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

Charles Bouveyron, Julien Jacques

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

Charles Bouveyron[1] & Julien Jacques[2]

[1]Laboratoire SAMM, University Paris I Panthéon-Sorbonne, Paris, France.
charles.bouveyron@univ-paris1.fr
[2] Laboratoire Paul Painlevé, UMR CNRS 8524, University Lille I, Lille, France.
julien.jacques@polytech-lille.fr

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 suggested. On the one hand, a parametric approach modeling the relationship between dependent variables of both populations is presented and the EM algorithm is used for the parameters 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. 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: Mixture of regressions, switching regression, adaptive learning, EM algorithm, Bayesian inference, MCMC algorithm.

1 Introduction

The mixture of regressions, introduced by [10] as the switching regression model and also named clusterwise linear regression model in [13], is a popular regression model for modeling complex system. In particular, the switching regression model is often used in Economics for modeling phenomena with different phases. This model assumes that the dependent variable $Y \in \mathbb{R}$ can be linked to a covariate $x = (1, x_1, ..., x_p) \in \mathbb{R}^{p+1}$ by one of K possible regression models:

$$Y = x^t \beta_k + \sigma_k \varepsilon, \; k = 1, ..., K \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0,1)$, $\beta_k = (\beta_{k0}, ..., \beta_{kp}) \in \{\beta_1, \ldots, \beta_K\}$ is the regression parameter vector in $\mathbb{R}^{p+1}$ and $\sigma_k^2 \in \{\sigma_1^2, \ldots, \sigma_K^2\}$ is the residual variance. The conditional density distribution of Y given x is therefore:

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

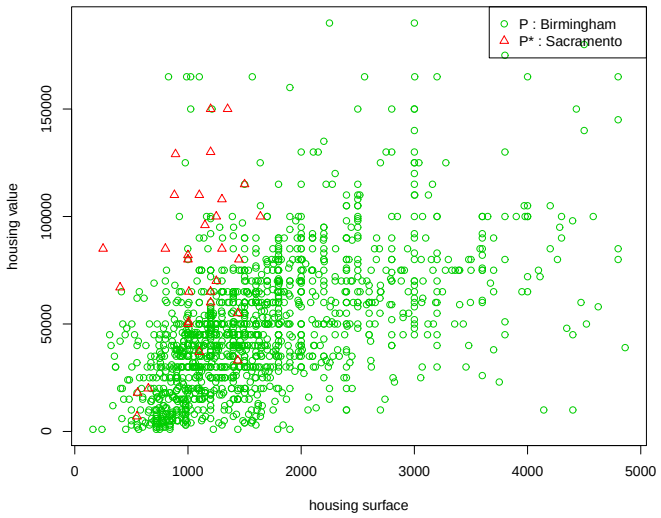

Figure 1: Housing value vs surface for Birmingham (AL, USA) and San Jose (CA, USA)

where $\pi_1, ..., \pi_K$ are the mixing proportions (with the classical constraint $\sum_{i=1}^{K} \pi_k = 1$), and $\phi(\cdot|x^t\beta_k, \sigma_k^2)$ is the Gaussian density parametrized by its mean $x^t\beta_k$ and variance σ_k^2. Among the works which focused on this model, we can emphasize the following ones which have contributed to the popularity of this model: [14] investigates the model identifiability, [15] proposes a Bayesian inference for the model estimation, [33] studies the asymptotic theory of parameter estimators in order to define hypothesis tests, and [17] considers variable selection for this specific regression model. Let us also mention that [18] presents 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, and we want to use it for predicting 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. Although very frequent in practical applications, this issue has unfortunately received very few attention in the literature. To our knowledge, only [5] has considered this situation in regression. In [5], the authors illustrate their adaptive regression model on a real-world situation: the prediction of 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 difference between these two cities is illustrated by Figure 1 which presents the value of the houses according to their surfaces. In this example, the difference between the training and the prediction populations is geographical, but it could also be temporal (as in Section 4) or due to species evolutions in a biological context (as in [4]).

