 5. The study of the different indicators is performed using a regression model that explains the diastolic blood pressure with a set of covariates where one variable (the physical activity) was not directly observed. Information about this latter variable is available from other variables that do not appear in the regression.

2.3 Introducing the joint predictive clustering model

Regression model Let a loss function be $\mathcal{L}(\cdot)$ and $\rho(\cdot)$ its piecewise derivative. The loss function $\mathcal{L}$ allows the regression model of Y on $\mathbf{V}$ to be specified with a fixed group effect given by

$$Y = \mathbf{V}^\top \beta + \varepsilon \text{ with } \mathbb{E}[\rho(\varepsilon)|\mathbf{V}] = 0, \tag{1}$$

where $\beta = (\gamma^\top, \delta^\top)^\top \in \mathbb{R}^{d_V}$, $\gamma \in \mathbb{R}^{d_U}$ are the coefficients of $\mathbf{U}$, $\delta = (\delta_1, \ldots, \delta_K)^\top \in \mathbb{R}^K$ are the coefficients of $\mathbf{Z}$ (*i.e.,* the parameters of the group effect), and ε is the noise. Note that for reasons of identifiability, the model does not have an intercept. The choice of $\mathcal{L}$ allows many models to be considered and, among them, one can cite the mean regression (with $\mathcal{L}(t) = t^2$ and $\rho(t) = 2t$), the τ-quantile regression (with $\mathcal{L}(t) = |t| + (2\tau - 1)t$ and $\rho(\varepsilon) = \tau - \mathbb{1}_{\{\varepsilon \leq 0\}}$; Koenker and Bassett (1978)), the τ-expectile regression (with $\mathcal{L}(t) = |\tau - \mathbf{1}\{t \leq 0\}|t^2$ and $\rho(t) = 2t((1-\tau)\mathbf{1}\{t \leq 0\} + \tau \mathbf{1}\{t > 0\})$; Newey and Powell (1987)), *etc.*

The restriction on the conditional moment of $\rho(\varepsilon)$ given $\mathbf{V}$ is sufficient to define a model and allows for parameter estimation. However, obtaining a maximum likelihood estimate (MLE) needs specific assumptions on the noise distribution. For instance, parameters of the mean regression can be consistently estimated with MLE by assuming centered Gaussian noise. Similarly, the parameters of τ-quantile (or τ-expectile) regression can be consistently estimated with MLE by assuming that the noise follows an asymmetric Laplace (or an asymmetric normal) distribution (Yu and Moyeed, 2001; Xing and Qian, 2017). Hereafter, we denote the density of the noise ε by f_ε.

Clustering model The distribution of $\mathbf{X}$ given $Z_k = 1$ is defined by the density $f_k(\cdot)$. Therefore, the marginal distribution of $\mathbf{X}$ is a mixture model defined by the density

$$f(\boldsymbol{x}_i; \psi) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i), \tag{2}$$

where $\psi = \pi \cup \{f_1, \ldots, f_K\}$, $\pi = (\pi_1, \ldots, \pi_K)^\top$ is the vector of proportions defined on the simplex of dimension K (*i.e.,* $\pi_k > 0$ and $\sum_{k=1}^K \pi_k = 1$) and where f_k is the density of component k. In a parametric approach, f_k is assumed to be parametric so it is denoted by $f_k(\cdot; \alpha_k)$ where α_k are the parameters of component k. In a semi-parametric approach, some assumptions are required to ensure model identifiability (see for instance Chauveau et al. (2015)). In the following, the semi-parametric approaches are considered with the assumption that each f_k is a product of univariate densities (see Section 3.3).

Joint clustering and regression model The joint model assumes that $\mathbf{Z}$ explains the dependency between Y and $\mathbf{X}$ (*i.e.,* Y and $\mathbf{X}$ are conditionally independent given $\mathbf{Z}$) and that $\mathbf{U}$ and $(\mathbf{X}^\top, \mathbf{Z}^\top)^\top$ are independent. Moreover, the conditional distribution of $\mathbf{W} = (\mathbf{X}^\top, Y)^\top$ given $\mathbf{U}$ is also a mixture model defined by the density

$$f(\boldsymbol{w}_i|\boldsymbol{u}_i;\theta) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i) f_\varepsilon(y_i - \boldsymbol{u}_i^\top \gamma - \delta_k), \tag{3}$$

where $\theta = \pi \cup \{\phi_1, \ldots, \phi_K\} \cup \varsigma_\varepsilon$, ϕ_k grouping the parameters specific to component k (*i.e.,* the finite parameter δ_k and the infinite parameters f_k) and ς_ε grouping the parameters shared among the components (*i.e.,* the finite parameter γ and the infinite parameter f_ε), we have

$$\mathbb{E}[\rho(Y - \mathbf{U}^\top \gamma - \mathbf{Z}^\top \delta)|\mathbf{V}] = 0, \tag{4}$$

Note that (3) is a particular mixture of regressions models where the mixture weights are proportional to $\pi_k f_k(\boldsymbol{x}_i)$ (thus depending on covariates that do not appear in the regressions) and where only the intercepts (*i.e.,* $\delta_1, \ldots, \delta_K$) are different among the regressions. Contrary to Grün and Leisch (2008) who consider the density $f(y_i|\boldsymbol{u}_i, \boldsymbol{x}_i;\theta)$ thus focusing on the regression framework, here we propose considering the density $f(\boldsymbol{w}_i|\boldsymbol{u}_i;\theta)$ which balances the regression and the clustering frameworks.

Moment condition The following lemma gives the moment equation verified on the joint model and only consider observed variables in conditioning. It will be used later to justify the need for a simultaneous approach.

