



Simultaneous semi-parametric estimation of clustering and regression

Matthieu Marbac, Mohammed Sedki, Christophe Biernacki, Vincent Vandewalle

► To cite this version:

Matthieu Marbac, Mohammed Sedki, Christophe Biernacki, Vincent Vandewalle. Simultaneous semi-parametric estimation of clustering and regression. *Journal of Computational and Graphical Statistics*, In press. hal-03090573v1

HAL Id: hal-03090573

<https://inria.hal.science/hal-03090573v1>

Submitted on 29 Dec 2020 (v1), last revised 7 Jan 2022 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Simultaneous semi-parametric estimation of clustering and regression

Matthieu Marbac¹, Mohammed Sedki², Christophe Biernacki³, and Vincent Vandewalle⁴

¹Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France

²Univ. Paris-Sud and Inserm, France

³Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille

⁴Univ. Lille, CHU Lille, ULR 2694 - METRICS : Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France

December 28, 2020

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 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 claim for 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.

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 a 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 fit 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 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 a 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 no direct focus is taken to explain a particular variable given other ones, and the approach 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 effects model, which allows for arbitrary covariate effects, as well as direct modelling 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. But, these approaches are rather focused on the mixture of regressions task than on clustering data based on concomitant variables.

We propose a new procedure (hereafter referred to as the *simultaneous approach*) that estimates simultaneously the clustering and the regression models in a semi-parametric frameworks (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 model were ill-specified, its bias modeling could contaminate the results of the other one. Thus we focus on 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 Vandekerckhove, 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 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 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 $(V^\top, X^\top, Y)^\top$ be the set of the random variables where $V = (U^\top, Z^\top)^\top$ is a $d_V = d_U + K$ dimensional vector used as covariates for the prediction of the univariate variable $Y \in \mathbb{R}$, X is a d_X dimensional vector and $Z = (Z_1, \dots, Z_K)^\top \in \mathcal{Z}$ is a categorical variable with K levels. The variable 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 $(U^\top, X^\top, Y)^\top$ are observed but the realizations of Z are unobserved. Thus, X is a set of proxy variables used to

estimate the realizations of Z . Considering the high blood pressure example, Y corresponds to the diastolic blood pressure, U is the set of observed covariates (gender, age, alcohol consumption, obesity and sleep quality), X is the set of covariates measuring the level of physical activity and Z indicates the membership of a group of subjects with similar physical activity behaviours. The observed data are n independent copies of $(U^\top, X^\top, Y)^\top$ denoted by $\mathbb{U} = (u_1, \dots, u_n)^\top$, $\mathbb{X} = (x_1, \dots, x_n)^\top$ and $\mathbb{Y} = (y_1, \dots, y_n)^\top$ respectively. The n unobserved realizations of Z are denoted by $\mathbb{Z} = (z_1, \dots, z_n)^\top$.

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 (V). However, the level of physical activity (Z) of a patient is not directly measured and we only have a set of variables which describes the physical activity (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 V to be specified with a fixed group effect given by

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

where $\beta = (\gamma^\top, \delta^\top)^\top \in \mathbb{R}^{d_V}$, $\gamma \in \mathbb{R}^{d_U}$ are the coefficients of U , $\delta = (\delta_1, \dots, \delta_K)^\top \in \mathbb{R}^K$ are the coefficients of 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 $\rho(\varepsilon) = \tau - \mathbf{1}_{\{\varepsilon \leq 0\}}$; 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 V is sufficient to define a model and allows for parameter estimation. However, obtaining 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 a centred 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 X given $Z_k = 1$ is defined by the density $f_k(\cdot)$. Therefore, the marginal distribution of X is a mixture model defined by the density

$$f(x; \vartheta) = \sum_{k=1}^K \pi_k f_k(x), \quad (2)$$

where $\vartheta = \{\pi_k, f_k; k = 1, \dots, K\}$, $\pi_k > 0$ 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 Z explains the dependency between Y and X (*i.e.*, Y and X are conditionally independent given Z) and

that U and (X^\top, Z^\top) are independent. Moreover, the distribution of (X, Y) given U is also a mixture model defined by the density (noting $\theta = \{\vartheta\} \cup \{\delta_k; k = 1, \dots, K\} \cup \{\gamma, f_\varepsilon\}$)

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

where, for $k = 1, \dots, K$ we have

$$\mathbb{E}[\rho(Y - U^\top \gamma - \delta_k)|U, Z_k = 1] = 0, \quad (4)$$

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

Moment condition The following lemma gives the moment equation verified on the joint model. It will be used later to justify the need for a simultaneous approach.