1.1 Related works

As mentioned before, only few papers have investigated the original problem considered in this work in the context of mixture of regressions. In the machine learning community, a related problem in a non-mixture regression background, named *Covariate Shift*, considers that the probability density of the covariates in the new population is different from the one of the training population, but that the relationship between covariates and dependent variable has not changed. 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. Several recent works [25, 27, 28, 29, 30] have contributed to analyze this context. Furthermore the fact that we consider mixtures of regression and Covariate Shift does not, the focus of the present work is however more general and does not assume that the relationship between explanatory and response variables is conserved from the learning data to the new data. In addition, the situation under review in this paper considers that only few learning data are available for the new situation, which is not enough to correctly estimate in practice their probability distribution. In supervised classification, a similar problem was studied in [4] on quantitative variables and in [16] on binary variables. For this, the authors introduce model-based discriminant rules for classifying individuals from a prediction population which differs from the training one. The parsimony of these rules is obtained by considering families of linear links modeling the transformation between the reference population and the new one. An extension of this work to logistic regression was also proposed in [2]. In unsupervised classification, [20] recently proposes Gaussian models for simultaneous clustering on two different populations. Finally, some other applied works cover the problematic of knowledge transfer in specific industrial contexts. For instance, [9] gives a good overview of solutions for model transfer in the field of Chemometrics. Among the proposed transfer models, the most used models are the piecewise direct standardization [32] and the neural network based nonlinear transformation [11]. Several works [3, 31] have also considered this problem in the field of semiconductor industry.

1.2 Problem formulation

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

$$\begin{aligned} Y^* &= x^{*t}\beta_k^* + \sigma_k^*\varepsilon^* \\ p(y^*|x^*) &= \sum_{k=1}^{K^*} \pi_k^* \phi(y^*|x^{*t}\beta_k^*, {\sigma_k^*}^2) \end{aligned} \tag{3}$$

with $\varepsilon^* \sim \mathcal{N}(0,1)$, $\beta_k^* \in \{\beta_1^*, \ldots, \beta_{K^*}^*\}$ and $\sigma_k^* \in \{\sigma_1^*, \ldots, \sigma_{K^*}^*\}$. Let us now precise the focus of this paper by making the three following assumptions. Firstly, the variables (Y,x) and (Y^*,x^*) are assumed to be the same but measured on two different populations. Secondly, 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 reference population P. Otherwise, the mixture regression model could be estimated directly without using the training population. Thirdly, 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 Organisation 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 prediction population by defining parametric models for the link between mixture regression models of both populations. An alternative Bayesian approach is then presented in Section 3 in which the link between regression models is formulated through prior densities on the new population. In Section 4, the performance of both the parametric and the Bayesian approaches is first illustrated on simulations and the proposed strategies are then compared to classical methods on an environmental application.

2 Parametric approach for adaptive mixture of regressions

This section presents a parametric approach which consists in modeling 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:

$$Y^*_{|x,z^*_{ik}=1} \sim \psi_k(Y_{|x,z_{ik}=1}) \tag{4}$$

with ψ_k a function from $\mathbb{R}$ to $\mathbb{R}$. By only assuming that the function ψ_k is $\mathscr{C}^1$, [4] proves that ψ_k is necessarily affine:

$$Y^*_{|x,z^*_{ik}=1} \sim \lambda_{k1} + \lambda_{k2} Y_{|x,z_{ik}=1}$$

where $(\lambda_{k1}, \lambda_{k2}) \in \mathbb{R}^2$. We therefore obtain the following relationship between the model parameters of P and P^*:

$$\begin{aligned} \beta^*_k &= (\lambda_{k1} + \lambda_{k2}\beta_{k0}, \lambda_{k2}\beta_{k1}, \ldots, \lambda_{k2}\beta_{kp})^t, & (5)\\ \sigma^*_k &= \lambda_{k2}\sigma_k. & (6) \end{aligned}$$