Lemma 1. *Assume that the model is defined by* (3)*, that the condition* (4) *holds true, that the covariance matrix of* $\boldsymbol{U}$ *has full rank and finally that* f_{kj} *and* f_ε *are strictly positive. Denoting* β_0 *as the single parameter satisfying* (4) *and* $r_k^{\boldsymbol{U},\boldsymbol{X},Y}(\boldsymbol{u},\boldsymbol{x},y) = \frac{\pi_k f_k(\boldsymbol{x}) f_\varepsilon(y - \boldsymbol{u}^\top\gamma - \delta_k)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{x}) f_\varepsilon(y - \boldsymbol{u}^\top\gamma - \delta_\ell)}$*, we have*

$$\forall k = 1, \ldots, K,\ \mathbb{E}[r_k^{\boldsymbol{U},\boldsymbol{X},Y}(\boldsymbol{U},\boldsymbol{X},Y)\rho(Y - \boldsymbol{U}^\top\gamma - \delta_k)|\,\boldsymbol{U},\boldsymbol{X}] = 0 \Longleftrightarrow \beta = \beta_0. \tag{5}$$

3 The proposed simultaneous estimation procedure

3.1 Limits of the standard two-step approach estimation

The aim is to explain the distribution of Y given $\mathbf{V} = (\mathbf{U}^\top, \mathbf{Z}^\top)^\top$ from an observed sample. A direct estimation of the model (1) is not doable because the realizations of $\mathbf{Z}$ are unobserved. The standard approach considers the following two-steps:

1. **Clustering step** Perform a clustering of $\mathbf{x}$ to obtain an estimated hard classification rule $\hat{r}^{\mathbf{X}} : \mathbb{R}^{d_X} \to \mathcal{Z}$ or an estimated fuzzy classification rule $\hat{r}^{\mathbf{X}} : \mathbb{R}^{d_X} \to \tilde{\mathcal{Z}}_K$ where $\tilde{\mathcal{Z}}_K$ is the simplex of size K.

2. **Regression step** Estimation of the regression parameters given the estimator of the group memberships $\hat{\beta}^{\hat{r}^{\mathbf{X}}} := (\hat{\gamma}^{\hat{r}^{\mathbf{X}}\top}, \hat{\delta}^{\hat{r}^{\mathbf{X}}\top})^\top$ with

$$\hat{\beta}^{\hat{r}^{\mathbf{X}}} = \arg\min_{\beta} \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{r}_k^{\mathbf{X}}(\boldsymbol{x}_i) \mathcal{L}(y_i - \boldsymbol{u}_i^\top \gamma - \delta_k),$$

   where $\hat{r}_k^{\mathbf{X}}(\boldsymbol{x}_i)$ is the element k of vector $\hat{r}^{\mathbf{X}}(\boldsymbol{x}_i)$. Note that $\hat{r}_k^{\mathbf{X}}(\boldsymbol{x}_i)$ is an estimator of the conditional probability that observation i belongs to cluster k given $\boldsymbol{x}_i$, if the fuzzy classification rule is used.

The following lemma states that the two-step approach is suboptimal. Indeed, even if the optimal classification rule on $\mathbf{X}$ is used, its expected good-classification rate is strictly smaller than that obtained by the best approach (see statement 1) and the estimators of the regression parameters are asymptotically biased (see statement 2).

Lemma 2. *Let the model be defined by* (3)-(4) *where* f_k *and* f_ε *are continuous and strictly positive where there exists* (k, ℓ) *such* f_k *and* f_ℓ *have no disjoint support and also* $\delta_k \neq \delta_\ell$*, and finally where* f_ε *is not constant. Suppose that* f_ε *defines a random variable with finite variance and that* $\boldsymbol{U}$ *has a full rank covariance matrix. Then,*

1. *Any hard classification rule* $\tilde{r}^X : \mathbb{R}^{d_X} \to \mathcal{Z}$ *is suboptimal in the sense that*

$$\mathbb{E}\left[\sum_{k=1}^{K} \tilde{r}_k^{\boldsymbol{X}}(\boldsymbol{X}) Z_k\right] < \mathbb{E}\left[\sum_{k=1}^{K} r_k^{\boldsymbol{U},\boldsymbol{X},Y}(\boldsymbol{U}, \boldsymbol{X}, Y) Z_k\right].$$

2. *Consider the quadratic loss, the best classification rule* $r^{\boldsymbol{X}}$ *computed on* $\boldsymbol{X}$ *and its associated estimator of the regression parameters* $\hat{\beta}^{r^{\boldsymbol{X}}}$*. The estimator* $\hat{\gamma}^{r^{\boldsymbol{X}}}$ *is asymptotically unbiased but the estimator* $\hat{\delta}^{r^{\boldsymbol{X}}}$ *is asymptotically biased with an asymptotic bias equals to* $\frac{\sum_{\ell=1}^{K} \Delta_{k\ell}\delta_\ell}{\sum_{\ell=1}^{K} \Delta_{k\ell}} - \delta_k$*, where* $\Delta_{k\ell} = \mathbb{E}[r_k^{\boldsymbol{X}}(\boldsymbol{X}) r_\ell^{\boldsymbol{X}}(\boldsymbol{X})]$.

Thus the clustering step provides a suboptimal classification rule because the classification neglects the information given by Y. Consequently, the regression step provides estimators that are asymptotically biased and implies fitting the parameters of a regression model with measurement errors in the covariates (for instance, considering the hard assignment, we have no guarantee of obtaining a perfect recovery of the partition, *i.e.*,, $\hat{r}^{\mathbf{X}}(\boldsymbol{x}_i) = \boldsymbol{z}_i$, for $i = 1, \ldots, n$). The measurement errors generally produce biases in the estimation. Finally, the quality of the estimated classification rule directly influences the quality of the estimator of the regression parameters.

