Simultaneous Semiparametric Estimation of Clustering and Regression

Matthieu Marbac\textsuperscript{a}, Mohammed Sedki\textsuperscript{b}, Christophe Biernacki\textsuperscript{c}, and Vincent Vandewalle\textsuperscript{d}\textsuperscript{*}

\textsuperscript{a}Univ. Rennes, Ensaï, CNRS, CREST - UMR 9194, Rennes, France; \textsuperscript{b}Univ. Paris-Saclay and Inserm, France; \textsuperscript{c}Inria, Univ. Lille, CNRS, UMR - Laboratoire Paul Painlevé, Lille; \textsuperscript{d}Inria, Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, Lille, France

\textbf{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 semiparametric 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 \texttt{R} package \texttt{ClusPred} available on CRAN. Supplementary materials containing the technical details and the \texttt{R} codes are available online.

\textbf{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; Ehlm et al. 2016; Daouia, Girard, and Stupfler 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, Paindaveine, and Saracco 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, Wang, and Zhu 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, Wall, and Amemiya 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 condi-
tional distribution of particular variable given other ones, and it
is limited to a parametric framework. Another related reference
is the work of Sammel, Ryan, and Legler (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 refer-
cences can be found in the field of concomitant variables (Dayton
and Macready 1988; Grün and Leisch 2008; Vaňková 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 mix-
ture 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 clus-
tering and the regression models in a semiparametric frame-
work (Hunter, Richards, and Rosenberger 2011) thus circum-
venting the limits of the standard procedure (biased estimators).
We demonstrate that this procedure improves both the estima-
tors of the partition and regression parameters. A full paramet-
ric 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 2020), 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, Wang, and
Hettmansperger 2007; Butucea and Vandekerkhove 2014) could
also be considered in a similar way. Semiparametric inference is
achieved by a maximum smoothed likelihood approach (Levine,
Hunter, and Chauveau 2011) via a maximization–minimization
(MM) algorithm (Hunter and Lange 2004). Note that selecting
the number of components in a semiparametric mixture is not
easy (Kasahara and Shimotsu 2014; Kwon and Mbakop 2021).
However, in our context, the number of components can be
selected according to the quality of the prediction of the target
variable.

This article is organized as follows. Section 2 introduces a
general context where a statistical analysis requires both meth-
ods 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 estima-
tion of the clustering and of the regression parameters gen-
erally outperforms the two-step approach. This section also
briefly presents the simultaneous procedure on a parametric
framework, then focuses on the semiparametric frameworks.
Section 4 presents numerical experiments on simulated data
showing the benefits of the proposed approach. Section 5 illus-
trates our proposition for problems associated with high blood
pressure prevention. Section 6 provides a conclusion and discus-
sion about extensions. The mathematical details are presented in
Appendix A.

2. Embedding Clustering and Prediction Models

2.1. Data Presentation

Let \( (V^T, X^T, Y)^T \) be the set of the random variables where
\( V = (U^T, Z^T)^T \) is a \( d_U + K \) dimensional vector used
as covariate for the prediction of the univariate variable \( Y \in \mathbb{R},
\) \( X \) is a \( d_X \)-dimensional vector and \( Z = (Z_1, \ldots, Z_k)^T \in \mathbb{Z}
\) is a categorical variable with \( K \) levels. The variable \( Z \) indi-
cates the group membership such that \( Z_k = 1 \) if the subject
belongs to cluster \( k \) and otherwise \( Z_k = 0 \). The realizations of
\( (U^T, X^T, Y)^T \) are observed but the realizations of \( Z \) are unob-
served. 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 behaviors. The observed
data are \( n \) independent copies of \( (U^T, X^T, Y)^T \) denoted by \( u_1,
\ldots, u_n \), \( x_1, \ldots, x_n \), and \( y_1, \ldots, y_n \) respectively. The \( n \) unob-
served realizations of \( Z \) are denoted by \( z = (z_1, \ldots, z_n)^T \).