Lemma 1. *Let an identifiable model defined by (3) and (4), where, for any x and k . Then, noting $r_k^{X,Y}(x, y) = \frac{\pi_k f_k(x) f_\varepsilon(y - u^\top \gamma - \delta_k)}{\sum_{\ell=1}^K \pi_\ell f_\ell(x) f_\varepsilon(y - u^\top \gamma - \delta_\ell)}$, $\beta = (\delta^\top, \gamma^\top)^\top$ is the single parameter satisfying*

$$\forall k = 1, \dots, K, \quad \mathbb{E}[r_k^{X,Y}(X, Y) \rho(Y - u^\top \gamma - \delta_k)|U, X] = 0. \quad (5)$$

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

1. **Clustering step** Perform a clustering of \mathbb{X} to obtain an estimated hard classification rule $\hat{r}^X : \mathbb{R}^{d_X} \rightarrow \mathcal{Z}$ or an estimated fuzzy classification rule $\hat{r}^X : \mathbb{R}^{d_X} \rightarrow \tilde{\mathcal{Z}}_K$ 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 $\hat{\beta}^{\hat{r}^X} := (\hat{\gamma}^{\hat{r}^X \top}, \hat{\delta}^{\hat{r}^X \top})^\top = \arg \min_{\beta} \sum_{i=1}^n \sum_{k=1}^K \hat{r}_k^X(x_i) \mathcal{L}(y_i - u_i^\top \gamma - \delta_k)$ where $\hat{r}_k^X(x_i)$ is the element k of vector $\hat{r}^X(x_i)$. Note that $\hat{r}_k^X(x_i)$ is an estimator of the conditional probability that observation i belongs to cluster k given x_i , if the fuzzy classification rule is used.

The following lemma states that the two-step approach produces asymptotically biased estimators of the partition and regression parameters, even if the optimal classification rule on X is used.

Lemma 2. *Let an identifiable model defined by (3) and (4), then*

1. *If classification rule \hat{r}^X converges to a classification rule \tilde{r}^X (i.e., the best classification rule based on X), thus \hat{r}^X is asymptotically suboptimal since $\mathbb{E} \left[\sum_{k=1}^K \tilde{r}_k^X(X) Z_k \right] < \mathbb{E} \left[\sum_{k=1}^K r_k^{X,Y}(X, Y) Z_k \right]$, with $\tilde{r}_k^X(x) \propto \pi_k f_k(x)$.*
2. *Let f_ε defines a random variable with finite variance and U to have a covariance matrix with non-zero eigenvalues. Then, considering the quadratic loss and a consistent fuzzy classification rule \hat{r}^X , the estimator $\hat{\gamma}^{\hat{r}^X}$ is asymptotically unbiased but the estimator $\hat{\delta}^{\hat{r}^X}$ is asymptotically biased since $\lim_{n \rightarrow \infty} \text{bias}(\hat{\delta}_k^{\hat{r}^X}) = \frac{\sum_{\ell=1}^K \Delta_{k\ell} \delta_\ell}{\sum_{h=1}^K \Delta_{kh}} - \delta_k$, where $\Delta_{k\ell} = \mathbb{E}[r_k^X(X) r_\ell^X(X)]$.*

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 having a perfect recovery of the partition, i.e., $\hat{r}^X(x_i) = z_i$, for $i = 1, \dots, n$). 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 . 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). The aim of the simultaneous procedure can be achieved by maximizing the log-likelihood of \mathbb{Y}, \mathbb{X} given \mathbb{U} with respect to θ

$$\ell(\theta; \mathbb{Y}, \mathbb{X} \mid \mathbb{U}) = \sum_{i=1}^n \ln \left(\sum_{k=1}^K \pi_k f_k(x_i; \alpha_k) f_\varepsilon(y_i - u_i^\top \gamma - \delta_k) \right),$$

where $\theta = (\pi_1, \dots, \pi_K, \alpha_1^\top, \dots, \alpha_K^\top, \beta^\top)$ groups all the model parameters. Indeed, the maximum likelihood inference using $\ell(\theta; \mathbb{Y}, \mathbb{X} \mid \mathbb{U})$ allows for similarly learning the classification rule based on (X, Y) and the regression coefficients. This function cannot be directly maximized so we consider the complete-data log-likelihood with data $(\mathbb{Y}, \mathbb{X}, \mathbb{Z})$ given \mathbb{U} defined by

$$\ell(\theta; \mathbb{Y}, \mathbb{X}, \mathbb{Z} \mid \mathbb{U}) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \ln (\pi_k f_k(x_i; \alpha_k) f_\varepsilon(y_i - u_i^\top \gamma - \delta_k)).$$

The MLE $\hat{\theta}$ can be obtained via an EM algorithm presented in Appendix A.2. Let an identifiable model defined by (3) and (4), 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.