3.2 Limits of a parametric simultaneous procedure

In this section, we consider a probabilistic approach with a parametric point-of-view. Thus, the family of distributions of each component k is supposed to be known and parameterized by α_k and thus we have $\phi_k = (\alpha_k^\top, \delta_k)^\top$. Moreover, the distribution of the noise f_ε is chosen according to the type of the regression under consideration (see the discussion in Section 2.3) and thus the parameters shared among the components are restricted to $\varsigma_\varepsilon = \gamma$. The aim of the simultaneous procedure can be achieved by maximizing the log-likelihood of $\mathbf{x}, \mathbf{y}$ given $\mathbf{u}$ with respect to θ

$$\ell(\theta; \mathbf{x}, \mathbf{y} \mid \mathbf{u}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i; \alpha_k) f_\varepsilon(y_i - \boldsymbol{u}_i^\top \gamma - \delta_k) \right).$$

Indeed, the maximum likelihood inference using $\ell(\theta; \mathbf{x}, \mathbf{y} \mid \mathbf{u})$ simultaneously allows for learning the classification rule based on $(\mathbf{X}^\top, Y)^\top$ and the regression coefficients. This function cannot be directly maximized, so we consider the complete-data log-likelihood with data $\mathbf{x}, \mathbf{y}$

and $\mathbf{z}$ given $\mathbf{u}$ defined by

$$\ell(\theta; \mathbf{x}, \mathbf{y}, \mathbf{z} \mid \mathbf{u}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln \left(\pi_k f_k(\boldsymbol{x}_i; \alpha_k) f_\varepsilon(y_i - \boldsymbol{u}_i^\top \gamma - \delta_k) \right).$$

The MLE $\hat{\theta}$ can be obtained via an EM algorithm presented in Appendix 1.2 of the supplementary materials. Moreover, if the model defined by (3)-(4) is identifiable, then

1. If all the parametric distributions are well-specified, then properties of the MLE imply that the classification rule is asymptotically optimal and $\hat{\beta}$ is asymptotically unbiased.

2. If at least one parametric distribution is misspecified, then the classification rule is generally asymptotically suboptimal and $\hat{\beta}$ is generally asymptotically biased.

It should be noticed that the distribution of the noise appears at the E-step and thus influences the classification rule. Hence, the classification rule is deteriorated if the distribution of the noise is misspecified. This is not the case when estimation is performed using the two-step approach, since clustering is performed prior to regression, and regression can still be unbiased if the moment condition (see Lemma 1) is well-specified. Thus, in the next section, we propose a semi-parametric approach that circumvents this issue because it does not assume a specific family of distributions for the noise and the components.

3.3 Advised simultaneous semi-parametric procedure

Semi-parametric model In this section, we consider the semi-parametric version of the model defined by (3) where the densities of the components are assumed to be a product of univariate densities (*i.e.,* $f_k(\boldsymbol{x}_i) = \prod_{j=1}^{d_X} f_{kj}(x_{ij})$). Therefore the parameters specific to component k, denoted by ϕ_k, are δ_k and $f_{k1}, \ldots, f_{kd_X}$. We have

$$f(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \theta) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \phi_k, \varsigma_\varepsilon) \text{ with } f_k(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \phi_k, \varsigma_\varepsilon) = \prod_{j=1}^{d_{\mathbf{X}}} f_{kj}(x_{ij}) f_\varepsilon(y_i - \boldsymbol{u}_i^\top \gamma - \delta_k).$$

A sufficient condition implying model identifiability is that the covariance matrix of $\mathbf{U}$ has full rank and that the marginal distribution of $\mathbf{X}$ is identifiable and thus a sufficient condition

is to consider linearly independent densities f_{kj}'s and $d_X \geq 3$ (Allman et al., 2009). Thus, if d_X is less than three, other semi-parametric mixture models should be considered to achieve clustering (*i.e.,* location-scale models; see Hunter et al. (2007); Chauveau et al. (2015)).

Smoothed log-likelihood Let $\mathcal{S}$ be the smoothing operator defined by $\mathcal{S}f_k(\boldsymbol{w} \mid \boldsymbol{u}; \phi_k, \varsigma_\varepsilon) = \int K_h(\boldsymbol{w} - \tilde{\boldsymbol{w}}) f_k(\tilde{\boldsymbol{w}} \mid \boldsymbol{u}; \phi_k, \varsigma_\varepsilon) d\tilde{\boldsymbol{w}}$, where $K_h(\boldsymbol{a}) = \prod_{j=1}^{d} K_h(a_j)$ with $\boldsymbol{a} \in \mathbb{R}^d$ and with $K_h(a_j)$ is a rescale kernel function defined by $K_h(a_j) = h^{-1}K(h^{-1}a_j)$ where h is the bandwidth. The estimation is achieved by maximizing the smoothed log-likelihood (Levine et al., 2011) defined by $\ell(\theta) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k \left(\mathcal{N} f_k\right)(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \phi_k, \varsigma_\varepsilon) \right)$ where $\left(\mathcal{N} f_k\right)(\boldsymbol{w} \mid \boldsymbol{u}; \phi_k, \varsigma_\varepsilon) = \exp\left\{\mathcal{S} \ln f_k(\boldsymbol{w} \mid \boldsymbol{u}; \theta_k)\right\} = \exp\left\{\int K_h(\boldsymbol{w} - \tilde{\boldsymbol{w}}) \ln f_k(\tilde{\boldsymbol{w}} \mid \boldsymbol{u}; \phi_k, \varsigma_\varepsilon) d\tilde{\boldsymbol{w}}\right\}$, subject to the empirical counterpart of (5):

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{f_k(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \phi_k, \varsigma_\varepsilon)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \phi_\ell, \varsigma_\varepsilon)} \rho(y_i - \boldsymbol{u}_i^\top \gamma - \delta_k) = 0.$$