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^*: using this link it decreases to $3K-1$ whereas for a complete mixture of regressions on P^* it is $(3+p)K-1$. This assumption is however relatively strong: if there is no real link between P and P^*, it will be no more possible to correctly estimate the mixture regression model (3) since relations (5) and (6) lead to loose all freedom on the parameters ($3K-1$ is strictly lower than $(3+p)K-1$). In order to introduce more flexibility, it is possible to introduce additional models for the link between populations by allowing the effects of different covariables on the dependent variable to be differently transformed from P to P^*. A second class of link models, including the first one, is then taken into consideration:

$$\begin{aligned} \beta^*_k &= \Lambda_k \beta_k, \text{ where } \Lambda_k = diag(\lambda_{k0}, \lambda_{k1}, \ldots, \lambda_{kp}) & (7)\\ \sigma^*_k &\quad \text{is free,} \end{aligned}$$

where $diag(\lambda_{k0}, \lambda_{k1}, \ldots, \lambda_{kp})$ is the diagonal matrix containing $(\lambda_{k0}, \lambda_{k1}, \ldots, \lambda_{kp})$ on its diagonal completed by zeros. In the following, some constraints on Λ_k will be introduced in order to define a family of parcimonious models:

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_1 assumes that both populations are the same population: $\Lambda_k = I_d$ is the identity matrix,
- M_2 assumes that the link between 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 assumes that the link between populations is covariate independent:
  - $M_{3a} : \lambda_{k0} = 1$, $\lambda_{kj} = \lambda_k$ and $\sigma_k^* = \lambda_k\sigma_k \quad \forall 1 \leq j \leq p$,
  - $M_{3b} : \lambda_{k0} = \lambda_k$, $\lambda_{kj} = 1$ 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 $\sigma_k^* = \lambda_{k1}\sigma_k \quad \forall 1 \leq j \leq p$,

  Note that M_{3d}, which is the most general model among M_2 and M_3 classes of models, is equivalent to the model defined by relations (5) and (6).
- M_4 assumes that the link between populations is mixture component independent:
  - $M_{4a} : \lambda_{k0} = 1$ 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 that Λ_k is unconstrained, which leads to estimate the mixture regression model for P^* by using only S^*.

Moreover, the mixing proportions are allowed to be the same in each population or to be different between both populations P and P^*. In the latter case, they consequently have to be estimated using the sample S^*. Corresponding notations for the models are respectively $M_\cdot$ and $pM_\cdot$. 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.

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) for the reference 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 (7)

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 [7]. This technique is certainly the most popular approach for inference in mixtures of regressions (see [18] for instance). Conditionally to a sample $S^* = (\mathbf{y}^*, \mathbf{x}^*)$ of observations, where $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$ and $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$, the log-likelihood of model (3) is given by:

$$L(\theta; \mathbf{y}^*, \mathbf{x}^*) = \sum_{i=1}^{n^*} \ln \left(\sum_{k=1}^{K} \pi_k^* \phi(y_i^* | x_i^t \Lambda_k \beta_k, \sigma_k^{*2}) \right), \tag{8}$$

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

$$L_c(\theta; \mathbf{y}^*, \mathbf{x}^*, \mathbf{z}^*) = \sum_{i=1}^{n^*} \sum_{k=1}^{K} z_{ik}^* \ln \left(\pi_k^* \phi(y_i^* | x_i^t \Lambda_k \beta_k, \sigma_k^{*2}) \right), \tag{9}$$

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 $\mathscr{M}(1, \pi_1^*, \ldots, \pi_K^*)$.

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 completed log-likelihood:

$$\begin{aligned} Q(\theta, \theta^{(q)}) &= E_{\theta^{(q)}}[L_c(\theta; \mathbf{y}^*, \mathbf{x}^*, \mathbf{z}^*) | \mathbf{y}^*, \mathbf{x}^*] \\ &= \sum_{i=1}^{n^*} \sum_{k=1}^{K} t_{ik}^{(q)} \left(\ln(\pi_k^*) + \ln(\phi(y_i^* | x_i^t \Lambda_k \beta_k, \sigma_k^{*2})) \right), \end{aligned} \tag{10}$$

where:

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

is the posterior probability that the observation i comes from the k-th mixture component.

