

Adaptive mixtures of regressions: Improving predictive inference when population has changed

Charles Bouveyron^a, Julien Jacques^{b,c,d,*}

^aLaboratoire SAMM, EA4543, University Paris I Panthéon-Sorbonne, Paris, France

^bLaboratoire Paul Painlevé, UMR CNRS 8524, University Lille I, Lille, France

^cMODAL team, INRIA Lille-Nord Europe

^dPolytech'Lille

Abstract

When regression is carried out in a prediction purpose, one of the main assumptions is the absence of evolution in the modeled phenomenon between the training and the prediction stages. Unfortunately, this assumption turns out to be often false in practical situations. The present work investigates the estimation of regression mixtures when population has changed between the training and the prediction stages. The main idea of this work is to link the regression mixture of the prediction population with the known regression mixture of the training population. For this, two approaches are proposed. On the one hand, a parametric approach modelling the relationship between dependent variables of both populations is presented and the EM algorithm is used for parameter estimation. On the other hand, a Bayesian approach is also proposed in which the priors on the prediction population depend on the mixture regression parameters of the training population. In this latter case, a MCMC procedure is used for inference. Both approaches need nevertheless to observe at least some observations arising from the prediction population. The relevance of both the parametric and the Bayesian approaches is illustrated on simulations and then compared to classical strategies on an environmental dataset.

Keywords: Transfer learning, Mixture of regressions, Switching regression, EM algorithm, Bayesian inference, MCMC algorithm.

1. Introduction

The mixture of regressions, introduced by Goldfeld and Quandt (1973) as the switching regression model and also named clusterwise linear regression model in Hennig (1999), is a popular regression model for modelling complex system. In particular, the switching regression model is often used in Economics for modelling phenomena with different phases. This model

*Corresponding author. Tel.: +33 320 436 760, Fax: +33 320 434 302

Email addresses: charles.bouveyron@univ-paris1.fr (Charles Bouveyron), julien.jacques@polytech-lille.fr (Julien Jacques)

assumes that the dependent variable $Y \in \mathbb{R}$ can be linked to a covariate $x = (1, x_1, \dots, x_p) \in \mathbb{R}^{p+1}$ by one of K possible regression models:

$$Y = x^t \beta_k + \sigma_k \varepsilon, \quad k = 1, \dots, K \quad (1)$$

with prior probabilities π_1, \dots, π_K (with the classical constraint $\sum_{i=1}^K \pi_k = 1$), where $\varepsilon \sim \mathcal{N}(0, 1)$, $\beta_k = (\beta_{k0}, \dots, \beta_{kp}) \in \{\beta_1, \dots, \beta_K\}$ is the regression parameter vector in \mathbb{R}^{p+1} and $\sigma_k^2 \in \{\sigma_1^2, \dots, \sigma_K^2\}$ is the residual variance. The conditional density distribution of Y given x is therefore:

$$p(y|x) = \sum_{k=1}^K \pi_k \phi(y|x^t \beta_k, \sigma_k^2), \quad (2)$$

where $\phi(\cdot|x^t \beta_k, \sigma_k^2)$ is the univariate Gaussian density parametrized by its mean $x^t \beta_k$ and variance σ_k^2 . For such a model, the prediction of y for a new observed covariate x is usually carried out in two steps: first the component membership of the data is estimated by the maximum a posteriori (MAP) rule and then y is predicted using the selected regression model (see Section 2.2). Among the works which focused on this model, we can emphasize the following ones which have contributed to the popularity of this model: Hennig (2000) investigates the model identifiability, Hurn et al. (2003) proposes a Bayesian inference for the model estimation, Zhu and Zhang (2004) studies the asymptotic theory of parameter estimators in order to define hypothesis tests, and Khalili and Chen (2007) considers variable selection for this specific regression model. Let us also mention that Leisch (2004) develops a package for the R software devoted to the mixture of regressions.

The present paper focuses on the problem of using a mixture regression model for prediction when the modeled phenomenon has changed between the training stage, which has led to the parameter estimation, and the prediction stage. More precisely, we assume that model (1) has been estimated with a sample from a given training population (of size large enough to have an estimation of satisfying quality), and we want to use it to predict the dependent variable Y for a new population which could be different from the training one. For instance, the difference between both populations can be due to a switch in the covariate distribution or to a variation of the link between the covariates and the dependent variable. The goal is then to transfer the knowledge from the training (source) population to the prediction (target) population. This task is usually known as *transfer learning* (see Pan and Yang (2010) for a complete survey), and can be summarized by Figure 1 in the case of the regression mixture model.

We now give some application examples of transfer learning. In a biological context, Biernacki et al. (2002) and Jacques and Biernacki (2010) proposed models for clustering male and female birds: the source population consists of birds from a common species whereas the target population is composed of birds from a rarer species. Another application concerns the problem of sentiment classification as considered by Blitzer et al. (2007). As the review data can be very different among several type of products, there is a need to collect a large sample of labeled data for each product in order to train a specific review-classification model per product. The use of transfer learning techniques allows to adapt a sentiment classifier

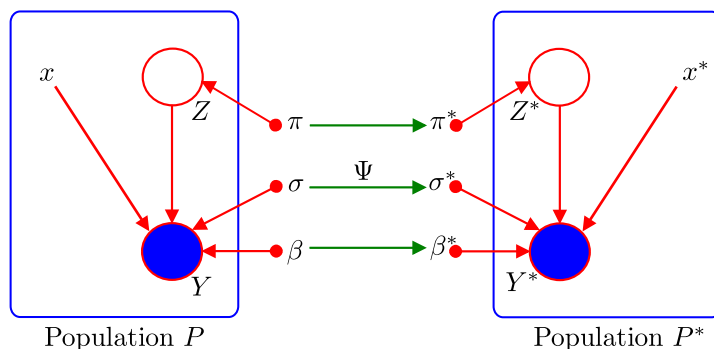


Figure 1: Learning process of transfer learning

from one product to another one. In Bouveyron and Jacques (2010), the authors predict house prices from house features for a city of the USA West Coast (San Jose, California) by adapting a regression model learned with data issued from another city stated on the East Coast (Birmingham, Alabama). The use of a transfer learning model allows to spare an expensive recollect of training data for the target population (San Jose housing in this application). Other examples can be found in Pan and Yang (2010).

1.1. Related work

Transfer learning is a particularly active research field since the NIPS-95 workshop “Learning to Learn”, in which the need for machine learning methods reusing previously learned knowledge was exhibited. Contrary to previously cited works in the classification context (Biernacki et al. (2002); Jacques and Biernacki (2010)), in which the data of the target population can be unlabelled, the regression purpose need to observe at least some couples (y_i, x_i) in the target population. In this case, we speak of *inductive* transfer learning. Readers interested in a comprehensive review can refer to Pan and Yang (2010).

Most of the methods allowing to treat such setting are especially designed for estimating simultaneously the parameters of both source and target populations (we speak of *multi-task learning*), but can easily be adapted for transfer learning. They consider either a Bayesian or a regularization framework. Typically, in the Bayesian approach, each task is assume to share the same prior (see Lawrence and Platt (2004) for instance). In the regularization framework, parameters between models for source and target population are assumed to be linked (see Evgeniou and Pontil (2004) in a SVM context for instance).

In the regression context, *Covariate Shift* is a specific transfer learning problem considering that the probability density of the covariates in the target population is different from the one of the source population. However, the relationship between covariates and dependent variable is assumed not to have changed (Shimodaira (2000); Storkey and Sugiyama (2007); Sugiyama (2006); Sugiyama and Müller (2005, 2007)). Thus, if the regression model is exactly known, a change in the probability distribution of the explanatory variables is not

a problem. Unfortunately, this is never the case in practice and the regression model estimated with the training data could be very disappointing when applied to data with a different probability distribution.

The originality of our work consists in introducing parametric models allowing to link the source and target populations. A more conventional Bayesian approach is also investigated, and comparison of both approaches are carried out on simulation and real data.

1.2. Problem formulation

Assuming that the target population P^* , for which we want to predict Y , is different from the source population P , the mixture regression model for P^* can be written as follows:

$$Y^* = x^{*t} \beta_k^* + \sigma_k^* \varepsilon^* \quad (3)$$

$$p(y^*|x^*) = \sum_{k=1}^{K^*} \pi_k^* \phi(y^*|x^{*t} \beta_k^*, \sigma_k^{*2})$$

with $\varepsilon^* \sim \mathcal{N}(0, 1)$, $\beta_k^* \in \{\beta_1^*, \dots, \beta_{K^*}^*\}$ and $\sigma_k^* \in \{\sigma_1^*, \dots, \sigma_{K^*}^*\}$. Let us now precise the focus of this paper by making the three following assumptions:

H_1 : the couples of variables (Y, x) and (Y^*, x^*) are assumed to be the same but measured on two different populations.