Majorization-Minimization algorithm Parameter estimation is achieved via a Majorization-Minimization algorithm. Given an initial value $\theta^{[0]}$, this algorithm iterates between a majorization and a minimization step. Thus, an iteration $[r]$ is defined by

- Majorization step:

$$t_{ik}^{[r-1]} = \frac{\pi_k^{[r-1]} \left(\mathcal{N} f_k\right)(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \phi_k^{[r-1]}, \varsigma_\varepsilon^{[r-1]})}{\sum_{\ell=1}^{K} \pi_\ell^{[r-1]} \left(\mathcal{N} f_\ell\right)(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \rho_\ell^{[r-1]}, \varsigma_\varepsilon^{[r-1]})}.$$

- Minimization step:

$$\pi_k^{[r]} = \frac{1}{n} \sum_i t_{ik}^{[r-1]}, \beta^{[r]} = \arg\min_{\beta} \sum_{i,k} t_{ik}^{[r-1]} \mathcal{L}(y_i - \boldsymbol{u}_i^\top \gamma - \delta_k),$$

$$f_{kj}^{[r]}(a) = \frac{1}{n\pi_k^{[r]}} \sum_i t_{ik}^{[r-1]} K_h(x_{ij} - a) \text{ and } f_\varepsilon^{[r]}(a) = \frac{1}{n} \sum_{i,k} t_{ik}^{[r-1]} K_h(y_i - \boldsymbol{u}_i^\top \gamma^{[r]} - \delta_k^{[r]} - a),$$

then set $\phi_k^{[r]} = \gamma_k^{[r]} \cup \{f_{k1}^{[r]}, \ldots, f_{kd_X}^{[r]}\}$ and $\varsigma_\varepsilon^{[r]} = \delta^{[r]} \cup f_\varepsilon^{[r]}$.

The Majorization-Minimization algorithm is monotonic for the smoothed log-likelihood. It is a direct consequence of the monotony of the algorithm of Levine et al. (2011) where we

use the fact that, in order to satisfy the moment condition defined in (5) of Lemma 1, we must have $\beta^{[r]} = \arg\min_\beta \sum_{i=1}^n \sum_{k=1}^K t_{ik}^{[r-1]} \mathcal{L}(y_i - \boldsymbol{u}_i^\top \gamma - \delta_k)$.

As in Hunter and Young (2012), the majorization step is not explicit. However, because it only implies univariate integrals, it can be efficiently assessed by numerical approximations. Finally, bandwidth selection can be performed as usual for semi-parametric mixtures (see Chauveau et al. (2015)). However, as in any supervised problem, we can use the cross-validated accuracy of the prediction of Y for bandwidth selection.

4 Numerical experiments

4.1 Simulation setup

Data are generated such that $\mathbf{U}_i \sim \mathcal{N}_2(0, \mathbf{I}_2)$ and such that $(\mathbf{X}_i, Y_i)^\top$ given $\mathbf{U}_i$ follows a K-component mixture with proportions $\pi_k = 1/2$ if $k = 1$ and $\pi_k = 1/2(K-1)$ otherwise. The density of $\mathbf{X}_i$ given $\mathbf{Z}_i$ is a product of univariate densities such that $X_{ij} = \xi \mathbf{Z}_i^\top \kappa_j + \eta_{ij}$ where $\kappa_j = (\kappa_{j1}, \ldots, \kappa_{jK})^\top$, $\kappa_{jk} = 1$ if $k = (j \bmod K) + 1$ and $\kappa_{jk} = 0$ otherwise. Finally, we have $Y_i = \mathbf{U}_i^\top \gamma + \mathbf{Z}_i^\top \delta + \varepsilon_i$ with $\gamma = (1, 1)^\top$ and $\delta_k = 2\xi k$. η_{ij} and ε_i are independently drawn from a standard Gaussian distribution or a Student distribution with 3 degrees of freedom. The parameter ξ is tuned according to the distributions η_{ij} and ε_i and allows three theoretical misclassification rates (5%, 10% and 15%) to be considered. The approaches are compared with respect to the Mean Square Error (MSE) of the estimator of β and the Adjusted Rand Index (ARI) between the true and the estimated partition on 100 replicates. The semi-parametric approach is used with a fixed bandwidth $h = n^{-1/5}$. Note that a tuning of this window could be considered as in Chauveau et al. (2015).

4.2 Method comparison

Considering the quadratic loss, the experiment shows that the simultaneous procedure outperforms the standard two-step procedure, in both parametric and semi-parametric frameworks, where the parametric approaches assume that η_{ij} and ε_i are Gaussian. We consider

four scenarios: $\eta_{ij} \sim \mathcal{N}(0,1)$ for the first two scenarios and $\eta_{ij} \sim \mathcal{T}(3)$ for the last two scenarios, and $\varepsilon_i \sim \mathcal{N}(0,1)$ for the scenarios 1 and 3 and $\varepsilon_i \sim \mathcal{T}(3)$ for scenarios 2 and 4. Figure 1 presents the results obtained when $K = 3$ and $d = 6$. When the parametric model is well-specified (scenario 1), results are equivalent to those obtained by the semi-parametric model. Moreover, if at least one parametric assumption is violated (scenarios 2, 3 and 4), the results of the parametric approach are deteriorated even if the moment condition of the regression model is well-specified. Thus, we advise using the semi-parametric model if the family of the distributions is unknown to prevent the bias in the estimation.