2.2. Motivating Example

We use the following example throughout the article, which
examines the general objective of high blood pressure preven-
tion. Here, we focus on the detection of indicators related to the
diastolic blood pressure (\( Y \)); see Berney, Burnier, and Wuerzner
(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 pres-
sure 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

2.3.1. Regression Model

Let a loss function be \( L(\cdot) \) 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

\[
Y = V^T \beta + \epsilon \quad \text{with} \quad \mathbb{E}[\rho(\epsilon)|V] = 0, \tag{1}
\]

where \( \beta = (\gamma^T, \delta^T)^T \in \mathbb{R}^{d_V}, \gamma \in \mathbb{R}^{d_U} \) are the coefficients
of \( U, \delta = (\delta_1, \delta_k)^T \in \mathbb{R}^Z \) are the coefficients of \( Z \) (i.e.,
the parameters of the group effect), and \( \epsilon \) 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 \( L(t) = t^2 \) and \( \rho(t) = 2t \)), the \( \tau \)-quantile regression (with \( L(t) = |t| + (2\tau - 1)t \) and \( \rho(t) = \tau - 1 \) for \( \tau \in [0, 1] \); Koenker and Bassett 1978), the \( \tau \)-expetile regression (with \( L(t) = |t - 1| |t \leq 0| t^2 \) and \( \rho(t) = 2t((1 - \tau) 1\{t \leq 0\} + \tau 1\{t > 0\} \); Newey and Powell 1987)), and so on.

The restriction on the conditional moment of \( \rho(e) \) given \( V \) 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 \( \tau \)-quantile (or \( \tau \)-expetile) 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 \( \varepsilon \) by \( f_\varepsilon \).

### 2.3.2. 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

\[
  f(x; \psi) = \sum_{k=1}^{K} \pi_k f_k(x),
\]

where \( \psi = \pi \cup \{f_1, \ldots, f_K\}, \pi = (\pi_1, \ldots, \pi_K)^T \) 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 \( \alpha_k \) are the parameters of component \( k \). In a semiparametric approach, some assumptions are required to ensure model identifiability (see, for instance, Chauveau et al. 2015). In the following, the semiparametric approaches are considered with the assumption that each \( f_k \) is a product of univariate densities (see Section 3.3).

### 2.3.3. 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 \( (X^T, Z^T)^T \) are independent. Moreover, the conditional distribution of \( W = (X^T, Y^T)^T \) given \( U \) is also a mixture model defined by the density

\[
  f(w; u; \theta) = \sum_{k=1}^{K} \pi_k f_k(x) f_\varepsilon(y_i - u_i^T \gamma - \delta_k),
\]

where \( \theta = \pi \cup \{\phi_1, \ldots, \phi_K\} \cup \varsigma, \phi_k \) grouping the parameters specific to component \( k \) (i.e., the finite parameter \( \delta_k \)) and the infinite parameters \( f_k \) and \( \varsigma \) grouping the parameters shared among the components (i.e., the infinite parameter \( \gamma \) and the infinite parameter \( f_\varepsilon \)). We have

\[
  E[L(\theta) | V] = 0.
\]

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 \( U \) has full rank and finally that \( f_{\delta_j} \) and \( f_\varepsilon \) are strictly positive. Denoting \( \beta_0 \) as the single parameter satisfying (4) and

\[
  \tau_k^U X Y (u, x, y) = \sum_{\ell=1}^{K} \pi_\ell f_\ell(x)(y - u^T \gamma - \delta_\ell),
\]

we have

\[
  \forall k = 1, \ldots, K, \ E[\tau_k^U X Y (U, X, Y) | \rho(Y - U^T \gamma - \delta_k)|U, X] = 0 \iff \beta = \beta_0.
\]

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

1. **Clustering step.** Perform a clustering of \( U \) to obtain an estimated hard classification rule \( \hat{\pi}^X : \mathbb{R}^d_x \to \mathbb{Z} \) or an estimated fuzzy classification rule \( \tau^X : \mathbb{R}^d_x \to \mathbb{Z}_K \) where \( \mathbb{Z}_K \) is the simplex of size \( K \).

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

   \[
   \hat{\beta}^X = \arg\min_{\beta} \sum_{i=1}^{n} \sum_{k=1}^{K} I_k(X_i) \mathcal{L}(y_i - u_i^T \gamma - \delta_k),
   \]

where \( I_k(X_i) \) is the element \( k \) of vector \( \tau^X(X_i) \). Note that \( I_k(X_i) \) 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 is suboptimal. Indeed, even if the optimal classification rule on \( X \) 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 \( f_\varepsilon \) are continuous and strictly positive where there exists \((k, \ell)\) such \( f_k \) and \( f_\ell \) have no disjoint support and also \( \delta_k \neq \delta_\ell \), and finally where \( f_\varepsilon \) is not constant. Suppose that \( f_\varepsilon \) defines a random variable with finite variance and that \( U \) has a full-rank covariance matrix. Then,
1. Any hard classification rule \( \hat{r}^{X} : \mathbb{R}^{d_{X}} \rightarrow \mathcal{Z} \) is suboptimal in the sense that
   \[
   \mathbb{E} \left[ \sum_{k=1}^{K} \hat{r}^{X}(X)Z_{k} \right] < \mathbb{E} \left[ \sum_{k=1}^{K} U^{X,Y}(U, X, Y)Z_{k} \right].
   \]