H_2 : the size n^* of the observation sample $S^* = (y_i^*, x_i^*)_{i=1, n^*}$ of population P^* is assumed to be small compared to the number of observations of the source population P . Otherwise, the mixture regression model could be estimated directly without using the source population.

H_3 : as both populations have the same nature, each mixture is assumed to have the same number of components ($K^* = K$).

Under these assumptions, the goal is then to predict Y^* for some new x^* by using both samples $S = (y_i, x_i)_{i=1, n}$ and S^* . The challenge consists therefore in exhibiting a link between both populations.

1.3. Organization of the manuscript

The reminder of this work is organised as follows. Section 2 proposes a first solution to improve the predictive inference on the target population by defining parametric models for the link between mixture regression models of both populations. This approach has the advantage to lead to interpretable results, which should help the practitioner in analyzing the differences between the source and target populations. An alternative Bayesian approach, most frequent in transfer learning, is presented in Section 3. The link between regression models is then formulated through prior densities on the target population. The advantage of this strategy is its flexibility which can fit into different situations, if the prior densities are well chosen. In Section 4, the performance of both the parametric and the Bayesian approaches

is first illustrated on simulations. Then, the proposed strategies are compared to classical methods on an environmental dataset. Section 5 finally proposes some concluding remarks and directions for future works.

2. Parametric approach for adaptive mixture of regressions

This section presents a parametric approach which consists in modelling the link between training and test populations by a parametric relationship between the regression parameters.

2.1. Parametric models for linking the reference and test populations

Let us introduce a latent variable $Z^* \in \{0, 1\}^K$ representing the belonging of observations to the K mixture components, *i.e.* $z_{ik}^* = 1$ indicates that the i -th observation (x_i^*, y_i^*) comes from the k -th component and $z_{ik}^* = 0$ otherwise. Conditionally to an observation x of the covariates, we would like to exhibit a distributional relationship between the dependent variables of the same mixture component such that $Y_{|x, z_{ik}^*=1}^*$ and $\psi_k(Y_{|x, z_{ik}^*=1})$ have the same probability distribution, with ψ_k a function from \mathbb{R} to \mathbb{R} .

Let β_k and β_k^* ($1 \leq k \leq K$) be respectively the parameters of the mixture regression models in the source and the target populations (Equations (1) and (3)). We assume in this section that the function ψ_k , exhibiting the link between the source and target populations, is such that:

$$\begin{aligned} \beta_k^* &= \Lambda_k \beta_k, \text{ where } \Lambda_k = \text{diag}(\lambda_{k0}, \lambda_{k1}, \dots, \lambda_{kp}) \\ \sigma_k^* &\text{ is free,} \end{aligned} \quad (4)$$

where $\text{diag}(\lambda_{k0}, \lambda_{k1}, \dots, \lambda_{kp})$ is the diagonal matrix containing $(\lambda_{k0}, \lambda_{k1}, \dots, \lambda_{kp})$ on its diagonal. The interest of introducing such a link lies in the reduction of the number of parameters to estimate for the mixture regression model for P^* . In the following, we go further by introducing some constraints on Λ_k and σ_k^* in order to define a family of parsimonious models, which includes many of the situations that may be encountered in practice:

- M_1 assumes both populations are the same: $\Lambda_k = I_d$ is the identity matrix ($\sigma_k^* = \sigma_k$),
- M_2 models assume the link between both populations is covariate and mixture component independent:
 - M_{2a} : $\lambda_{k0} = 1$, $\lambda_{kj} = \lambda$ and $\sigma_k^* = \lambda \sigma_k \quad \forall 1 \leq j \leq p$,
 - M_{2b} : $\lambda_{k0} = \lambda$, $\lambda_{kj} = 1$ and $\sigma_k^* = \sigma_k \quad \forall 1 \leq j \leq p$,
 - M_{2c} : $\Lambda_k = \lambda I_d$ and $\sigma_k^* = \lambda \sigma_k$,
 - M_{2d} : $\lambda_{k0} = \lambda_0$, $\lambda_{kj} = \lambda_1$ and $\sigma_k^* = \lambda_1 \sigma_k \quad \forall 1 \leq j \leq p$,
- M_3 models assume the link between both populations is covariate independent:
 - M_{3a} : $\lambda_{k0} = 1$, $\lambda_{kj} = \lambda_k$ and $\sigma_k^* = \lambda_k \sigma_k \quad \forall 1 \leq j \leq p$,

Model	M_1	M_{2a}	M_{2b}	M_{2c}	M_{2d}	M_{3a}	M_{3b}	M_{3c}	M_{3d}	M_{4a}	M_{4b}	M_5
Param.	0	1	1	1	2	K	K	K	$2K$	$p+K$	$p+K+1$	$K(p+2)$

Table 1: Number of parameters to estimate for each model of the proposed family.

- M_{3b} : $\lambda_{k0} = \lambda_k$, $\lambda_{kj} = 1$ and $\sigma_k^* = \sigma_k \quad \forall 1 \leq j \leq p$,
- M_{3c} : $\Lambda_k = \lambda_k I_d$ and $\sigma_k^* = \lambda_k \sigma_k$,
- M_{3d} : $\lambda_{k0} = \lambda_{k0}$, $\lambda_{kj} = \lambda_{k1}$ and $\sigma_k^* = \lambda_{k1} \sigma_k \quad \forall 1 \leq j \leq p$,
- M_4 models assume the link between both populations is mixture component independent (σ_k^* free):
 - M_{4a} : $\lambda_{k0} = 1$ and $\lambda_{kj} = \lambda_j \quad \forall 1 \leq j \leq p$,
 - M_{4b} : $\Lambda_k = \Lambda$ with Λ a diagonal matrix,
- M_5 assumes Λ_k is unconstrained, which leads to estimate the mixture regression model for P^* by using only S^* (σ_k^* free).

Let us remark that transformation other models could be defined, in particular by considering that only the variance component is different between the source and target populations. Even though, only the previous models are investigated in this paper, the practitioner can easily introduced other models if needed, by following the strategy presented here.

Moreover, the mixing proportions are allowed to be the same in each population or to be different. In the latter case, they consequently have to be estimated using the sample S^* . Corresponding notations for the models are respectively pM . when the mixing proportion of P^* have to be estimated and M . when not. Table 1 gives the number of parameters to estimate for each model. If the mixing proportions are different from P to P^* , $K - 1$ parameters to estimate must be added to these values. The estimation of the models M_2 to M_4 are derived in the next subsection.

Let us also remark that by only assuming that the function ψ_k (defined at the beginning of this section) is \mathcal{C}^1 , rather than assuming (4), Biernacki et al. (2002) proves that ψ_k is necessarily affine, and then $Y_{|x, z_{ik}=1}^*$ have the same probability distribution $\lambda_{k1} + \lambda_{k2} Y_{|x, z_{ik}=1}$, where $(\lambda_{k1}, \lambda_{k2}) \in \mathbb{R}^2$. We therefore obtain the following relationship between the model parameters of P and P^* :

$$\beta_k^* = (\lambda_{k1} + \lambda_{k2} \beta_{k0}, \lambda_{k2} \beta_{k1}, \dots, \lambda_{k2} \beta_{kp})^t, \quad (5)$$

$$\sigma_k^* = \lambda_{k2} \sigma_k. \quad (6)$$

The model M_{3d} previously defined, which is the most general model among the M_2 and M_3 classes of models, is equivalent to the model defined by relations (5) and (6). M_4 -type models allow to introduce more flexibility in the proposed model.

2.2. Parameter estimation

In the situation under review in this paper, the mixture of regressions is assumed to be known (β_k and σ_k will be estimated in practice from a sample of sufficient size) for the source population P , and the goal is to estimate the mixture of regressions for P^* . This will be done in two steps. In the first step, the link parameters Λ_k and the mixing proportions π_k^* are estimated as well as the residual variances σ_k^{*2} when necessary (models M_4). In the second step, the estimation of the mixture regression parameters β_k^* and the residual variances σ_k^{*2} (for models M_2 and M_3) are deduced by plug-in through equations (4) and (6). In the following, only the situation where mixing proportions are different from those of population P is considered.

