Simultaneous semi-parametric estimation of clustering and regression

Matthieu Marbac, Mohammed Sedki, Christophe Biernacki, Vincent Vandewalle

Simultaneous semi-parametric estimation of clustering and regression

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

[1] *Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France.*
matthieu.marbac-lourdelle2@ensai.fr
[2] *Univ. Paris-Sud and Inserm, France. mohammed.sedki@universite-paris-saclay.fr*
[3] *Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille.*
christophe.biernacki@inria.fr
[4] *Univ. Lille, CHU Lille, ULR 2694 - METRICS and Inria, F-59000 Lille, France.*
vincent.vandewalle@univ-lille.fr

Résumé. Nous étudions l'estimation des paramètres des modèles de régression avec des effets fixes par classe, lorsque la variable de classe est manquante alors que des variables liées à la classe sont disponibles. Ce problème peut être résolu en modélisant la distribution jointe de la variable cible et des variables liées à la classe. La stratégie habituelle d'estimation des paramètres pour ce modèle joint est une approche en deux étapes, commençant par l'apprentissage de la variable de la classe (étape de clustering) et ensuite l'insertion de son estimateur pour ajuster le modèle de régression (étape de régression). Toutefois, cette approche est sous-optimale car à la fois les estimateurs des paramètres de la régression sont biaisés et aussi elle n'utilise pas la variable cible pour le clustering. Ainsi, nous plaidons pour une approche d'estimation simultanée du clustering et de la régression, dans un cadre semi-paramétrique. Des expériences numériques illustrent les avantages de notre proposition en considérant différents modèles pour la distribution dans les classes et différents modèles de régression.

Mots-clés. clustering ; modèles de mélange ; modèles de régression ; modèles semi-paramétriques.

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 can be solved by modeling the joint distribution 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 since both regression estimates are biased and 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.

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

1 Introduction

The regression model with a fixed group effect 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.

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. 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. However, since the group variable is estimated with error (class overlap), it is well-known that the resulting estimators of the parameters of regression are biased (Bertrand et al., 2017). The bias depends on the accuracy of the clustering step. Note that, although 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 sub-optimal procedures.

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. We focus on semi-parametric mixture where the component densities are defined as a product of univariate densities (Chauveau et al., 2015), which is identifiable if the univariate densities are linearly independent and if at least three variables are used for clustering (Allman et al., 2009). 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).

The presentation 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. Section 4 discusses about numerical experiments, which are not given in this long summary but will be presented during the talk. More details about the work presented can be found in Marbac et al. (2020).

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 $Z = (Z_1, \ldots, Z_K)^\top \in \mathcal{Z}$ 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 $\mathbb{U} = (u_1, \ldots, u_n)^\top$, $\mathbb{X} = (x_1, \ldots, x_n)^\top$ and $\mathbb{Y} = (y_1, \ldots, y_n)^\top$ respectively. The n unobserved realizations of Z are denoted by $\mathbb{Z} = (z_1, \ldots, z_n)^\top$.

2.2 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

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

where $\beta = (\gamma^\top, \delta^\top)^\top \in \mathbb{R}^{d_V}$, $\gamma \in \mathbb{R}^{d_U}$ are the coefficients of U, $\delta = (\delta_1, \ldots, \delta_K)^\top \in \mathbb{R}^K$ 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 $\rho(\varepsilon) = \tau - \mathbb{1}_{\{\varepsilon \leq 0\}}$), the τ-expectile regression (with $\mathcal{L}(t) = |\tau - \mathbf{1}\{t \leq 0\}|t^2$ and $\rho(t) = 2t((1-\tau)\mathbf{1}\{t \leq 0\} + \tau \mathbf{1}\{t > 0\})$.

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. 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

$$f(x;\vartheta) = \sum_{k=1}^{K} \pi_k f_k(x) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d_X} f_{kj}(x_j), \tag{2}$$

where $\vartheta = \{\pi_k, f_k; k = 1, \ldots, K\}$, $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$ and where f_k is the density of component k defined as the product of univariate densities f_{kj} in the semi-parametric setting.

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 $\theta = \{\vartheta\} \cup \{\delta_k; k = 1, \ldots, K\} \cup \{\gamma, f_\varepsilon\}$)

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

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

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

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. It shows an equivalence between the moment equation which permits to understand why the two-step approach is biased and which justifies the use of the unified procedure.

Lemma 1. *Let an identifiable model defined by* (3) *and* (4)*, 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*

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

3 The proposed simultaneous estimation procedure

Based on Lemma 1, we have shown in Marbac et al. (2020) that performing a two-step approach, thus performing the regression based on the clustering step output, provides a suboptimal classification rule because the classification neglects the information given by Y. Consequently, we circumvent this issue by using a simultaneous semi-parametric approach, also avoiding bias of parametric miss-specified models.

Semi-parametric model In this section, we consider the semi-parametric version of the model defined by (3) where the densities of the components are assumed to be a product of univariate densities. Thus, we have

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

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 $f(y, x \mid u; \theta) = \sum_{k=1}^{K} \pi_k f_k(w \mid u; \theta)$.

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 steps. Thus, an iteration $[r]$ is defined by

- Majorization step: $t_{ik}^{[r-1]} \propto \pi_k^{[r-1]} \left(\mathcal{N} f_k^{[r-1]}\right)(w_i \mid u_i; \theta^{[r-1]})$, with $\mathcal{N} f_k$ the exponential of the smoothed log-density in class k (Levine et al., 2011).
- 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

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

     with K_h the rescale kernel function of bandwith h considered in the smoothing.

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

4 Numerical experiments and conclusion

In experiments presented in Marbac et al. (2020), we have considered several types of regressions in a semi-parametric framework. In a first time we have compared the simultaneous and the two-step approaches in a parametric and semi-parametric framework. In a second time we have shown that robust regressions can be easily used with the semi-parametric approach and improve the estimators of the regression parameters. In a third time, we have shown that the semi-parametric method permits to consider asymetric losses (quantile or expectile regressions). An application the high blood pressure has also been studied where we have considered simultaneously the clustering of subjects based on their physical activity and the use of this variable in a regression model on the diastolic blood pressure.

The main conclusion is that simultaneously performing the clustering and the estimation of the regression model improves the accuracy of both the partition and of the regression parameters. The approach can be applied to a wide range of regression models, and avoids bias in the estimation compared with parametric models.

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.

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.

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.

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.

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

Marbac, M., Sedki, M., Biernacki, C., and Vandewalle, V. (2020). Simultaneous semi-parametric estimation of clustering and regression. preprint, `https://hal.inria.fr/hal-03090573`.