Let notice 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 model defined by (3) where the densities of the components are assumed to be a product of univariate densities. Thus, we have

$$f(y, x \mid u; \theta) = \sum_{k=1}^K \pi_k f_k(y, x \mid u; \theta) = \sum_{k=1}^K \pi_k \prod_{j=1}^{d_X} f_{kj}(x_j) f_\varepsilon(y - u^\top \gamma - \delta_k), \quad (6)$$

where θ groups all the finite and infinite parameters and β is such that (5) holds. A sufficient condition implying identifiability for model (6) is that the marginal distribution of 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). For sake of simplicity we will note $w = (x^\top, y)^\top$ with $w \in \mathbb{R}^{d_X+1}$, such that $f(y, x \mid u; \theta) = \sum_{k=1}^K \pi_k f_k(w \mid u; \theta)$.

Smoothed log-likelihood Let \mathcal{S} be the smoothing operator defined by $\mathcal{S}f_k(w \mid u; \theta) = \int K_h(w - \tilde{w}) f_k(\tilde{w} \mid u; \theta) d\tilde{w}$, where $K_h(a) = \prod_{j=1}^d K_h(a_j)$ with $a \in \mathbb{R}^d$ 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. For the semi-parametric approach, the estimation can be performed by maximizing the smoothed log-likelihood (Levine et al., 2011) defined by $\ell(\theta) = \sum_{i=1}^n \ln \left(\sum_{k=1}^K \pi_k (\mathcal{N}f_k)(w \mid u; \theta) \right)$ subject to the empirical counterpart of (5), namely $\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \frac{f_k(w \mid u; \theta)}{\sum_{\ell=1}^K \pi_\ell f_\ell(w \mid u; \theta)} \rho(y_i - u_i^\top \gamma - \delta_k) = 0$, where $(\mathcal{N}f_k)(w \mid u; \theta) = \exp \{ \mathcal{S} \ln f_k(w \mid u; \theta) \} = \exp \left\{ \int K_h(w - \tilde{w}) \ln f_k(\tilde{w} \mid u; \theta) d\tilde{w} \right\}$.

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

- Majorization step: $t_{ik}^{[r-1]} \propto \pi_k^{[r-1]} \left(\mathcal{N} f_k^{[r-1]} \right) (w_i \mid u_i; \theta^{[r-1]}).$
- Minorization step:
 1. Updating the parametric elements

$$\pi_k^{[r]} = \frac{1}{n} \sum_i t_{ik}^{[r-1]} \text{ and } \beta^{[r]} = \arg \min_{\beta} \sum_{i,k} t_{ik}^{[r-1]} \rho(y_i - u_i^\top \gamma - \delta_k).$$

2. Updating the nonparametric elements

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

The Majorization-Minorization 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 $\beta^{[r]} = \arg \min_{\beta} \sum_{i=1}^n \sum_{k=1}^K t_{ik}^{[r-1]} \rho(y_i - u_i^\top \gamma - \delta_k).$

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, like 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 from a bicomponent mixture with equal proportions (*i.e.*, $K = 2$ and $\pi_1 = \pi_2 = 1/2$) such that the density of X_i given Z_i , is a product of univariate densities. For $j = 1, \dots, 4$, $X_{ij} = (-1)^{Z_{i1}} \xi + \eta_{ij}$ where the group variable Z_i is sampled according to a multinomial distribution of parameters π_1 and π_2 , and η_{ij} are independent and identically distributed random variables. Moreover, we have $U_i \sim \mathcal{N}_2(0, \mathbf{I}_2)$ and $Y_i = (Z_{i1}, Z_{i2}, U_{i1}, U_{i2})\beta + \varepsilon_i$, with $\beta = (-1, 1, 1, 1)^\top$. Different distributions are used for

η_{ij} and ε_i . 100 samples are generated for each case, and the parameter ξ , controlling the class overlap, is tuned to have a theoretical misclassification of 0.10. The approaches are compared with the MSE of the coefficient of the regression β and the adjusted rand index (ARI) between the true and the estimated partition. The semi-parametric approach is applied with a fixed bandwidth $h = n^{-1/5}$, this choice is considered for sake of simplicity, the tuning of this window could be considered as in Chauveau et al. (2015).