The estimation of the link parameters is carried out by maximum likelihood using a missing data approach *via* the EM algorithm (Dempster et al., 1977). This technique is certainly the most popular approach for inference in mixtures of regressions (see Leisch (2004) for instance). Conditionally to a sample $S^* = (\mathbf{y}^*, \mathbf{x}^*)$ of observations, where $\mathbf{y}^* = (y_1^*, \dots, y_n^*)$ and $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$, the log-likelihood of model (3) is given by:

$$L(\theta; \mathbf{y}^*, \mathbf{x}^*) = \sum_{i=1}^{n^*} \ln \left(\sum_{k=1}^K \pi_k^* \phi(y_i^* | x_i^{*t} \Lambda_k \beta_k, \sigma_k^{*2}) \right), \quad (7)$$

with $\theta = (\pi_1^*, \dots, \pi_K^*, \Lambda_1, \dots, \Lambda_K, \sigma_1^*, \dots, \sigma_K^*)$, and the complete log-likelihood is:

$$L_c(\theta; \mathbf{y}^*, \mathbf{x}^*, \mathbf{z}^*) = \sum_{i=1}^{n^*} \sum_{k=1}^K z_{ik}^* \ln (\pi_k^* \phi(y_i^* | x_i^{*t} \Lambda_k \beta_k, \sigma_k^{*2})), \quad (8)$$

where $\mathbf{z}^* = (z_{ik}^*)_{i=1, n^*, k=1, K}$ is the unobserved latent variable, introduced in Section 2, and assumed to be distributed as a one order multinomial $\mathcal{M}(1, \pi_1^*, \dots, \pi_K^*)$.

The E step. From a current value $\theta^{(q)}$ of the parameter θ , the E step of the EM algorithm consists in computing the conditional expectation of the complete log-likelihood:

$$\begin{aligned} Q(\theta, \theta^{(q)}) &= E_{\theta^{(q)}}[L_c(\theta; \mathbf{y}^*, \mathbf{x}^*, \mathbf{z}^*) | \mathbf{y}^*, \mathbf{x}^*] \\ &= \sum_{i=1}^{n^*} \sum_{k=1}^K t_{ik}^{(q)} (\ln(\pi_k^*) + \ln(\phi(y_i^* | x_i^{*t} \Lambda_k \beta_k, \sigma_k^{*2}))), \end{aligned} \quad (9)$$

where:

$$t_{ik}^{(q)} = E[z_{ik}^* | \mathbf{y}^*, \mathbf{x}^*] = P(z_{ik}^* = 1 | \mathbf{y}^*, \mathbf{x}^*) = \frac{\pi_k^{*(q)} \phi(y_i^* | x_i^{*t} \Lambda_k^{(q)} \beta_k, \sigma_k^{*2(q)})}{\sum_{l=1}^K \pi_l^{*(q)} \phi(y_i^* | x_i^{*t} \Lambda_l^{(q)} \beta_l, \sigma_l^{*2(q)})} \quad (10)$$

is the conditional probability for the observation i to belong to the k -th mixture component.

The *M* step. The *M* step of the EM algorithm consists in choosing the value $\theta^{(q+1)}$ which maximizes the conditional expectation Q computed in the *E* step:

$$\theta^{(q+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}} Q(\theta; \theta^{(q)}) \quad (11)$$

where Θ is a parameter space depending on the model at hand.

For the mixing proportions, the maximum is as usual reached for:

$$\pi_k^{(q+1)} = \frac{1}{n^*} \sum_{i=1}^{n^*} t_{ik}^{(q)}. \quad (12)$$

For the residual variances (models M_4), we have:

$$\sigma_k^{*2(q+1)} = \frac{1}{\sum_{i=1}^{n^*} t_{ik}^{(q)}} \sum_{i=1}^{n^*} t_{ik}^{(q)} (y_i^* - x_i^{*t} \Lambda_k^{(q)} \beta_k)^2. \quad (13)$$

The reminder of this section details the maximisation of the link parameters:

- for model pM_{2a} : $\lambda^{(q+1)}$ is the positive solution of the quadratic equation

$$n^* \lambda^2 + \lambda \sum_{i=1}^{n^*} \sum_{k=1}^K \frac{t_{ik}^{(q)} (y_i^* - \beta_{k0}) x_{i\sim 0}^{*t} \beta_{k\sim 0}}{\sigma_k^2} - \sum_{i=1}^{n^*} \sum_{k=1}^K \frac{t_{ik}^{(q)} (y_i^* - \beta_{k0})^2}{\sigma_k^2} = 0$$

where $x_{i\sim 0}^* = (x_{i1}^*, \dots, x_{ip}^*)$ is the vector x_i^* without its first component x_{i0}^* , and similarly $\beta_{k\sim 0} = (\beta_{k1}, \dots, \beta_{kp})$,

- for model pM_{3a} : $\lambda_k^{(q+1)}$ is the positive solution of the quadratic equation

$$n_k^* \lambda_k^2 + \lambda_k \sum_{i=1}^{n^*} \frac{t_{ik}^{(q)} (y_i^* - \beta_{k0}) x_{i\sim 0}^{*t} \beta_{k\sim 0}}{\sigma_k^2} - \sum_{i=1}^{n^*} \frac{t_{ik}^{(q)} (y_i^* - \beta_{k0})^2}{\sigma_k^2} = 0$$

where $x_{i\sim 0}^* = (x_{i1}^*, \dots, x_{ip}^*)$ is the vector x_i^* without its first component x_{i0}^* , similarly $\beta_{k\sim 0} = (\beta_{k1}, \dots, \beta_{kp})$, and $n_k^* = \sum_{i=1}^{n^*} t_{ik}^{(q)}$,

- for model pM_{2b} : $\lambda^{(q+1)} = \left(\sum_{i=1}^{n^*} \sum_{k=1}^K \frac{t_{ik}^{(q)}}{\sigma_k^2} \beta_{k0}^2 \right)^{-1} \sum_{i=1}^{n^*} \sum_{k=1}^K \frac{t_{ik}^{(q)}}{\sigma_k^2} (y_i^* - x_{i\sim 0}^{*t} \beta_{k\sim 0}) \beta_{k0}$,

- for model pM_{3b} : $\lambda_k^{(q+1)} = \left(\sum_{i=1}^{n^*} \frac{t_{ik}^{(q)}}{\sigma_k^2} \beta_{k0}^2 \right)^{-1} \sum_{i=1}^{n^*} \frac{t_{ik}^{(q)}}{\sigma_k^2} (y_i^* - x_{i\sim 0}^{*t} \beta_{k\sim 0}) \beta_{k0}$,

- for model pM_{2c} : $\lambda^{(q+1)}$ is the positive solution of the quadratic equation

$$n^* \lambda^2 + \lambda \sum_{i=1}^{n^*} \sum_{k=1}^K \frac{t_{ik}^{(q)} y_i^* x_i^{*t} \beta_k}{\sigma_k^2} - \sum_{i=1}^{n^*} \sum_{k=1}^K \frac{t_{ik}^{(q)} y_i^{*2}}{\sigma_k^2} = 0$$

- for model pM_{3c} : $\lambda_k^{(q+1)}$ is the positive solution of the quadratic equation

$$n_k^* \lambda_k^2 + \lambda_k \sum_{i=1}^{n^*} \frac{t_{ik}^{(q)} y_i^* x_i^{*t} \beta_k}{\sigma_k^2} - \sum_{i=1}^{n^*} \frac{t_{ik}^{(q)} y_i^{*2}}{\sigma_k^2} = 0$$

For the model pM_{2d} , as two interdependent scalar parameters λ_0 and λ_1 are considered, no analytical formulae are available for the global maximum on both λ_0 and λ_1 . In such a situation, an easy way to carry out the maximization in this case is to consider a descending algorithm in which λ_0 and λ_1 are alternatively maximized. Using such a strategy incorporated in a EM algorithm is very frequent and, in such a case, the algorithm is called GEM (generalized EM, (Dempster et al., 1977)). Update formulas for these two parameters are consequently:

$$\lambda_0^{(q+1)} = \frac{\sum_{i=1}^{n^*} \sum_{k=1}^K t_{ik}^{(q)} \beta_{k0} (y_i^* - \lambda_1^{(q+1)} x_{i \sim 0}^{*t} \beta_{k \sim 0}) \sigma_k^{-2}}{\sum_{i=1}^{n^*} \sum_{k=1}^K t_{ik}^{(q)} \beta_{k0}^2 \sigma_k^{-2}},$$

and $\lambda_1^{(q+1)}$ is the positive solution of the quadratic system

$$n^* \lambda_1^2 + \lambda_1 \sum_{i=1}^{n^*} \sum_{k=1}^K \frac{t_{ik}^{(q)}}{\sigma_k^2} x_{i \sim 0}^{*t} \beta_{k \sim 0} (y_i^* - \lambda_0^{(q+1)} \beta_{k0}) - \sum_{i=1}^{n^*} \sum_{k=1}^K \frac{t_{ik}^{(q)}}{\sigma_k^2} (y_i^* - \lambda_0^{(q+1)} \beta_{k0})^2 = 0$$