2. Consider the quadratic loss, the best classification rule \( r^{X} \) computed on \( X \) and its associated estimator of the regression parameters \( \hat{\beta}^{r^{X}} \). The estimator \( \hat{r}^{X} \) is asymptotically unbiased but the estimator \( \hat{\beta}^{r^{X}} = \Delta_k \) is asymptotically biased with an asymptotic bias equals to \( \frac{\sum_{k=1}^{K} \Delta_k \delta_k}{\sum_{k=1}^{K} \delta_k} \), where \( \Delta_k \) is defined.

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, that is, \( \hat{r}^{X}(x_i) = z_i \), 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 \( \alpha_k \) and thus we have \( \phi_k = (\alpha_k, x_i) \). Moreover, the distribution of the noise \( f_e \) 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 \( \gamma = \gamma \). The aim of the simultaneous procedure can be achieved by maximizing the log-likelihood of \( x, y \) given \( u \) with respect to \( \theta \)

\[
\ell(\theta; x, y | u) = \sum_{i=1}^{n} \ln \left( \sum_{k=1}^{K} \pi_k f_k(x_i; \alpha_k) f_e(y_i - u_i^\top \gamma - \delta_k) \right).
\]

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

\[
\ell(\theta; x, y, z | u) = \sum_{i=1}^{n} \sum_{k=1}^{K} \ln \left( \pi_k f_k(x_i; \alpha_k) f_e(y_i - u_i^\top \gamma - \delta_k) \right).
\]

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 semiparametric 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 Semiparametric Procedure

##### 3.3.1. Semiparametric Model

In this section, we consider the semiparametric version of the model defined by (3) where the densities of the components are assumed to be a product of univariate densities (i.e., \( f_k(x_i) = \prod_{j=1}^{d} f_j(x_{ij}) \)). Therefore, the parameters specific to component \( k \), denoted by \( \phi_k \), are \( \delta_k \) and \( f_{k1}, \ldots, f_{kd} \). We have

\[
f_k(w_i | u_i; \varphi, \gamma) = \prod_{j=1}^{d} f_j(x_{ij}) f_e(y_i - u_i^\top \gamma - \delta_k).
\]

A sufficient condition implying model identifiability is that the covariance matrix of \( U \) has full rank and that the marginal distribution of \( X \) is identifiable and thus a sufficient condition is to consider linearly independent densities \( f_j \)'s and \( d_X \geq 3 \) (Allman et al. 2009). Thus, if \( d_X \) is less than three, other semiparametric mixture models should be considered to achieve clustering (i.e., location-scale models; see Hunter, Wang, and Hettmansperger 2007; Chauveau et al. 2015).

##### 3.3.2. Smoothed Log-Likelihood

Let \( S \) be the smoothing operator defined by

\[
Sf_k(w | u; \phi_k, \gamma) = \int K_h(w - \tilde{w})f_k(\tilde{w} | u; \phi_k, \gamma) d\tilde{w},
\]

where \( K_h(a) = \prod_{i=1}^{d} K(h^{-1}a_i) \) with \( a \in \mathbb{R}^d \) and with \( K_h(a) \) is a rescale kernel function defined by \( K_h(a_i) = h^{-1}K(h^{-1}a_i) \) where \( h \) is the bandwidth. The estimation is achieved by maximizing the smoothed log-likelihood (Levine, Hunter, and Chauveau 2011) defined by

\[
\ell(\theta) = \sum_{i=1}^{n} \ln \left( \sum_{k=1}^{K} \pi_k (N\tilde{f}_k)(w_i | u_i; \phi_k, \gamma) \right)
\]

where \( N\tilde{f}_k \) (\( w_i; \phi_k, \gamma \)) = \( \exp \{ S \ln f_k(w | u; \theta_i) \} = \exp \{ \int K_h(w - \tilde{w})f_k(\tilde{w} | u; \theta_i) d\tilde{w} \} \), subject to the empirical counterpart of (5):

\[
\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{f_k(w_i | u_i; \phi_k, \gamma)}{\pi_k f_k(w_i | u_i; \phi_i, \gamma)} \rho(y_i - u_i^\top \gamma - \delta_k) = 0.
\]