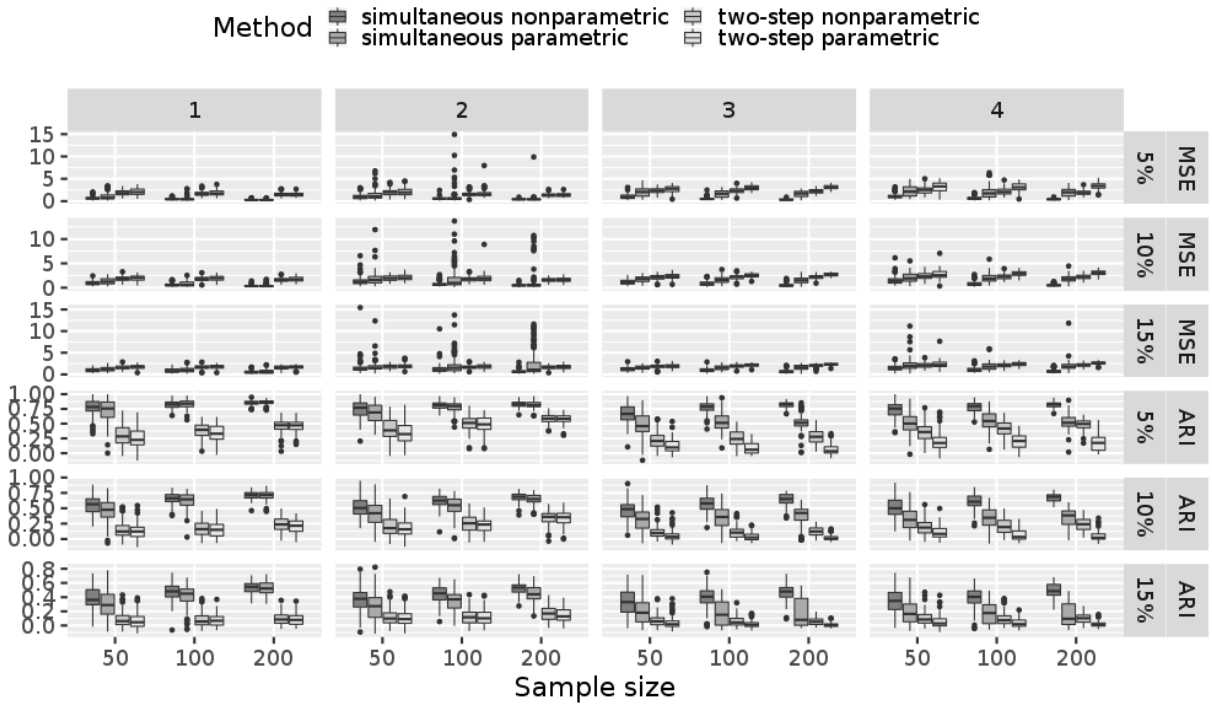

Figure 1: Boxplots of the MSE of the estimators of the regression parameters and ARI according to the theoretical misclassification (rows), the scenario (columns) and the sample size obtained when $K = 3$ and $d = 6$.

4.3 Robust regression

When the noise of a regression follows an heavy-tail distribution, robust regressions allow the estimators of the regression coefficients to be improved compared to the ordinary least square estimators. Despite this, with a suitable assumption on the noise distribution, the simultaneous parametric approach could consider such regressions. The parametric assumptions made on the noise distribution would be quite unrealistic (*e.g.,* Laplace distribution for the median regression). Thus, we now illustrate that the simultaneous approach can easily consider robust regressions, in a semi-parametric framework, and that the resulting estimators are better than those obtained with the quadratic loss. In this experiment, we consider scenario 4 (*i.e.,* η_{ij} and ε_i both follow independent $\mathcal{T}(3)$) and we consider different robust regressions (median, Huber with parameter 1 and logcosh). Figure 2 presents the results obtained when $K = 2$ and $d = 4$. It shows that the simultaneous approach improves the estimators (according to the MSE and the ARI) for any type of regression and any sample size. Moreover, robust regressions improve the accuracy of the estimator of the regression parameters. However, for this simulation setup, this improvement does not affect the accuracy of the estimated partitions.

4.4 Asymmetric losses

Expectile and quantile regressions respectively, generalize the mean and the median regression by focusing on the tails of the distribution of the target variable given the covariates. To illustrate the fact that the semi-parametric simultaneous method allows these regression models to be easily managed, data are generated with $K = 2$ and $d = 4$ such that $\eta_{ij} \sim \mathcal{N}(0, 1)$ and $\varepsilon_i \sim \mathcal{N}(-c_\tau, 1)$. The scalar c_τ is defined according to the regression model. Thus, c_τ is the 0.75-expectile, 0.9-expectile, 0.75-quantile and 0.9-quantile of the standard Gaussian distribution for the 0.75-expectile, 0.9-expectile, 0.75-quantile and 0.9-quantile regression respectively. Figure 3 shows that the simultaneous semi-parametric approach improves the estimators compared to those provided by the two-step approach.