For the model pM_{3d} , the same algorithm is considered with the following update formulas:

$$\lambda_{k0}^{(q+1)} = \frac{\sum_{i=1}^{n^*} t_{ik}^{(q)} (y_i^* - \lambda_{k1}^{(q+1)} x_{i \sim 0}^{*t} \beta_{k \sim 0})}{\sum_{i=1}^{n^*} t_{ik}^{(q)} \beta_{k0}},$$

and $\lambda_{k1}^{(q+1)}$ is the positive solution of the quadratic system

$$n_k^* \lambda_{k1}^2 + \lambda_{k1} \sum_{i=1}^{n^*} \frac{t_{ik}^{(q)}}{\sigma_k^2} x_{i \sim 0}^{*t} \beta_{k \sim 0} (y_i^* - \lambda_{k0}^{(q+1)} \beta_{k0}) - \sum_{i=1}^{n^*} \frac{t_{ik}^{(q)}}{\sigma_k^2} (y_i^* - \lambda_{k0}^{(q+1)} \beta_{k0})^2 = 0$$

Models M_{4a} and M_{4b} have respectively p and $p + 1$ scalar parameters plus the residual variance. A descending algorithm has to be used for alternatively maximizing the variances (by (13)) and each scalar link parameter. Update formulas for the link parameters are the following:

- model M_{4a} , $\forall 1 \leq J \leq p$:

$$\lambda_J^{(q+1)} = \left(\sum_{i=1}^n \sum_{k=1}^K \frac{t_{ik}^{(q)}}{\sigma_k^{*2}} (x_{iJ}^* \beta_{kJ})^2 \right)^{-1} \sum_{i=1}^n \sum_{k=1}^K \frac{t_{ik}^{(q)}}{\sigma_k^{*2}} x_{iJ}^* \beta_{kJ} \left(y_i^* - \beta_{k0} - \sum_{j=1, j \neq J}^p \lambda_j^{(q+1)} x_{ij}^* \beta_{kj} \right),$$

- model M_{4b} , $\forall 0 \leq J \leq p$:

$$\lambda_J^{(q+1)} = \left(\sum_{i=1}^n \sum_{k=1}^K \frac{t_{ik}^{(q)}}{\sigma_k^{*2}} (x_{iJ}^* \beta_{kJ})^2 \right)^{-1} \sum_{i=1}^n \sum_{k=1}^K \frac{t_{ik}^{(q)}}{\sigma_k^{*2}} x_{iJ}^* \beta_{kJ} \left(y_i^* - \sum_{j=0, j \neq J}^p \lambda_j^{(q+1)} x_{ij}^* \beta_{kj} \right).$$

with $x_{i0} = 1$ for all $1 \leq i \leq n$.

The EM algorithm stops when the difference of the likelihood value of two consecutive steps is lower than a given threshold ε (typically $\varepsilon = 10^{-6}$).

Prediction rule. Once the model parameters have been estimated, the prediction \hat{y}_i^* of the response variable corresponding to an observation x_i of X is obtained by a two step procedure. First, the component membership z_i of x_i is estimated by the maximum a posteriori rule. It consists in assigning x_i to the k th component ($z_{ik} = 1$) which maximizes the conditional probability (10) obtained at the last step of the EM algorithm. Then, \hat{y}_i^* is predicted using the k th regression model of the mixture:

$$\hat{y}_i^* = x_i \hat{\beta}_k \quad \text{where} \quad k = \underset{1 \leq k \leq K}{\operatorname{argmax}} t_{ik}^{(q_{last})} \quad (14)$$

with q_{last} the number of EM steps until convergence.

2.3. Model selection

In order to select among the 24 transformation models defined in Section 2 the most appropriate model of transformation between the populations P and P^* , we propose to use two well known criteria. The reader interested in a comparison of the respective performances of models selection criteria could refer to Hastie et al. (2001) for instance. The first considered criterion is the PRESS criterion (Allen, 1974), which represents the mean squared prediction error computed on a cross-validation scheme, formally defined by:

$$PRESS = \sum_{i=1}^{n^*} (y_i^* - \hat{y}_{(i)}^*)^2$$

where $\hat{y}_{(i)}^*$ is the prediction of y_i^* obtained by the mixture regression model estimated without using the i th observation of the sample S^* . This criterion is one of the most often used for model selection in regression analysis, and we encourage its use when it is computationally feasible. The second considered criterion is the Bayesian Information Criterion (BIC, Schwarz

(1978)), which is a penalized likelihood criterion with a less computation cost. The BIC criterion is defined by:

$$BIC = -2 \ln \ell + v \ln n^*,$$

where ℓ is the maximum log-likelihood value and v is the number of estimated parameters (see Table 1). It consists in selecting the models leading to the highest likelihood while penalizing models with a large number of parameters. Let us precise that, for both criteria, the most adapted model is the one with the smallest criterion value.

3. Bayesian approach for adaptive mixture of regressions

The previous section has considered the modelling and the estimation of parametric adaptive models for mixture of regressions with the classical frequentist point of view. This section adopts a Bayesian approach for inferring adaptive mixture of regressions and Gibbs sampling is considered for the estimation of the posterior distribution.

3.1. A Bayesian view of the problem

The classical treatment of the mixture regression problem seeks a point estimate of the unknown regression parameters. By contrast, the Bayesian approach (Hurn et al., 2003; Robert, 2007) characterizes the uncertainty on parameters through a probability distribution, called a prior distribution. Bayesian analysis combines the prior information on the parameters (carried out by the prior distribution) with information on the current sample (through the likelihood function) to provide estimates of the parameters using the posterior distribution. In the context of adaptive mixture of regressions, the Bayesian approach makes particularly sense since there is an actual prior on the model parameters of population P^* . Indeed, even though source and target populations differ, they are here assumed to have a strong link and it is therefore natural to define the prior on parameters of population P^* according to the ones of population P .

In the context of mixture of regressions, it is usual to assume the conditional independence between the mixing parameters π^* and both component parameters $\beta^* = \{\beta_1^*, \dots, \beta_K^*\}$ and $\sigma^{*2} = \{\sigma_1^{*2}, \dots, \sigma_K^{*2}\}$. The independence between $(\beta_k^*, \sigma_k^{*2})$ and $(\beta_\ell^*, \sigma_\ell^{*2})$ is as well assumed for all $k \neq \ell$, $k, \ell = 1, \dots, K$. For simplicity, only conjugate priors are considered in this work and, since model parameters of the reference population P are assumed to be known, prior distributions of the parameters of population P^* will depend on model parameters of the population P . We therefore propose to assume that, for all $k = 1, \dots, K$, the prior distribution for β_k^* is a normal distribution centered in β_k :

$$\beta_k^* \sim \mathcal{N}(\beta_k, \sigma_k^{*2} A_k),$$

where A_k is a $(p+1) \times (p+1)$ covariance matrix. The prior distribution of σ_k^{*2} , for all $k = 1, \dots, K$, is assumed to be an inverse-gamma distribution:

$$\sigma_k^{*2} \sim \mathcal{IG}(\gamma_k, \nu_k).$$

The prior distribution for parameters $\pi^* = \{\pi_1^*, \dots, \pi_K^*\}$ is assumed to be a Dirichlet distribution centered in the mixing proportions (π_1, \dots, π_K) of population P :

$$\pi^* \sim \mathcal{D}(\pi_1, \dots, \pi_K).$$

With such a modelling, the regression coefficients β_k^* and the mixing proportions $\pi^* = \{\pi_1^*, \dots, \pi_K^*\}$ of population P^* are naturally linked to the ones of population P . The variance terms $\sigma_k^{*2} A_k$ control how the regression coefficients β_k^* differ from the ones of the reference population P . In the experiments presented in Section 4, the prior parameters ν_k , γ_k and A_k , $k = 1, \dots, K$, were respectively set to 1, 2 and the identity matrix.

Finally, by combining the likelihood of the mixture of regressions model and the priors, we end up with the joint posterior distribution:

$$p(\theta^* | Y^*) \propto \prod_{i=1}^{n^*} \left[\sum_{k=1}^K \pi_k^* \phi(y_i^* | x_i^{*t} \beta_k^*, \sigma_k^{*2}) \right] p(\pi^*) \prod_{k=1}^K [p(\beta_k^* | \sigma_k^{*2}) p(\sigma_k^{*2})],$$

where $\theta^* = (\pi_k^*, \beta_k^*, \sigma_k^{*2})_{k=1, \dots, K}$. However, since the posterior distribution $p(\theta^* | Y^*)$ takes into account all possible partitions of the sample into K groups, the maximization of $p(\theta^* | Y^*)$ is intractable even with moderately large sample size and Markov Chain Monte Carlo methods have to be used.