4.2 Method comparison

In this experiment, we show that the simultaneous procedure outperforms the standard two-step procedure, in both parametric and semi-parametric framework. In the case where the parametric model is well-specified, we show that its results are equivalent to those obtained by the semi-parametric model. Moreover, we show that if one distribution is misspecified (component distribution or noise distribution), the results of the parametric approach are deteriorated even if the moment condition of the regression model is well-specified. Thus, we advise to use the semi-parametric model if the family of the distributions is unknown to prevent the bias in the estimation.

In this experiment, we focus on the quadratic loss. We consider one case where the parametric model is well-specified (case-1) where η_{ij} and ε_i follow standard Gaussian distributions. We also consider one case where the parametric model is misspecified on the noise distribution (case-2) where η_{ij} follows a standard Gaussian distribution and $\varepsilon_i = \tau_i - 1$ with $\tau_i \sim \mathcal{Exp}(1)$, and one case where the parametric model is misspecified on the component distributions (case-3) where η_{ij} follows a Student distribution with 3 degrees of freedom and ε_i follows a standard Gaussian distribution. Finally, we consider one case where the parametric model is misspecified on both of the component and noise distributions (case-4), where η_{ij} and ε_i follows Student distributions with 3 degrees of freedom. Figure 1 shows the MSE of the regression parameters and the ARI obtained.

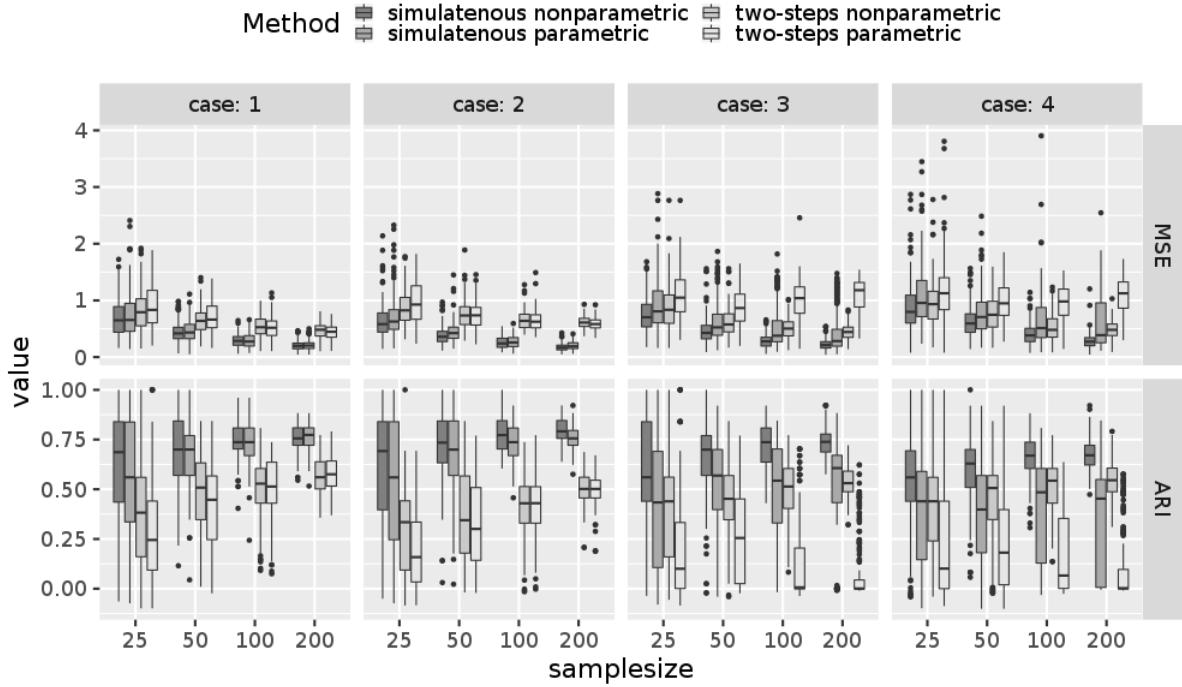


Figure 1: Boxplots of the MSE of the estimators of the regression parameters and ARI obtained by the simultaneous and two-step methods, in a parametric and semi-parametric framework, on 100 samples of size n .

4.3 Robust regression

When the noise of a regression follows an heavy-tail distribution, robust regressions (median, Huber and logcosh regressions) allow for improvement of the estimators of the regression coefficients compare to the ordinary least square estimators. Despite that, with a suitable assumption on the noise distribution, the simultaneous parametric approach could consider such regressions, the parameteric 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 that η_{ij} and ε_i follow independent Student distributions

with three degrees of freedom. Figure 2 shows the MSE of the regression parameters and the ARI obtained by the mean regression, the median regression, the Huber regression with parameter 1 and the logcosh regression. Results show 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.