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

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

where Θ is a parameter space depending on the model at hand. This maximization is now described for each component of $\theta = (\pi_k^*, \Lambda_k, \sigma_k^*)_{k=1,K}$. For the mixing proportions, the maximum is as usual reached for:

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

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

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

The reminder of this section details only the maximization according to the link parameters for the model M_{3d} and we refer to Appendix A for update formulae of models M_2 and M_4. As model M_{3d} considers two interdependent scalar parameters λ_{k0} and λ_{k1}, no analytical formulae are available for the global maximum on both λ_{k0} and λ_{k1}. In such a situation, an easy way to carry out the maximization in this case is to consider a descending algorithm in which λ_{k0} and λ_{k1} 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, [7]). Update formulas for these two parameters are consequently:

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

and

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

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

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 [12] for instance. The first considered criterion is the PRESS criterion [1], which represents the mean squared prediction error computed on a cross-validation scheme, formally defined by:

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

where $\hat{y^*}_i^{-i}$ is the prediction of y_i^* obtained by the mixture regression model estimated without using the ith 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, [24]), which is a penalized likelihood criterion which has a less computation cost. The BIC criterion is defined by:

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

where ℓ is the maximum log-likelihood value and ν 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 remark 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 modeling 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 [15, 23] 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 a real prior on the model parameters of population P^*. Indeed, even though training and prediction populations differ, they have a strong link and it is 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 $\beta^* = \{\beta_1^*, ...\beta_K^*\}$ and $\sigma^{*2} = \{\sigma_1^{*2}, ..., \sigma_K^{*2}\}$. The independence between $(\beta_k^*, \sigma_k^{*2})$ and $(\beta_\ell^*, \sigma_\ell^{*2})$ is as well assumed for all $k \neq \ell$, $k, \ell = 1, ..., K$. 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 $k = 1, ..., K$, 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 $k = 1, ..., K$, is assumed to be an inverse-gamma distribution:

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

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

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

With such a modelling, the regression coefficients β_k^* and the mixing proportions $\pi^* = \{\pi_1^*, ..., \pi_K^*\}$ 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, $k = 1, ..., K$, 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:

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

where $\theta^* = (\pi_k^*, \beta_k^*, \sigma_k^*)_{k=1,K}$. 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 ([8]). In Gibbs sampling, the vector parameter θ^* is partitioned into s groups of parameters $\{\theta_1^*,...,\theta_s^*\}$ 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:

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

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:

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

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 $k = 1,...,K$. With these notations, the conditional posterior distribution of σ_k^{*2} is an inverse gamma:

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

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:

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

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.

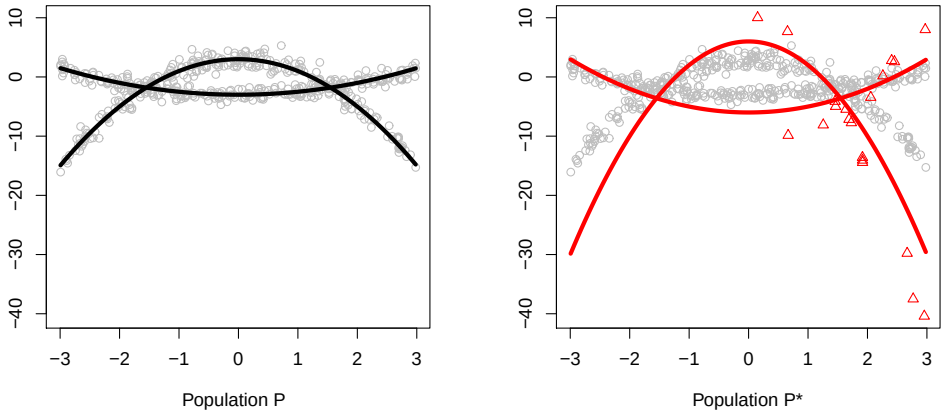

Figure 2: Populations P and P^* used for the introductory example. Curves black (left) and red (right) indicates respectively the actual mixture regression of populations P and P^*.

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 [22], consists in adding indenfiability constraints on model parameters such as an order relation in mixing proportions. Unfortunately, this approach does not work very well as showed by [6]. By contrast, some authors like Celeux *et al.* [6] and Stephens [26] 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 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 in a real situation.