3.2. Gibbs sampler for adaptive mixture of regressions

Markov Chain Monte Carlo methods allow to approximate a complicated distribution by using samples drawn indirectly from this distribution. Among MCMC methods, the Gibbs sampler is the most commonly used approach when dealing with mixture distribution (Diebolt and Robert (1994)). In Gibbs sampling, the vector parameter θ^* is partitioned into s groups of parameters $\{\theta_1^*, \dots, \theta_s^*\}$ and a Markov chain is generated by iteratively sampling from the conditional posterior distributions. Once a Markov chain of length Q has been generated, sample values can be averaged on the last sampling iterations to provide consistent estimates of model parameters. In the context of inference for mixture distribution, the Gibbs sampler requires to add a latent variable $Z^* \in \{0, 1\}^K$ representing the allocation of observations to the K mixture components (introduced in Section 2). Since the latent variable Z^* is not observed, Z^* can be viewed as unknown and should be estimated along with the other model parameters. Consequently, given estimates $\hat{\beta}$ and $\hat{\pi}$ of respectively regression parameters and mixing proportions of population P and starting from initial values $\pi^{*(0)}$, $\beta^{*(0)}$ and $\sigma^{*2(0)}$, the Gibbs algorithm generates, at iteration q , parameter values from the conditional posterior distributions:

$$\begin{aligned} Z^{*(q)} &\sim p(Z | Y^*, \hat{\beta}, \hat{\pi}, \pi^{*(q-1)}, \beta^{*(q-1)}, \sigma^{*2(q-1)}), \\ \pi^{*(q)} &\sim p(\pi^* | Y^*, \hat{\beta}, \hat{\pi}, Z^{*(q)}, \beta^{*(q-1)}, \sigma^{*2(q-1)}), \\ \sigma_k^{*2(q)} &\sim p(\sigma_k^{*2} | Y^*, \hat{\beta}, \hat{\pi}, Z^{*(q)}, \pi^{*(q)}, \beta^{*(q-1)}), \\ \beta_k^{*(q)} &\sim p(\beta_k^* | Y^*, \hat{\beta}, \hat{\pi}, Z^{*(q)}, \pi^{*(q)}, \sigma^{*2(q-1)}). \end{aligned}$$

According to the priors given in the previous paragraph, the conditional posterior distribution of Z^* is a multinomial distribution:

$$z_i^* | Y^*, \hat{\beta}, \hat{\pi}, \pi^*, \beta^*, \sigma^{*2} \sim \mathcal{M}(1, t_{i1}, \dots, t_{iK}),$$

where $t_{ik} = \pi_k^* \phi(y_i^* | x_i^{*t} \beta_k^*, \sigma_k^{*2}) / \sum_{\ell=1}^K \pi_\ell^* \phi(y_i^* | x_i^{*t} \beta_\ell^*, \sigma_\ell^{*2})$, and the conditional posterior distribution of π^* is a Dirichlet distribution:

$$\pi^* | Y^*, \hat{\beta}, \hat{\pi}, Z^*, \beta^*, \sigma^{*2} \sim \mathcal{D}(\hat{\pi}_1 + n_1^*, \dots, \hat{\pi}_K + n_K^*),$$

with $n_k^* = \sum_{i=1}^n z_{ik}^*$. Once the component belongings of each observation are known, the observations of the same component k can be gathered into the matrices x_k^* and Y_k^* , for all $k = 1, \dots, K$. With these notations, the conditional posterior distribution of σ_k^{*2} is an inverse gamma:

$$\sigma_k^{*2} | Y^*, \hat{\beta}, \hat{\pi}, Z^*, \pi^*, \beta_k^* \sim \mathcal{IG}(\gamma_k + n_k/2, \nu_k + S_k/2),$$

where $S_k = (Y_k^* - x_k^{*t} \beta_k^*)^t (Y_k^* - x_k^{*t} \beta_k^*) + (\hat{\beta}_k - \beta_k^*)^t (A_k + (x_k^{*t} x_k^*)^{-1})^{-1} (\hat{\beta}_k - \beta_k^*)$, and the conditional posterior distribution of β_k^* is a normal distribution:

$$\beta_k^* | Y^*, \hat{\beta}, \hat{\pi}, Z^*, \pi^*, \sigma_k^{*2} \sim \mathcal{N}(m_k, \Delta_k),$$

with

$$\begin{aligned} m_k &= (A_k^{-1} + x_k^{*t} x_k^*)^{-1} (x_k^{*t} Y_k^* + A_k^{-1} \hat{\beta}_k), \\ \Delta_k &= \sigma_k^{*2} (x_k^{*t} x_k^* + A_k^{-1})^{-1}. \end{aligned}$$

Finally, consistent estimates of model parameters π^* , β^* and σ^{*2} are obtained by averaging on the last $Q - q_0$ sampling iterations, where q_0 defines the number of iterations of the so called ‘‘burning phase’’ of the Gibbs sampler.

3.3. The label switching problem

When simulating a Markov chain to estimate parameters of a mixture model, the label switching problem frequently arises and is due to the multimodality of the likelihood. Indeed, if the prior distributions are symmetric, the posterior distribution inherits the multimodality of the likelihood. In such a case, the Markov chain can move from one mode to another and it is difficult to deduce consistent estimators of model parameters. The earliest solution, proposed by Richardson and Green (1997), consists in adding indenifiability constraints on model parameters such as an order relation in mixing proportions. Unfortunately, this approach does not work very well as showed by ?. By contrast, some authors like ? and Stephens (2000) propose to work *a posteriori* on the generated Markov chain in order to reorganize it according to a specific criterion. The Stephens’ procedure reorganizes the Markov chain by searching the correct permutations of mixture component which minimizes a divergence criterion. The solutions proposed by Celeux *et al.* are in the same spirit and, among the different proposed criteria, they propose in particular to reorganize the Markov chain using a sequential k -means

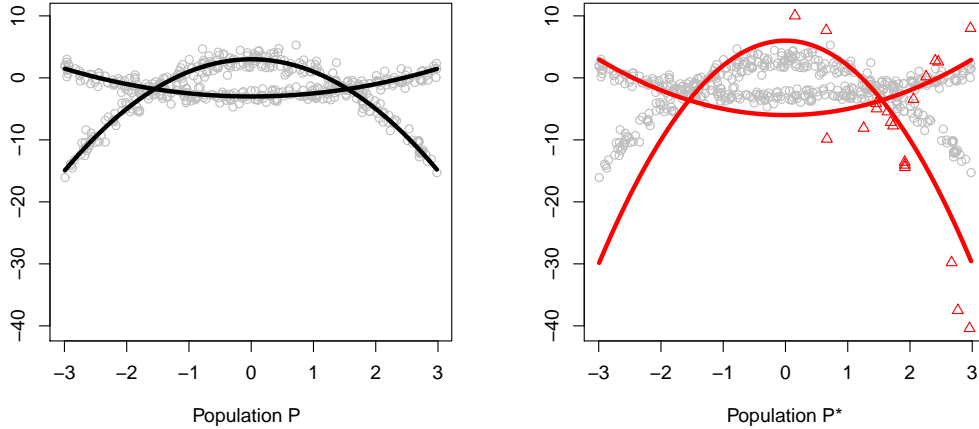


Figure 2: Populations P and P^* used for the introductory example. Curves black (left) and red (right) indicates respectively the actual mixture regression of populations P and P^* .

algorithm. Both the Stephens and Celeux's approaches are efficient to deal with the label switching problem. However, the sequential k -means algorithm has the advantage to be less memory consuming and, in the experiments presented in Section 4, this approach is used to overcome the label switching problem.

4. Experimental results

This section proposes experiments on simulated and real data in order to highlight the main features of the adaptive models proposed in the previous sections. After an introductory example, the behavior of adaptive mixtures of regressions (parametric and Bayesian) is compared to the one of classical mixtures of regressions on simulated data. The last experiment will demonstrate the interest of using adaptive mixtures of regressions on an illustrative real dataset, and where the size of the target population sample will be artificially moved from small to larger sizes.