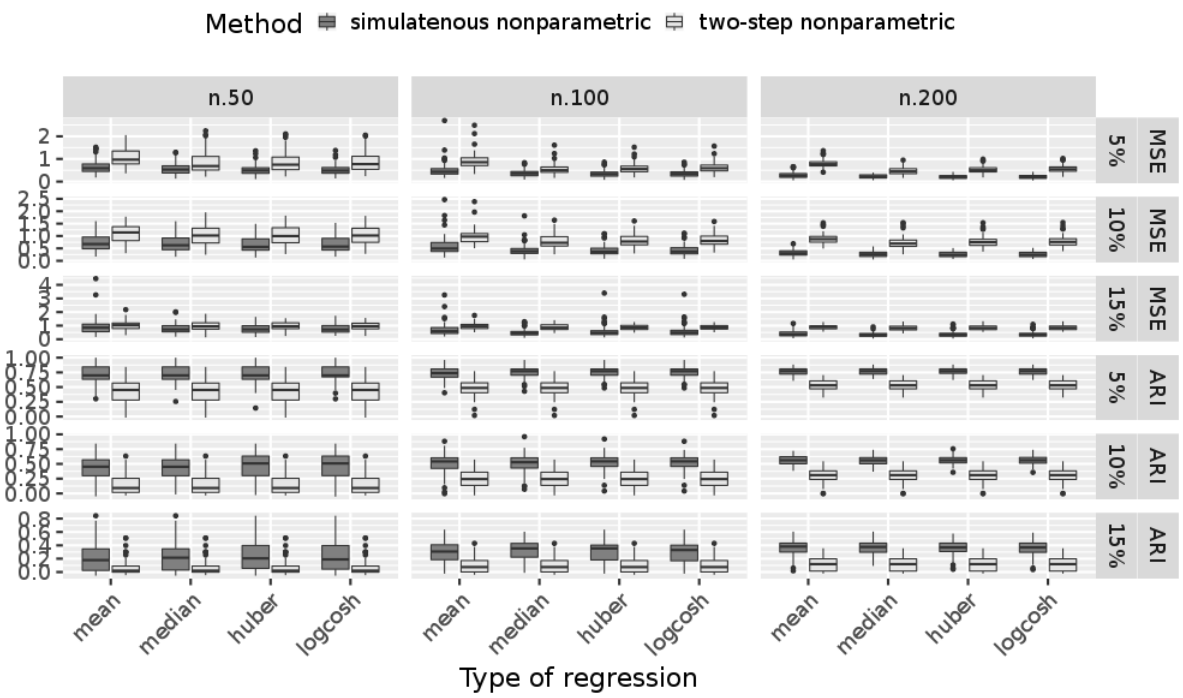

Figure 2: Boxplots of the MSE of the estimators of the regression parameters and ARI according to the theoretical misclassification (rows), sample size (columns) and the type of regression obtained when $K = 2$ and $d = 4$ for scenario 4.

5 High blood pressure prevention data set

Problem summary We consider the problem of high blood pressure prevention where we focus on the detection of indicators related to the diastolic blood pressure. The indicators we want to consider are gender, age, alcohol consumption, obesity, sleep quality and level of physical activity. However, the level of physical activity of a patient is not directly measured and we only have a set of variables that describe the physical activity. Thus, we want to cluster the subjects based on this set of variables to obtain patterns of similar physical activities and we want to use these patterns in the prediction of the diastolic blood pressure.

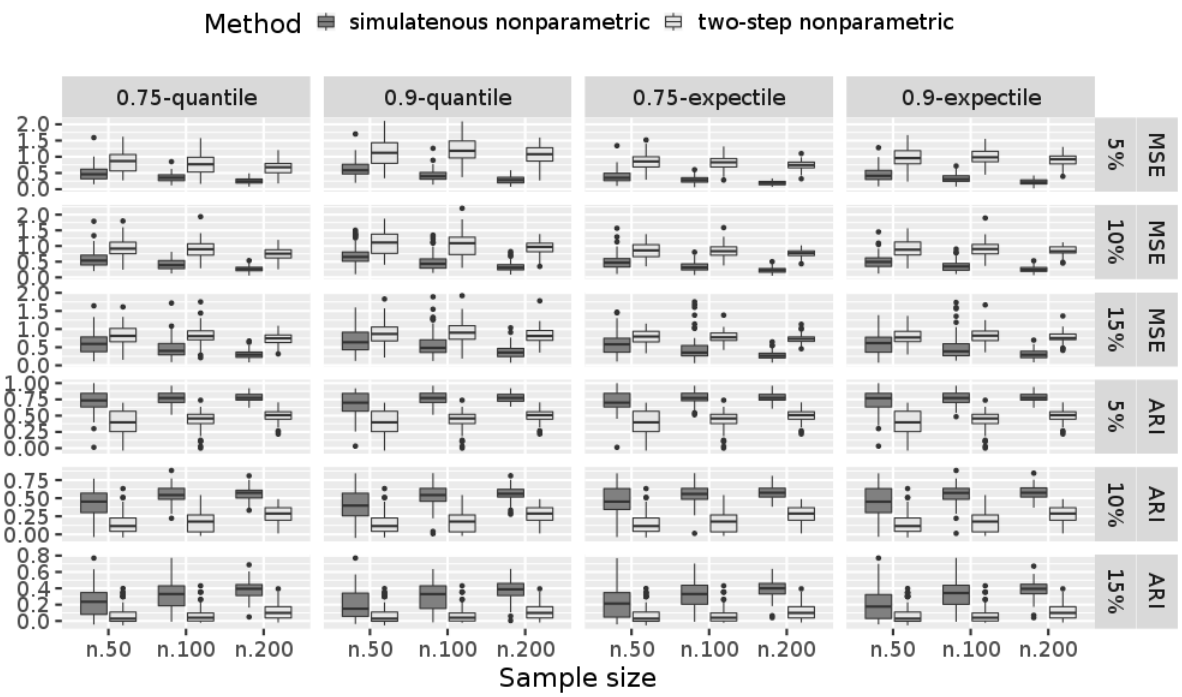

Figure 3: Boxplots of the MSE of the estimators of the regression parameters and ARI according to the theoretical misclassification (rows), the type of regression (columns) and regression obtained when $K = 2$ and $d = 4$.