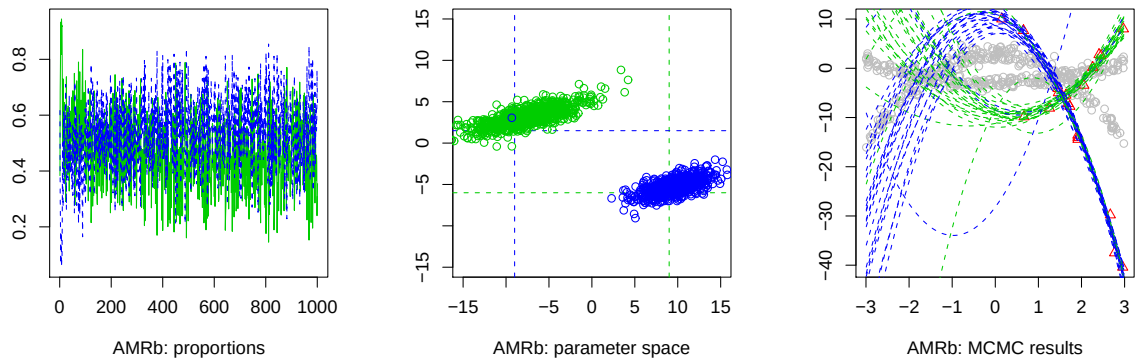

Figure 3: Results obtained for the introductory example with the Bayesian approach of adaptive mixture of regressions (AMRb). From left to right: mixing proportions over the MCMC iterations, Gibbs sampling in the parameter space and some of the generated regression curves. See text for details.

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 $\beta_2^* = (-9,0,1.5)$. 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 1000 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

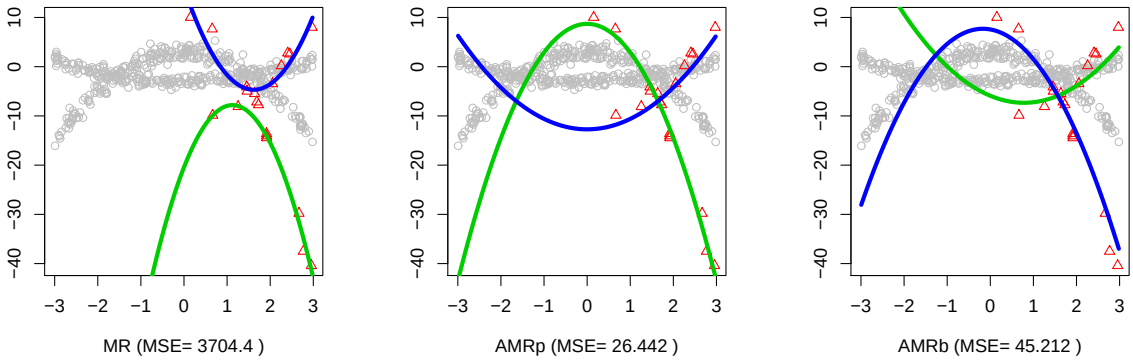

Figure 4: Results obtained for the introductory example with classical mixture of regressions (MR), parametric adaptive mixture of regressions (AMRp) and Bayesian adaptive mixture of regressions (AMRb) methods. See text for details.

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 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 a independent validation set, is consequently high (3704.4). Conversely, the parametric (with the most general 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 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 a 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 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 the size of S^*. Indeed, 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