4.1. An introductory example

This first experiment aims to compare the basic behaviors of adaptive mixtures of regressions (parametric and Bayesian), hereafter referred to as AMR (respectively AMRp and AMRb), and classical mixtures of regressions, referred to as MR. For this study, the reference population P is modeled by a 2 component mixture of quadratic polynomial regressions with parameters $\beta_1 = (3, 0, -2)$ and $\beta_2 = (-3, 0, 0.5)$. The left panel of Figure 2 shows the mixture regression of population P as well as some observations simulated from this model. The mixture model of population P^* has then been obtained from the previous model by multiplying all regression parameters of population P by a factor 3. It follows that $\beta_1^* = (9, 0, -6)$ and

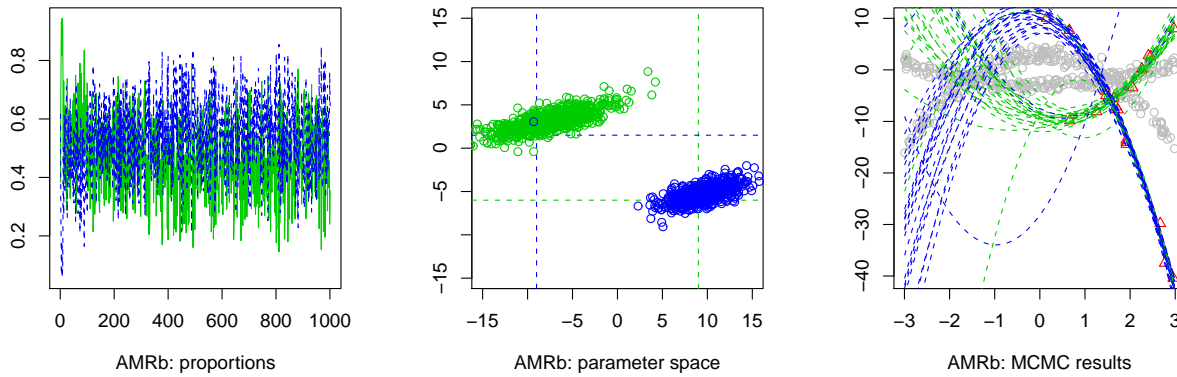


Figure 3: Results obtained for the introductory example with the Bayesian approach of adaptive mixture of regressions (AMRb). From left to right: mixing proportions over the MCMC iterations, Gibbs sampling in the parameter space and some of the generated regression curves. See text for details.

$\beta_2^* = (-9, 0, 1.5)$. Finally, 20 observations of population P^* have been simulated using the latter model on $[0, 3]$. The right panel of Figure 2 shows the actual mixture regression model of population P^* as well as the 20 simulated observations (red triangles). These 20 observations of P^* were used by the three studied regression methods to estimate the regression model of P^* and to predict the value of 5000 validation observations of P^* . The mean square error (MSE), computed on the validation sample, has been chosen to evaluate the predicting ability of each regressions method in this introductory example.

Figure 3 illustrates the estimation procedure of the Bayesian approach on this toy dataset. The MCMC procedure was made of 1 000 sampling iterations including a burning phase of 100 iterations. The left panel of Figure 3 shows the sampled proportions over the MCMC iterations. As one can see, after the burning phase, the proportions of both mixture components stabilize in the neighborhood of 0.5 which is the actual value of π_1 and π_2 . The central panel presents the sampled values for regression parameters β_1 and β_2 in the parameter space (restricted to β_{k1} and β_{k3} for $k = 1, 2$ because both β_{12} and β_{22} are both equal to 0). The blue and green dashed lines indicate at the intersections the actual values of regression parameters. It appears that the Bayesian approach succeeds in estimating the conditional distributions of regression parameters. Finally, the right panel exhibits some of the 1 000 regression models generated during the MCMC iterations which are then used to provide by averaging the final estimated regression model of P^* .

Figure 4 presents the results obtained for the considered example with the classical mixture of regressions (MR), parametric adaptive mixture of regressions (AMRp) and Bayesian adaptive mixture of regressions (AMRb). The MR method used only the 20 observations sampled from P^* whereas AMR and AMRb combines the informations carried by these observations with the knowledge on P to build their estimation of the mixture regression model of P^* . In order to not favor the adaptive approaches, the actual number of components and

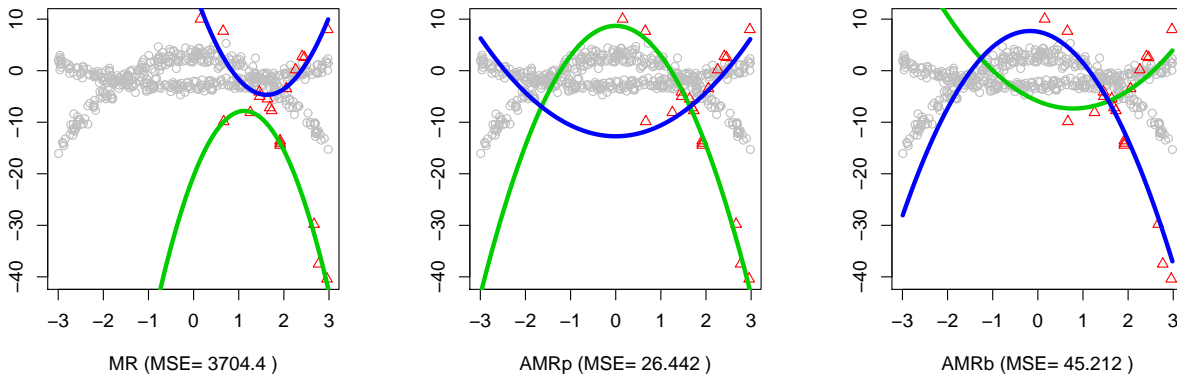


Figure 4: Results obtained for the introductory example with classical mixture of regressions (MR), parametric adaptive mixture of regressions (AMRp) and Bayesian adaptive mixture of regressions (AMRb) methods. See text for details.

dimension of the polynomial regression were also provided to the MR method. Nevertheless, the MR method provides a poor estimate of the regression model and its mean square error (MSE) value, computed on an independent validation set, is consequently high (3704.4). Conversely, the parametric (with the model pM_{3c}) and Bayesian approaches of AMR give good estimations of the P^* model (they should be compared to the red curves of Fig. 3). The associated MSE values are naturally much lower than the one of the classical MR method (26.4 for AMRp and 45.2 for AMRb). Nonetheless, the Bayesian approach performs less than the parametric AMRp. This could be due to the fact that AMRb favors too much the prior (the regression parameters of P) in this situation with only few observations of the new population. This introductory example has shown that adaptive regression models succeed in transferring the knowledge of a reference population to a new population.

4.2. Influence of the size of S^*

The second experiment focuses on the influence of the number of observations n^* from the new population P^* on the estimation quality of mixture regression models for the MR, AMRp and AMRb methods. The experimental setup is the same as for the previous experiment except that the number of observations n^* from the new population P^* varies from 6 to 200. For each value of n^* , the regression model of P^* has been estimated with the three studied methods and the associated MSE values have been computed again on an independent validation set of 5 000 observations. Finally, the experiment has been replicated 50 times in order to average the results. Figure 5 shows the evolution of the median logarithm of the MSE value according to the size of S^* for the classical mixture of regressions (MR), parametric adaptive mixture of regressions (AMRp) and Bayesian adaptive mixture of regressions (AMRb) methods. For the parametric approach of the AMR method, the model used is pM_{3c} . Associated boxplots are presented by Figure 6 on a logarithmic scale. On view of Figure 5, it can be first noticed that the performance of the classical MR method is sensitive to the size of S^* . Indeed,

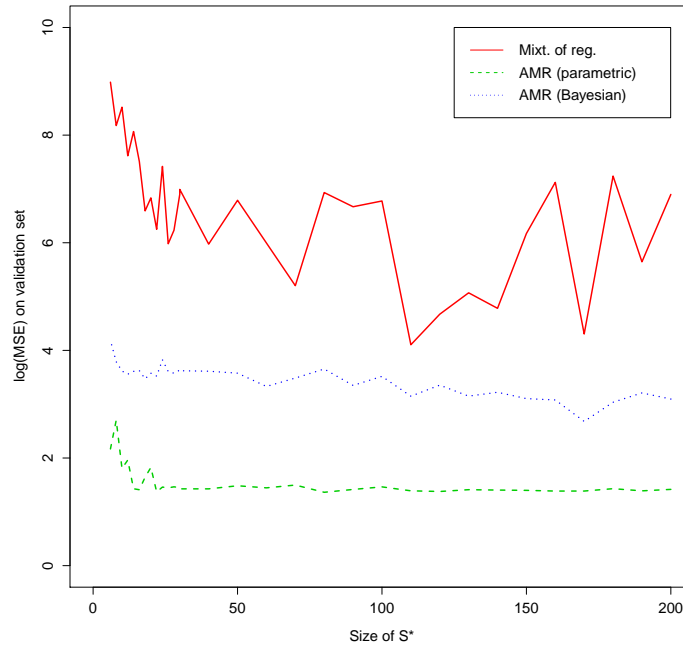


Figure 5: Average logarithm of the MSE value according to the the size of S^* for the classical mixture of regressions (MR), parametric adaptive mixture of regressions (AMRp) and Bayesian adaptive mixture of regressions (AMRb) methods.

for small sample sizes, the MR method provides poor estimates of the mixture regression model of population P^* and this consequently yields poor prediction performances (large MSE values). As one can expect, the model estimation and the prediction improve when the number of observations n^* from the new population P^* increases. More surprisingly, as it can be observed on the left panel of Figure 6, the variance of the prediction performance of the MR method remains large even for sample sizes bigger than 100. This remind us that the fitting of a mixture regression model is always a difficult task. Conversely, the adaptive methods AMRp and AMRb which exploit their knowledge on the reference population obtain on average good prediction results (low MSE values) and this even for very small numbers of observations n^* . In particular, the parametric approach AMRp provides very stable prediction results and its variance decreases quickly when n^* increases. The Bayesian approach AMRb, even though it is much efficient and stable than the classical MR method, appears to be slightly less efficient than the parametric approach AMRp. To summarize, this study on simulations has shown that adaptive regression models greatly improve the prediction and reduce the predictor variance compared to the classical mixture regression approach when the number of observations of the new population is small.