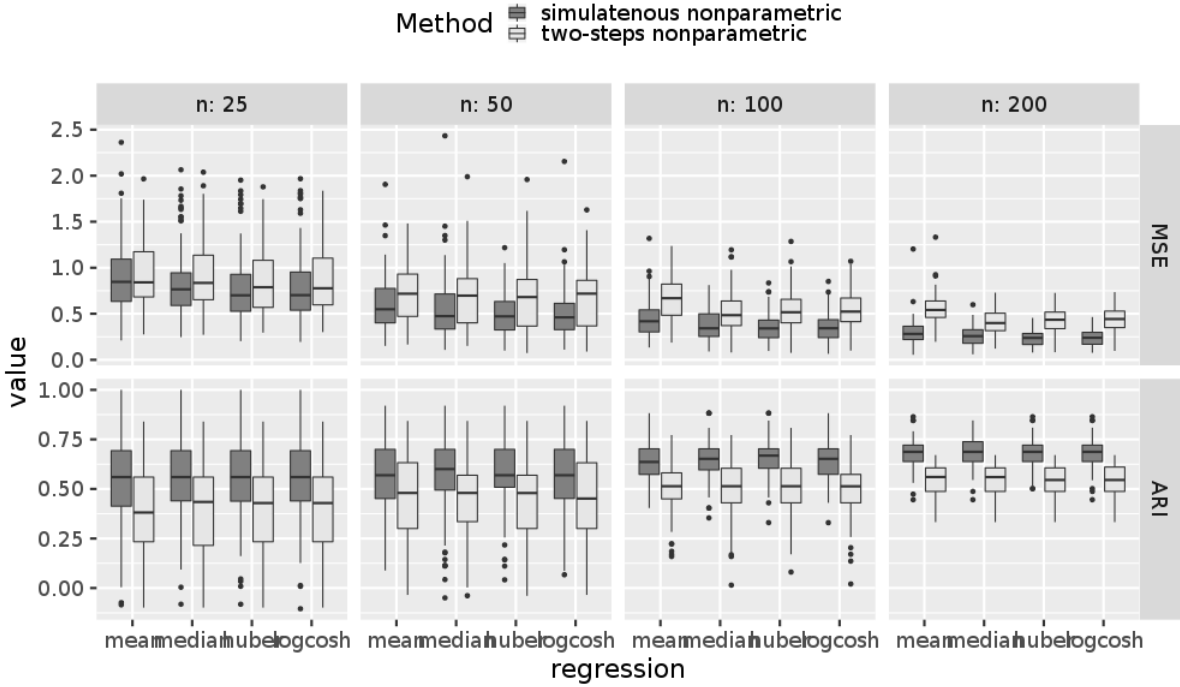


Figure 2: Boxplots of the MSE of the estimators of the regression parameters and ARI obtained by different regressions (mean, median, Huber with parameter 1 and logcosh) estimated with the simultaneous and two-step semi-parametric methods on 100 samples of size n generated with heavy-tail distributions.

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 for easily considering these regression models, data are generated 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 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 compare to those provided by the two-step approach.