##### 3.3.3. 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:
\[ f^{[r−1]}_{ik} = \frac{\pi^{[r−1]}_{ik} (N f_{ik} (w_i \mid u_i, \phi^{[r−1]}_{ik}, \xi^{[r−1]}), \sum_{l=1}^{K} \pi^{[r−1]}_{il} (N f_{il} (w_i \mid u_i, \beta^{[r−1]}_{il}, \xi^{[r−1]}))}{\pi^{[r−1]}_{ik} (N f_{ik} (w_i \mid u_i, \phi^{[r−1]}_{ik}, \xi^{[r−1]}), \sum_{l=1}^{K} \pi^{[r−1]}_{il} (N f_{il} (w_i \mid u_i, \beta^{[r−1]}_{il}, \xi^{[r−1]}))} \]

• Minimization step:
\[
\pi^{[r]}_{k} = \frac{1}{n} \sum_{i} f^{[r−1]}_{ik}, \\
\beta^{[r]} = \arg\min_{\beta} \sum_{i,k} \ell^{[r−1]}_{ik} \mathcal{L}(y_i - u_i \Gamma^{[r]} - \delta_k), \\
\phi^{[r]}_{k} = \frac{1}{n} \sum_{i,k} \ell^{[r−1]}_{ik} K_h(x_{ij} - a) \\
\phi^{[r]}_{k} = \frac{1}{n} \sum_{i,k} \ell^{[r−1]}_{ik} K_h(y_i - u_i \Gamma^{[r]} - \delta_k - a),
\]

then set \( \phi^{[r]}_{k} = \gamma^{[r]}_{k} + \phi^{[r]}_{k} \) and \( \xi^{[r]}_{k} = \delta^{[r]}_{k} \).

The Majorization–Minimization algorithm is monotonic for the smoothed log-likelihood. It is a direct consequence of the monotonicity of the algorithm of Levine, Hunter, and Chauveau (2011) where we use the fact that, in order to satisfy the moment condition defined in (5) of Lemma 1, we must have \( \beta^{[r]} = \arg\min_{\beta} \sum_{i,k} \ell^{[r−1]}_{ik} \mathcal{L}(y_i - u_i \Gamma^{[r]} - \delta_k) \).

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 semiparametric 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 \( U_i \sim \mathcal{N}(0, I_2) \) and such that \( (X_i, Y_i) \) given \( U_i \) 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 \( X_i \) given \( Z_i \) is a product of univariate densities such that \( X_{ikj} = \xi Z_{ikj} + \eta_{ij} \) where \( \kappa_j = (\kappa_{j1}, \ldots, \kappa_{jK}) \), \( \kappa_{jk} = 1 \) if \( k = (j \mod K) + 1 \) and \( \kappa_{jk} = 0 \) otherwise. Finally, we have \( Y_i = U_i \Gamma^{[r]} + Z_i \delta^{[r]} + \epsilon_i \) with \( \Gamma^{[r]} = (1, 1) \) and \( \delta^{[r]} = 2\xi k \). \( \eta_{ij} \) and \( \epsilon_i \) are independently drawn from a standard Gaussian distribution or a Student distribution with 3 degrees of freedom. The parameter \( \xi \) is tuned according to the distributions \( \eta_{ij} \) and \( \epsilon_i \) 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 \( \beta \) and the Adjusted Rand Index (ARI) between the true and the estimated partition on 100 replicates. The semiparametric 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 semiparametric frameworks, where the parametric approaches assume that \( \eta_{ij} \) and \( \epsilon_i \) are Gaussian. We consider four scenarios: \( \eta_{ij} \sim \mathcal{N}(0, 1) \) for the first two scenarios and \( \eta_{ij} \sim T(3) \) for the last two scenarios, and \( \epsilon_i \sim \mathcal{N}(0, 1) \) for the scenarios 1 and 3 and \( \epsilon_i \sim T(3) \) for Scenarios 2 and 4. Figure 1 presents the results obtained

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 \( d = 6 \).
when \( K = 3 \) and \( d = 6 \). When the parametric model is well-specified (Scenario 1), results are equivalent to those obtained by the semiparametric 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 semiparametric model if the family of the distributions is unknown to prevent the bias in the estimation.