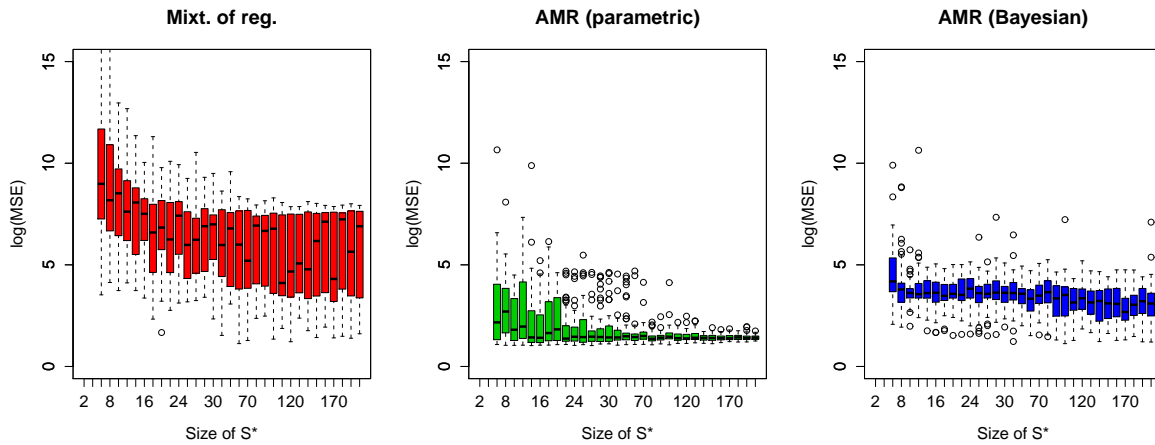


Figure 6: Boxplots of MSE values (on logarithmic scale) according to the the size of S^* for the classical mixture of regressions (left), parametric adaptive mixture of regressions (center) and Bayesian adaptive mixture of regressions (right) methods.

4.3. Illustration on real data: CO_2 emissions vs gross national product

In this last experiment, the link between CO_2 emission and gross national product (GNP) of various countries is investigated. The sources of the data are *The official United Nations site for the Millennium Development Goals Indicators* and the *World Development Indicators of the World Bank*. Figure 7 plots the CO_2 emission per capita *versus* the logarithm of GNP per capita for 111 countries, in 1980 (left) and 1999 (right). A mixture of second order polynomial regressions seems to be particularly well adapted to fit these data and will be used in the following. Let remark that regression model with heteroscedasticity could also be appropriated for such data, but these kind of models are out of the topic of the present work. For the 1980's data, two groups of countries are easily distinguishable: a first minority group (about 25% of the whole sample) is made of countries for which a grow in the GNP is linked to a high grow of the CO_2 emission, whereas the second group (about 75%) seems to have more environmental political orientations. As pointed out by Hurn et al. (2003), the study of such data could be particularly useful for countries with low GNP in order to clarify in which development path they are embarking. This country discrimination in two groups is more difficult to obtain on the 1999's data: it seems that countries which had high CO_2 emission in 1980 have adopted a more environmental development than in the past, and a two-component mixture regression model could be more difficult to exhibit.

In order to help this distinction, parametric adaptive mixture models are used to estimate the mixture regression model on the 1999's data. The ten AMRp models, with free component proportions π_k^* , pM_{2a} to pM_{4b} , AMRb model, classical mixture of second order polynomial regressions with two components (MR) and usual second order polynomial regression (UR) are considered. Different sample size of the 1999's data are tested: 30%, 50%, 70% and 100% of the S^* size ($n^* = 111$). The experiments have been repeated 20 times in order to average

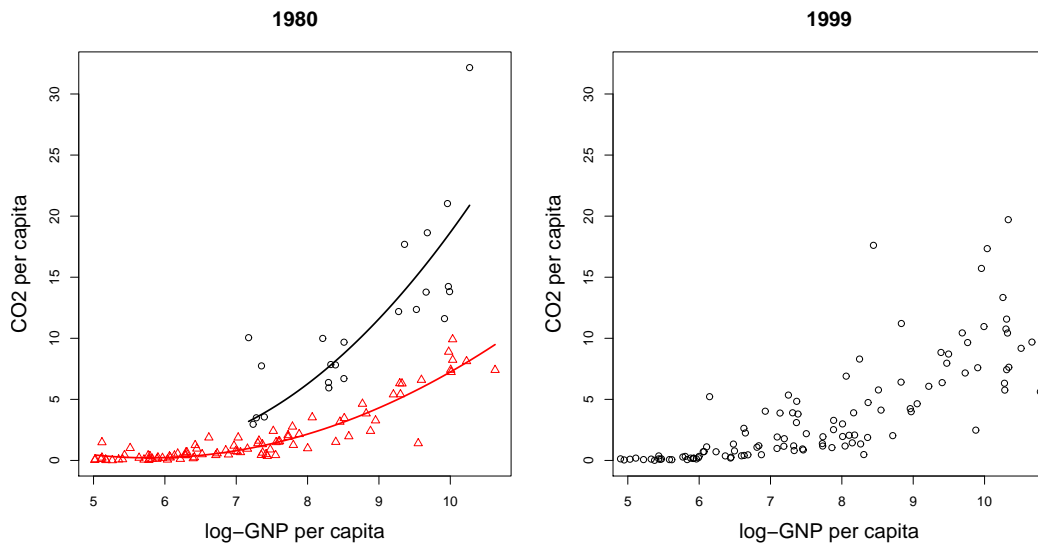


Figure 7: Emission of CO₂ per capita *versus* GNP per capita in 1980 (left) and 1999 (right).

the results. Table 2 summarizes these results: MSE corresponds to the mean square error, whereas PRESS and BIC are the model selection criteria introduced in Section 2.3. In this application, the total number of available data in the 1999 population is not sufficiently large to separate them into two training and test samples. For this reason, MSE is computed on the whole S^* sample, even though a part of it has been used for the training (from 30% for the first experiment to 100% for the last one). Consequently, MSE is a significant indicator of predictive ability of the model when 30% and 50% of the whole dataset are used as training set since 70% and 50% of the samples used to compute the MSE remain independent from the training stage. However, MSE is a less significant indicator of predictive ability for the two last experiments and the PRESS should be preferred in these situations as indicator of predictive ability.

Table 2 first allows to remark that the 1999's data are actually made of two components as in the 1980's data since both PRESS and MSE are better for MR (2 components) than UR (1 component) for all sizes n^* of S^* . This first result validates the assumption that both the reference population P and the new population P^* have the same number $K = 2$ components, and consequently the use of adaptive mixture of regression makes sense for this data. Secondly, AMRp turns out to provide very satisfying predictions for all values of n^* and particularly outperforms the other approaches when n^* is relatively small (less than 77 here). Indeed, both BIC, PRESS and MSE testify that the models of AMRp provide better predictions than the other studied methods when n^* is equal to 30%, 50% and 70% of the whole sample. Furthermore, it should be noticed that AMRp provide stable results according to variations on n^* . In particular, the models pM_2 are those which appear the most efficient on this dataset and this means that the link between both populations P and P^* is mixture component independent. On

30% of the 1999's data ($n^* = 33$)				50% of the 1999's data ($n^* = 55$)			
model	BIC	PRESS	MSE	model	BIC	PRESS	MSE
AMRp (pM_{2a})	13.09	3.38	3.40	AMRp (pM_{2a})	10.18	4.11	3.44
AMRp (pM_{2b})	12.73	3.89	3.32	AMRp (pM_{2b})	13.54	3.73	3.37
AMRp (pM_{2c})	12.79	5.48	3.68	AMRp (pM_{2c})	13.89	4.25	3.45
AMRp (pM_{2d})	11.54	4.99	3.73	AMRp (pM_{2d})	22.35	4.38	4.80
AMRp (pM_{3a})	12.14	4.20	3.76	AMRp (pM_{3a})	12.00	3.84	4.49
AMRp (pM_{3b})	11.72	4.87	4.00	AMRp (pM_{3b})	12.00	4.47	3.86
AMRp (pM_{3c})	11.50	5.09	3.86	AMRp (pM_{3c})	17.53	3.97	3.28
AMRp (pM_{3d})	22.83	5.52	3.64	AMRp (pM_{3d})	25.39	4.77	3.67
AMRp (pM_{4a})	18.72	5.15	4.01	AMRp (pM_{4a})	20.65	3.68	3.44
AMRp (pM_{4b})	22.01	6.21	5.04	AMRp (pM_{4b})	24.92	5.57	4.19
AMRb	-	(†)	5.99	AMRb	-	(†)	5.66
UR	27.08	7.46	7.66	UR	20.87	7.95	7.21
MR	32.89	5.54	5.11	MR	39.69	4.82	4.77