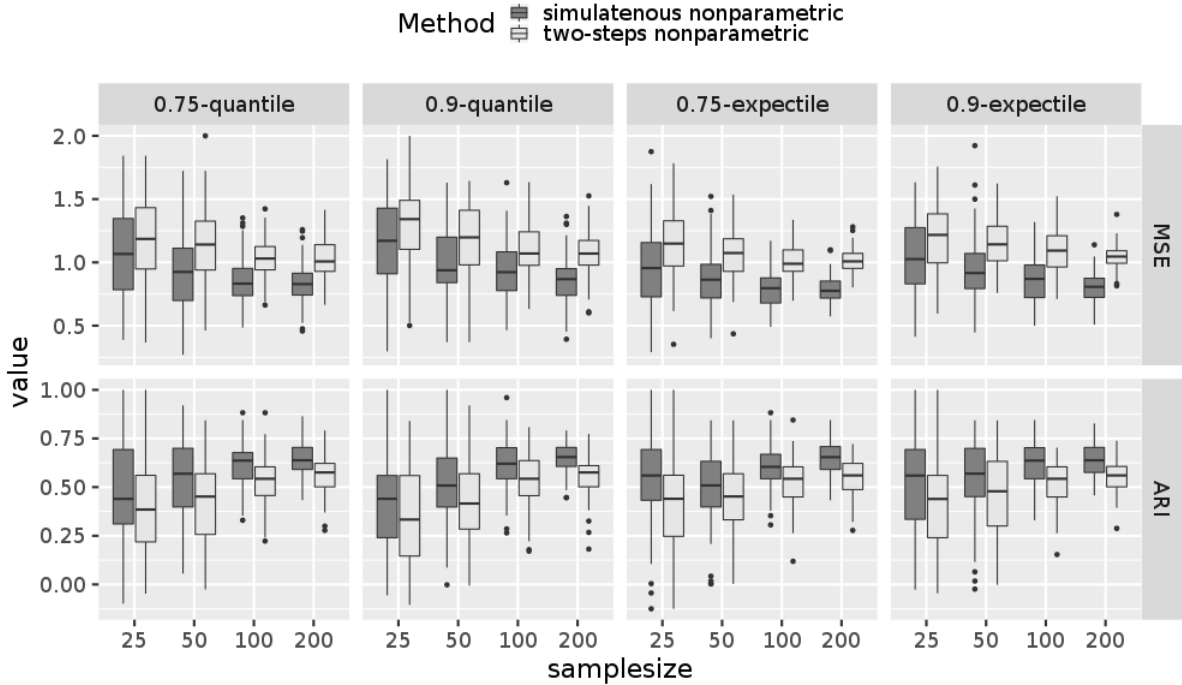


Figure 3: Boxplots of the MSE of the estimators of the regression parameters and ARI obtained by asymmetric regressions estimated with the simultaneous and two-step semi-parametric methods on 100 samples of size n .

5 High blood pressure prevention data set

Problem reminder We come back now to the problem of high blood pressure prevention. We focus on the detection of indicators related to the diastolic blood pressure. The indicators we want to consider are the 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.

Material and methods The data were obtained from National Health and Nutrition Examination Survey of 2011-2012¹. The target variable is the *diastolic blood pressure* in mmHg (code BPXDI1). The seven covariates in 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 values for those variables were removed. Seven variables are used in X to evaluate the level of physical activity. Among these variables, five variables are binary and indicated 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

¹The data are freely downloadable at

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

variables in 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 B. We consider 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 4 in Appendix B), 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*), regression model of Y on U (*regquadU*) with a selection of variables according to AIC (two variables are removed by the criterion: *alcohol* and *smoke*), regression model of Y on (U^\top, X^\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 MSE of prediction obtained on the testing samples are 122.55, 122.72 and 122.81 for *regquadUZ-K3*, *regquadUX* and *regquadU* respectively. Thus, the approach permits to summarize the information about the physical activity and slightly improve the prediction accuracy. Note that a Shapiro-Wilk’s normality test done on the residuals of *regquadUZ-K3* has a pvalue less than 10^{-4} for both of the learning and 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 similar partition since the adjusted rand index between all the couples of partition is more than 0.85. The regression parameters are presented in Table 1 of Appendix B. The signs of the coefficients are the same for the three losses. It appears that being a woman prevents the high blood pressure while age, alcohol consumption, overweight, lack of sleeping and cholesterol increase high blood pressure. One can be surprised that results claim that smoking limits the risk of high blood pressure, but this effect has already been brought out in Omvik (1996); Li et al. (2017). Note that the robust methods detect a more significant effect of the alcohol, smoking and physical activity to high blood pressure. Moreover, they slightly improve the prediction accuracy because the MSE obtained on the testing sample is 122.44 and 122.48 for the median and the logcosh losses respectively.

We now interpret the clustering results provided by the median loss. Class 1 ($\pi_1 = 0.15$) 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 by watching screens. Class 2 ($\pi_2 = 0.44$) groups the subjects having few physical activities. Class 3 ($\pi_3 = 0.44$) groups the subjects having high physical activities at work. These results show that having moderate physical activities (recreational activities, traveling by bike or foot, not spending many hours by watching screens) prevents the 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 fitting 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. Like 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 into account the regression 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 (X^\top, Z^\top) is independent of U . The crucial assumption of the model is the conditional independence of Y and X given (Z^\top, U^\top) .

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 X could be considered. This extension is straightforward if the different sub-groups of variables of X are known. However, the case of 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.

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 non-parametric 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 Vandekerckhove, 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.