Material and methods The data were obtained from National Health and Nutrition Examination Survey of 2011-2012[1]. The target variable is the *diastolic blood pressure* in mmHg (code BPXDI1). The seven covariates in $\mathbf{U}$ are *gender* which was equal to 1 for men et 0 for women (code RIAGENDR), *age* (RIDAGEYR), *alcohol* which indicates whether the subjects consume more than five drinks (for men) and four drinks (for women) of alcoholic beverages almost daily (computed from code ALQ151 and ALQ155), *obesity* which indicates if the body mass index is more than 30 (computed from code BMXBMI), *sleep* which indicates the number of hours of sleeping (computed from code SLD010H), *smoke* which indicates if the subjects used tobacco/nicotine in the last five days (code SMQ680) and *cholesterol* which indicates the total cholesterol in mg/dL (code LBXTC). All the subjects that had missing

[1]The data are freely downloadable at

https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2011

values for those variables were removed. Seven variables are used in $\mathbf{X}$ to evaluate the level of physical activity. Among these variables, five variables are binary and indicate whether the subject has a vigorous work activity (code PAQ605), whether the subject has a moderate work activity (code PAQ620), whether the subject usually travels on foot or by bike (code PAQ635), whether the subject has vigorous recreational activities (code PAQ650) and whether the subject has moderate recreational activities (code PAQ665). The two remaining variables in $\mathbf{X}$ have 7 levels and indicate the time spent watching TV (code PAQ710) and the time spent using a computer (code PAQ715). Finally, the studied population is composed of 2626 subjects between 18 and 60 years old. To investigate the performances of the different models, 67% of the sample (*i.e.,* 1760 subjects) is used for estimating the model parameters and 33% of the sample (*i.e.,* 866 subjects) is used for investigating the performances of the models. The smoothing is performed on the continuous variables with a Gaussian kernel and a bandwidth $h = n^{-1/5}$.

Results We present the main results of the application. Details used for the results interpretation are presented in Appendix 2 of the supplementary materials. We consider a proposed approach in a semi-parametric framework with a quadratic loss. According to the evolution of the smoothed log-likelihood with respect to the number of classes (see Figure 1 in Appendix 2 of the supplementary materials), the model is considered with $K = 3$ classes. To investigate the relevance of the activity level for explaining high blood pressure, we consider three models with a quadratic loss: the proposed approach in a semi-parametric framework (*regquadUZ-K3*), a regression model of Y on $\mathbf{U}$ (*regquadU*) with a selection of variables according to AIC (two variables are removed by the criterion: *alchohol* and *smoke*), a regression model of Y on $(\mathbf{U}^\top, \mathbf{X}^\top)^\top$ (*regquadUX*) with a selection of variables according to AIC (six variables are selected by the criterion: *gender*, *age*, *obesity*, *sleep*, *cholesterol* and the binary variable indicating whether the subject usually travels on foot or by bike). Considering the activity levels seems to be relevant for explaining high blood pressure, since the MSEs of the prediction obtained on the testing samples are 122.34, 122.72 and 122.81 for *regquadUZ-K3*, *regquadUX* and *regquadU* respectively. Thus, the approach allows the

information about the physical activity to be summarized and slightly improves the prediction accuracy. Note that a Shapiro-Wilk's normality test performed on the residuals of *regquadUZ-K3* has a pvalue less than 10^{-5} for the learning sample and 0.003 for the testing sample. Thus, the semi-parametric approach avoids the normality assumption which is not relevant for the residuals.

To prevent the variability due to outliers, we fit the proposed approach in a semi-parametric framework with the median loss and the logcosh loss. Again, evolution of the smoothed log-likelihood with respect to the number of classes, leads us to consider $K = 3$ classes for both losses. We now compare the results obtained by the proposed method with $K = 3$ classes in a semi-parametric framework with a quadratic loss, median loss (*regmedUZ-K3*) and logcosh loss (*reglogchUZ-K3*). The three models provided a similar partition since the ARIs between all the couples of partitions is more than 0.83. The regression parameters are presented in Table 1 of Appendix 2 of the supplementary materials. The signs of the coefficients are the same for the three losses. It appears that being a woman lessens the risk of high blood pressure while age, alcohol consumption, overweight, lack of sleeping and cholesterol increase high blood pressure. One can be surprised that the results claim that smoking limits the risk of high blood pressure, but this effect has already been revealed in Omvik (1996); Li et al. (2017). Note that the robust methods detect a more significant effect of alcohol, smoking and physical activity on high blood pressure. Moreover, they slightly change the prediction accuracy because the MSEs obtained on the testing sample are 122.88 and 123.00 for the median and the logcosh losses respectively.

We now interpret the clustering results provided by the median loss. Class 1 ($\pi_1 = 0.15$ and $\delta_1 = 59.06$) grouping the subjects having high physical activity is the smallest class and contains the subjects having recreational physical activities, traveling by foot or by bike, having no physical activity at work and spending few hours watching screens. Class 2 ($\pi_2 = 0.44$ and $\delta_2 = 59.29$) groups the subjects having few physical activities but spending little time watching screens. Class 3 ($\pi_3 = 0.37$ and $\delta_3 = 60.34$) groups those having some physical activities but spending a lot of time watching screens. These results show that

having moderate physical activities (recreational activities, traveling by bike or foot, not spending many hours watching screens) lessens the risk of high blood pressure.

6 Conclusion

In this paper, we propose an alternative to the two-step approach that starts by summarizing some observed variables by clustering and then fits a prediction model using the estimator of the partition as a covariate. Our proposition consists of simultaneously performing the clustering and the estimation of the prediction model to improve the accuracy of the partition and of the regression parameters. This approach can be applied to a wide range of regression models. Our proposition can be applied in a parametric and semi-parametric framework. We advise using the semi-parametric approach to avoid bias in the estimation (due to bias in the distribution modeling).