70% of the 1999's data ($n^* = 77$)				$(n^* = 111)$			
model	BIC	PRESS	MSE	model	BIC	PRESS	MSE
AMRp (pM_{2a})	14.76	3.65	3.35	AMRp (pM_{2a})	15.51	4.78	3.32
AMRp (pM_{2b})	14.73	3.91	3.39	AMRp (pM_{2b})	15.44	3.81	3.37
AMRp (pM_{2c})	14.53	4.49	3.53	AMRp (pM_{2c})	15.39	4.84	3.47
AMRp (pM_{2d})	18.90	4.30	3.72	AMRp (pM_{2d})	20.05	4.45	3.59
AMRp (pM_{3a})	18.84	4.33	3.85	AMRp (pM_{3a})	20.18	4.29	3.79
AMRp (pM_{3b})	18.80	4.40	3.85	AMRp (pM_{3b})	20.03	4.38	3.77
AMRp (pM_{3c})	18.81	4.41	3.26	AMRp (pM_{3c})	20.05	3.94	3.10
AMRp (pM_{3d})	27.05	3.91	3.17	AMRp (pM_{3d})	29.37	4.08	3.34
AMRp (pM_{4a})	22.29	5.25	4.00	AMRp (pM_{4a})	23.98	4.21	4.13
AMRp (pM_{4b})	26.55	4.92	4.03	AMRp (pM_{4b})	28.58	5.21	4.52
AMRb	-	(†)	5.99	AMRb	-	(†)	5.66
UR	22.08	8.00	7.10	UR	23.62	7.53	6.99
MR	43.91	5.06	3.33	MR	47.19	3.66	2.89

Table 2: MSE on the whole 1999's sample, PRESS and BIC criterion for the 10 parametric adaptive mixture models (AMRp pM_{2a} to pM_{4b}), AMRb model, usual regression model (UR) and classical regressions mixture model (MR), for 4 sizes of the 1999's sample: 33, 55, 77 and 111 (whole sample). Lower BIC, PRESS and MSE values for each sample size are in bold character. (†): Cross-validation on MCMC procedures is too computationally heavy to be computed in a reasonable time.

the other hand, the Bayesian approach AMRb appears to provide results as stable as the ones of AMRp but slightly less satisfying. The results of the Bayesian approach would probably be better with a more specific choice of the priors.

This application illustrates well the interest of combining informations on both past (1980) and present (1999) situations in order to analyse the link between CO₂ emissions and gross national product for several countries in 1999, especially when the number of data for the present situation is not sufficiently large. Moreover, the competition between the parametric AMR models is also informative. Effectively, it seems that three models are particularly well adapted to model the link between the 1980's data and those of 1999's data: pM_{2a} , pM_{2b} and pM_{2c} . The particularity of these models is that they consider the same transformation for both classes of countries, which means, conversely to what one might *prima facie* have thought, that all the countries have made an effort to reduce their CO₂ emissions and not only those which had the higher ones.

5. Conclusion

We proposed in this paper adaptive models for mixture of regressions in order to improve the predictive inference when the studied population has changed between training and prediction phases. The first class of models considers a parsimonious and parametric link between the mixture of regressions of both populations, whereas the second approach adopts a Bayesian point a view in which the populations are linked by the prior information imposed on the mixture regression parameters. On both simulated data and real data, models considering parametric link turn out to be the most powerful: all the interest of such adaptive methods consists in their sparsity, which leads to significantly decrease the number of observations of the target population required for the estimation. As the indispensable stage of data collecting is often expensive and time consuming, there is a real interest to consider adaptive mixture of regressions in practical applications. Moreover, as it has been showed in the illustration on real data, the competition between the parametric link models provides informations on the link between populations, which can be meaningful for the practitioner.

Regarding the further works, a first perspective concerns the Bayesian approach. In this paper, the prior hyperparameters for σ_k^{*2} were simply fixed to values seeming experimentally reasonable. The results of the Bayesian approach may be improved by working on the choice of these hyperparameters. One generic way to do this is to make similar assumptions as in the frequentist approach. For instance, the variance $\sigma_k^{*2}A_k$ of the regression parameters β_k^* could be assumed to be common between mixture components or to be equal to $\sigma_k^{*2}I_d$. The selection between the considered assumptions could then be done by choosing those maximizing the integrated likelihood (Raftery et al., 2007). A second working perspective is related to the joint estimation of the models of both populations P and P^* . Indeed, the reference regression model being only estimated in practice, the quality of this estimation, depending on the size n of the available sample, is directly responsible of the estimation quality of the mixture regression model for P^* . In some situations (typically when n is small compared to the model complexity), it could be interesting to consider a full likelihood estimation which consists

in estimating simultaneously both mixture regression models. Such an approach has been recently considered in Lourme and Biernacki (2008) in a supervised classification context. It must be emphasized that such a full likelihood estimation of both mixtures of regression must consider the same estimation method (parametric or Bayesian) for both populations.

References

- Allen, D. M., 1974. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16, 125–127.
- Biernacki, C., Beninel, F., Bretagnolle, V., 2002. A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics* 58 (2), 387–397.
- Blitzer, J., Dredze, M., Pereira, F., 2007. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, pp. 432–439.
- Bouveyron, C., Jacques, J., 2010. Adaptive linear models for regression: improving prediction when population has changed. *Pattern Recognition Letters* 31 (14), 2237–2247.
- Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data (with discussion). *Journal of the Royal Statistical Society. Series B* 39, 1–38.
- Diebolt, J., Robert, C., 1994. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B* 56 (2), 363–375.
- Evgeniou, T., Pontil, M., 2004. Regularized multi-task learning. In: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, Washington, USA, pp. 109–117.
- Goldfeld, M., Quandt, R., 1973. A markov model for switching regressions. *Journal of Econometrics* 1, 3–16.
- Hastie, T., Tibshirani, R., Friedman, J., 2001. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York.
- Hennig, C., 1999. *Classification in the Information Age*. Springer-Verlag, Heidelberg.
- Hennig, C., 2000. Identifiability of models for clusterwise linear regression. *Journal of Classification* 17, 273–296.
- Hurn, M., Justel, A., Robert, C. P., 2003. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics* 12 (1), 55–79.

- Jacques, J., Biernacki, C., 2010. Extension of model-based classification for binary data when training and test populations differ. *Journal of Applied Statistics* 37 (5), 749–766.
- Khalili, A., Chen, J., 2007. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* 102 (479), 1025–1038.
- Lawrence, N., Platt, J., 2004. Learning to learn with the informative vector machine. In: *Proceedings of the 21th International Conference on Machine Learning*. Banff, Alberta, Canada.
- Leisch, F., 2004. Flexmix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* 11 (8), 3–16.
- Lourme, A., Biernacki, C., 2008. Gaussian model-based classification when training and test population differ: Estimating jointly related parameters. In: *First joint meeting of the Société Francophone de Classification and of the Classification and Data Analysis Group of SIS*.
- Pan, S., Yang, Q., 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22 (10), 1345–1359.
- Raftery, A. E., Newton, M. A., Satagopan, J. M., Krivitsky, P. N., 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In: *Bayesian statistics 8*. Oxford Sci. Publ. Oxford Univ. Press, Oxford, pp. 371–416.
- Richardson, S., Green, P., 1997. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Serie B* 59, 731–792.
- Robert, C., 2007. *The Bayesian Choice*. Springer.
- Schwarz, G., 1978. Estimating the dimension of a model. *The Annals of Statistics* 6 (2), 461–464.
- Shimodaira, H., 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* 90 (2), 227–244.
- Stephens, M., 2000. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Serie B* 62 (4), 795–809.
- Storkey, A., Sugiyama, M., 2007. Mixture regression for covariate shift. *Advances in Neural Information Processing Systems* 19. MIT Press, Cambridge, pp. 1337–1344.
- Sugiyama, M., 2006. Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research* 7, 141–166.
- Sugiyama, M., Müller, K-R., K. M., 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* 8, 985–1005.

- Sugiyama, M., Müller, K.-R., 2005. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions* 23.
- Zhu, H.-T., Zhang, H., 2004. Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society. Series B.* 66 (1), 3–16.