A Technical details

A.1 Proofs of the Lemmas

Proof of Lemma 1. From (4), noting that $Y \perp X \mid Z$, we have for all $k = 1, \dots, K$

$$\begin{aligned} & \mathbb{E}[\rho(Y - U^\top \gamma - \delta_k) | U = u, Z_k = 1] = 0 \\ \Leftrightarrow & \forall x, \int \rho(Y - u^\top \gamma - \delta_k) p(y | Z_k = 1, u, x) dy = 0. \end{aligned}$$

Thus, noting that $Z \perp U \mid X$ and that $f_k(x) > 0$ implies $\mathbb{P}(Z_k = 1 | x, u) > 0$, we have, that for $k = 1, \dots, K$

$$\begin{aligned} & \mathbb{E}[\rho(Y - U^\top \gamma - \delta_k) | U = u, Z_k = 1] = 0 \\ \Leftrightarrow & \forall x, \mathbb{P}(Z_k = 1 | x, u) \int \rho(Y - u^\top \gamma - \delta_k) p(y | Z_k = 1, u, x) dy = 0 \\ \Leftrightarrow & \forall x, \int \rho(Y - u^\top \gamma - \delta_k) \mathbb{P}(Z_k = 1 | y, x, u) p(y | u, x) dy = 0 \end{aligned}$$

Proof is completed by noting that $\mathbb{P}(Z_k = 1 | y, x, u) = \mathbb{P}(Z_k = 1 | y, x) = r_k^{X,Y}(x, y)$ and that model identifiability involves that $\mathbb{E}[\rho(Y - U^\top \gamma - \delta_k) | U = u, Z_k = 1] = 0$ only for the true parameter β . \square

Proof of Lemma 2. Statement 1 holds because Y is discriminative (*i.e.*, its conditional distribution given Z_k is different to its unconditional distribution). Thus, the optimal classification error based on X is strictly more than the classification error based on (X^\top, Y) . Statement 2 is obtained by considering the quadratic loss and noting that $\mathbb{E}[Y | U = u, X = x, Z_k = 1] = u^\top \gamma + \delta_k$. Thus, we have $(\tilde{\gamma}, \tilde{\delta})^\top$ that statisfies

$$\begin{cases} \sum_{\ell=1}^K \Delta_{k\ell} \int g(u) (u^\top (\gamma - \tilde{\gamma}) + \delta_\ell - \tilde{\delta}_k) du & \forall k = 1, \dots, K \\ \sum_{k=1}^K \sum_{\ell=1}^K \Delta_{k\ell} \int g(u) u (u^\top (\gamma - \tilde{\gamma}) + \delta_\ell - \tilde{\delta}_k) du \end{cases}$$

where $g(u)$ is the density of U . Thus, we have

$$\begin{cases} \tilde{\delta}_k = \mathbb{E}[U]^\top (\gamma - \tilde{\gamma}) + \frac{\sum_{\ell=1}^K \Delta_{k\ell} \delta_\ell}{\sum_{v=1}^K \Delta_{kv}} & \forall k = 1, \dots, K \\ \text{Var}(U) (\gamma - \tilde{\gamma}) + \mathbb{E}[U] \sum_{k=1}^K \sum_{\ell=1}^K \Delta_{k\ell} (\delta_\ell - \tilde{\delta}_k) = 0. \end{cases}$$

Substituting, in the second line, $\tilde{\delta}_k$ by its definition, and noting that $\text{Var}(U)$ has only non-zero eigen values, we obtain

$$\begin{cases} \tilde{\delta}_k = \frac{\sum_{\ell=1}^K \Delta_{k\ell} \delta_\ell}{\sum_{v=1}^K \Delta_{kv}} & \forall k = 1, \dots, K \\ \tilde{\gamma} = \gamma. \end{cases}$$

□

A.2 EM algorithm for MLE estimation of the parametric simultaneous approach

Starting from the initial point $\theta^{[0]}$, this iterative algorithm alternates between the E-step and M-step. Its iteration $[r]$ is defined by

- **E-step** Computation of the expectation of the complete-data log-likelihood

$$t_{ik}(\theta^{[r-1]}) = \frac{\pi_k^{[r-1]} f_k(x_i; \alpha_k^{[r-1]}) f_\varepsilon(y_i - u_i^\top \gamma^{[r-1]} - \delta_k^{[r-1]})}{\sum_{\ell=1}^K \pi_\ell^{[r-1]} f_\ell(x_i; \alpha_\ell^{[r-1]}) f_\varepsilon(y_i - u_i^\top \gamma^{[r-1]} - \delta_\ell^{[r-1]})}.$$

- **M-step** Maximization of the expectation of the complete-data log-likelihood

$$\pi_k^{[r]} = \frac{1}{n} \sum_{i=1}^n t_{ik}(\theta^{[r-1]}), \quad \alpha_k^{[r]} = \arg \max_{\alpha_k} \sum_{i=1}^n t_{ik}(\theta^{[r-1]}) \ln f_k(x_i; \alpha_k),$$

$$\beta^{[r]} = \arg \max_{\beta} \sum_{i=1}^n \sum_{k=1}^K t_{ik}(\theta^{[r-1]}) \ln f_\varepsilon(y_i - u_i^\top \gamma - \delta_k).$$

The two maximizations at the M-step are standard. Indeed, the maximization problem provided by the computation of $\alpha_k^{[r]}$ already appears when the clustering is only made on \mathbb{X} . Moreover, the maximization problem provided by the computation of $\beta^{[r]}$ is the solution of the regression problem with weighted observations.