### 4.3. Robust Regression

When the noise of a regression follows an heavy-tailed distribution, robust regressions allow the estimators of the regression coefficients to be improved compared to the ordinary least-squares 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 semiparametric framework, and that the resulting estimators are better than those obtained with the quadratic loss. In this experiment, we consider Scenario 4 (i.e., \( \eta_{ij} \) and \( \epsilon_i \) both follow independent \( T(3) \)) 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 semiparametric simultaneous method allows these regression models to be easily managed, data are generated with \( K = 2 \) and \( d = 4 \) such that

\[
\eta_{ij} \sim \mathcal{N}(0,1) \quad \text{and} \quad \epsilon_i \sim \mathcal{N}(-c_\tau,1).
\]

The scalar \( c_\tau \) is defined according to the regression model. Thus, \( c_\tau \) 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 semiparametric approach improves the estimators compared to those provided by the two-step approach.

### 5. High Blood Pressure Prevention Dataset

#### 5.1. 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 the similar physical activities and we want to use these patterns in the prediction of the diastolic blood pressure.

![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.](image)
5.2. 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 \(U\) are gender which was equal to 1 for men and 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 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 \(X\) 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 \(h = n^{-1/5}\).

5.3. 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 semiparametric 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 semiparametric framework (regquadUZ-K3), a 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), a regression model of \(Y\) on \(\left(U^T, X^T\right)^T\) (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 p-value less than $10^{-5}$ for the learning sample and 0.003 for the testing sample. Thus, the semiparametric 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 semiparametric 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 semiparametric 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) and 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 article, 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 semiparametric framework. We advise using the semiparametric 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 semiparametric 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 semiparametric mixtures. Moreover, this could allow for a variable selection in clustering while this approach is only used in a parametric framework (Tadesse, Sha, and Vannucci 2005; Raftery and Dean 2006). The semiparametric 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, Wang, and Hettmansperger 2007) or by incorporating an independent component analysis structure (Zhu and Hunter 2019). Moreover, we can easily relax the assumption that $(X^T, Z^T)$ is independent of $U$. The crucial assumption of the model is the conditional independence of $Y$ and $X$ given $Z$.

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 $X$ could be considered. This extension is straightforward if the different sub-groups of variables of $X$ 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.

### 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).

### ORCID

Christophe Biernacki [http://orcid.org/0000-0002-8995-0568]

Vincent Vandewalle [http://orcid.org/0000-0003-2946-9059]

### References

Allman, E. S., Matias, C., and Rhodes, J. A., (2009), “Identifiability of Parameters in Latent Structure Models With Many Observed Variables,” The Annals of Statistics, 37, 3099–3132. [478,480]

Ando, T., and Bai, J. (2016). “Panel Data Models With Grouped Factor Structure Under Unknown Group Membership,” Journal of Applied Econometrics, 31, 163–191. [477]

Auray, S., Klutchikoff, N., and Rouviere, I. (2015), “On Clustering Procedures and Nonparametric Mixture Estimation,” Electronic Journal of Statistics, 9, 266–297. [477]

Berney, M., Burnier, M., and Wuerzner, G. (2018), “Isolated Diastolic Hypertension: Do We Still Have to Care About It?” Revue Medicale Suisse, 14, 1607–1610. [478]

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. [477]

Butucea, C., and Vandekerkhove, P. (2014), “Semiparametric Mixtures of Symmetric Distributions,” Scandinavian Journal of Statistics, 41, 227–239. [478]

Carroll, R. J., and Wand, M. P. (1991), “Semiparametric Estimation in Logistic Measurement Error Models,” Journal of the Royal Statistical Society, Series B, 53, 573–585. [477]

Charlier, I., Paindaveine, D., and Saracco, J. (2015), “Conditional Quantile Estimation Through Optimal Quantization,” Journal of Statistical Planning and Inference, 156, 14 – 30. [477]

Chauveau, D., Hunter, D. R., Levine, M. (2015), “Semi-Parametric Estimation for Conditional Independence Multivariate Finite Mixture Models,” Statistics Surveys, 9, 1–31. [478,479,480,481]