The quality of the prediction could be used as a tool for selecting the number of components and bandwidth, for semi-parametric mixtures. As in any regression problem, this criterion can also be used for selecting the variables (in the regression part but also in the clustering part). Thus, taking the regression into account is important in model selection for semi-parametric mixtures. Moreover, this could allow for a variable selection in clustering while this approach is only used in a parametric framework (Tadesse et al., 2005; Raftery and Dean, 2006). The semi-parametric approach has been presented by assuming that the components are products of univariate densities. However, the proposed approach can also be used by considering location scale symmetric distributions (Hunter et al., 2007) or by incorporating an independent component analysis structure (Zhu and Hunter, 2019). Moreover, we can easily relax the assumption that $(\mathbf{X}^\top, \mathbf{Z}^\top)$ is independent of $\mathbf{U}$. The crucial assumption of the model is the conditional independence of Y and $\mathbf{X}$ given $\mathbf{Z}$.

This approach has been introduced by considering only one latent categorical variable. However, more than one latent categorical variable explained by different sub-groups of variables of $\mathbf{X}$ could be considered. This extension is straightforward if the different sub-groups of variables of $\mathbf{X}$ are known. However, the cases where the sub-groups of variables

are also estimated (see the case of multiple partitions in clustering; Marbac and Vandewalle (2019)) could be considered in future work.

7 Supplementary Materials

Appendix: Technical details and details about the application on real data.

Codes.zip: Zipped archived containing all the R scripts related to the numerical experiments (see ReadMe.txt for details).

References

Allman, E. S., Matias, C., Rhodes, J. A., et al. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

Ando, T. and Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1):163–191.

Auray, S., Klutchnikoff, N., Rouviere, L., et al. (2015). On clustering procedures and nonparametric mixture estimation. *Electronic journal of statistics*, 9(1):266–297.

Berney, M., Burnier, M., and Wuerzner, G. (2018). Isolated diastolic hypertension: do we still have to care about it? *Revue medicale suisse*, 14(618):1607–1610.

Bertrand, A., Legrand, C., Léonard, D., and Van Keilegom, I. (2017). Robustness of estimation methods in a survival cure model with mismeasured covariates. *Computational Statistics & Data Analysis*, 113:3–18.

Butucea, C. and Vandekerkhove, P. (2014). Semiparametric mixtures of symmetric distributions. *Scandinavian Journal of Statistics*, 41(1):227–239.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):573–585.

Charlier, I., Paindaveine, D., and Saracco, J. (2015). Conditional quantile estimation through optimal quantization. *Journal of Statistical Planning and Inference*, 156:14 – 30.

Chauveau, D., Hunter, D. R., Levine, M., et al. (2015). Semi-parametric estimation for conditional independence multivariate finite mixture models. *Statistics Surveys*, 9:1–31.

Daouia, A., Girard, S., and Stupfler, G. (2018). Estimation of tail risk based on extreme expectiles. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2):263–292.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-Variable Latent-Class Models. *Journal of the American Statistical Association*, 83(401):173–178.

Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):505–562.

Grün, B. and Leisch, F. (2008). FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software*, 28(4).

Guo, J., Wall, M., and Amemiya, Y. (2006). Latent class regression on latent factors. *Biostatistics*, 7(1):145–163.

Horowitz, J. L. and Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, 100(472):1238–1249.

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

Hunter, D. R., Richards, D. S. P., and Rosenberger, J. L. (2011). *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P Hettmansperger, the Pennsylvania State University, USA, 23-24 May 2008*. World Scientific.

Hunter, D. R., Wang, S., and Hettmansperger, T. P. (2007). Inference for mixtures of symmetric distributions. *The Annals of Statistics*, pages 224–251.

Hunter, D. R. and Young, D. S. (2012). Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics*, 24(1):19–38.

Kasahara, H. and Shimotsu, K. (2014). Non-parametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):97–111.

Kim, M., Vermunt, J., Bakk, Z., Jaki, T., and Van Horn, M. L. (2016). Modeling predictors of latent classes in regression mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(4):601–614.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Kwon, C. and Mbakop, E. (2020). Estimation of the number of components of non-parametric multivariate finite mixture models. *Annals of Statistics (to appear)*.

Levine, M., Hunter, D. R., and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, pages 403–416.

Li, G., Wang, H., Wang, K., Wang, W., Dong, F., Qian, Y., Gong, H., Hui, C., Xu, G., Li, Y., et al. (2017). The association between smoking and blood pressure in men: a cross-sectional study. *BMC Public Health*, 17(1):797.

Marbac, M. and Vandewalle, V. (2019). A tractable multi-partitions clustering. *Computational Statistics & Data Analysis*, 132:167 – 179. Special Issue on Biostatistics.

Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics*, pages 829–838.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847.

Omvik, P. (1996). How smoking affects blood pressure. *Blood pressure*, 5(2):71–77.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178.

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):667–678.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617.

Vaňkátová, K. and Fišerová, E. (2017). The Evaluation of a Concomitant Variable Behaviour in a Mixture of Regression Models. *Statistika*, 97(4):16.

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

Wei, Y. and Carroll, R. J. (2009). Quantile regression with measurement error. *Journal of the American Statistical Association*, 104(487):1129–1143.

Wu, Q. and Yao, W. (2016). Mixtures of quantile regressions. *Computational Statistics & Data Analysis*, 93:162–176.

Xing, J.-J. and Qian, X.-Y. (2017). Bayesian expectile regression with asymmetric normal distribution. *Communications in Statistics-Theory and Methods*, 46(9):4545–4555.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Zhang, Y., Wang, H. J., and Zhu, Z. (2019). Quantile-regression-based clustering for panel data. *Journal of Econometrics*.

Zheng, C. and Wu, Y. (2019). Nonparametric estimation of multivariate mixtures. *Journal of the American Statistical Association*, pages 1–16.

Zhu, X. and Hunter, D. R. (2016). Theoretical grounding for estimation in conditional independence multivariate finite mixture models. *Journal of Nonparametric Statistics*, 28(4):683–701.

Zhu, X. and Hunter, D. R. (2019). Clustering via finite nonparametric ica mixture models. *Advances in Data Analysis and Classification*, 13(1):65–87.