B Details about the high blood pressure prevention data set

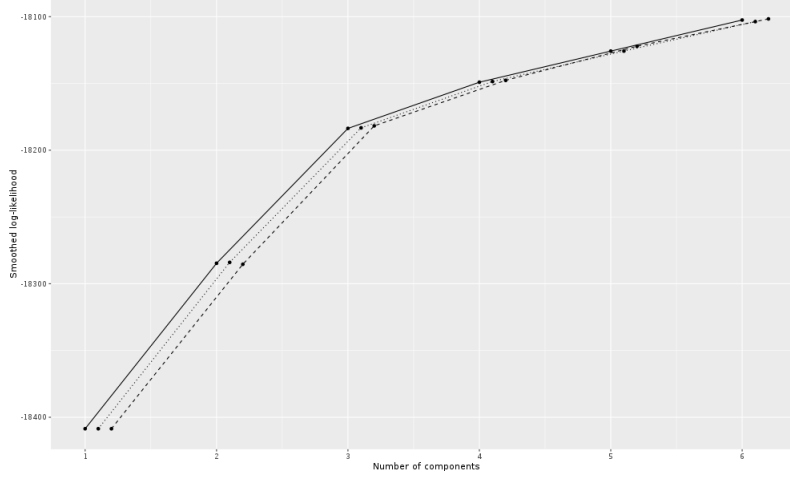


Figure 4: Value of the smoothed log-likelihood function obtained for different number of classes by the quadratic loss (in plain line), the median loss (in dotted line) and the logcosh loss (in dashed line). Note that the x-axis has been drift for the median and logcosh losses for ease of reading.

	<i>regquadUZ-K3</i>	<i>regmedUZ-K3</i>	<i>reglogchUZ-K3</i>
Gender (woman)	-3.07	-3.22	-3.13
Age	0.24	0.23	0.23
Alcohol (no)	-1.47	-2.48	-2.48
Obesity (yes)	3.00	2.72	2.72
Sleep	-0.48	-0.39	-0.31
Smoke (no)	0.09	0.42	0.51
Cholesterol	0.03	0.03	0.03
Class 1	60.28	59.04	58.58
Class 2	60.52	59.75	59.48
Class 3	61.09	60.14	59.62

Table 1: Estimators of the regression parameters given by the proposed semi-parametric approach with three losses.

	yes	no
class 1	0.03	0.97
class 2	0.10	0.90
class 3	0.99	0.01

Table 2: Estimated probabilities per classes for the levels of the variable PAQ605 (vigorous work activity)

	yes	no
class 1	0.29	0.71
class 2	0.23	0.77
class 3	0.94	0.06

Table 3: Estimated probabilities per classes for the levels of the variable PAQ620 (moderate work activity)

	yes	no
class 1	0.45	0.55
class 2	0.29	0.71
class 3	0.35	0.65

Table 4: Estimated probabilities per classes for the levels of the variable PAQ635 (usal travels on foot or by bike)

	yes	no
class 1	0.57	0.43
class 2	0.07	0.93
class 3	0.42	0.58

Table 5: Estimated probabilities per classes for the levels of the variable PAQ650 (vigorous recreational activities)

	yes	no
class 1	0.67	0.33
class 2	0.23	0.77
class 3	0.52	0.48

Table 6: Estimated probabilities per classes for the levels of the variable PAQ665 (moderate recreational activities)

	never	less than 1h	1h	2h	3h	4h	5h and more
class 1	0.02	0.15	0.24	0.25	0.18	0.08	0.07
class 2	0.02	0.10	0.13	0.26	0.17	0.12	0.20
class 3	0.01	0.14	0.22	0.33	0.13	0.10	0.07

Table 7: Estimated probabilities per classes for the levels of the variable PAQ710 (time spent watching TV)

	never	less than 1h	1h	2h	3h	4h	5h and more
class 1	0.08	0.22	0.27	0.20	0.10	0.05	0.08
class 2	0.30	0.24	0.13	0.12	0.07	0.05	0.09
class 3	0.21	0.31	0.21	0.14	0.04	0.03	0.06

Table 8: Estimated probabilities per classes for the levels of the variable PAQ715 (time spent using a computer)