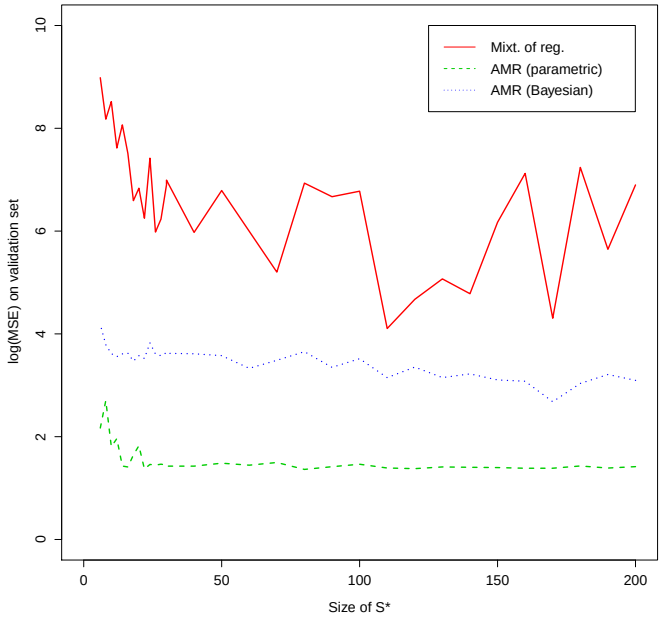

Figure 5: Average logarithm of the MSE value according to the the size of S^* for the classical mixture of regressions (MR), parametric adaptive mixture of regressions (AMRp) and Bayesian adaptive mixture of regressions (AMRb) methods.

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 performance and stable than the classical MR method, appears to be slightly less efficient than the parametric approach AMR. 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.

4.3 Real data study: 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. 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

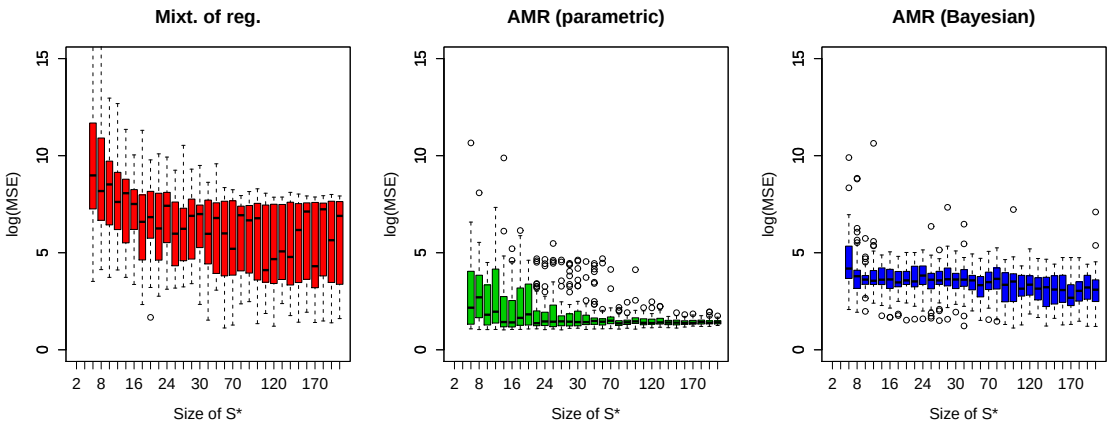

Figure 6: Boxplots of MSE values (on logarithmic scale) according to the the size of S^* for the classical mixture of regressions (left), parametric adaptive mixture of regressions (center) and Bayesian adaptive mixture of regressions (right) methods.

CO_2 emission, whereas the second group (about 75%) seems to have more environmental political orientations. As pointed out by [15], 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 eight AMRp models pM_{2a} to pM_{3d} (since pM_{4a} and pM_{4b} are equivalent to pM_{2a} and pM_{2c} for $p = 1$), 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 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, although 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 small. 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%

30% of the 1999's data ($n^* = 33$)

model	BIC	PRESS	MSE
AMRp (pM_{2a})	13.21	**4.01**	4.77
AMRp (pM_{2b})	12.89	4.57	**3.66**
AMRp (pM_{2c})	**12.57**	4.16	4.55
AMRp (pM_{2d})	17.13	4.38	4.77
AMRp (pM_{3a})	15.92	4.49	4.66
AMRp (pM_{3b})	16.01	5.59	4.11
AMRp (pM_{3c})	15.75	6.17	4.23
AMRp (pM_{3d})	22.72	4.49	4.66
AMRb	-	(†)	5.99
UR	27.08	7.46	7.66
MR	32.89	5.54	5.11

50% of the 1999's data ($n^* = 55$)

model	BIC	PRESS	MSE
AMRp (pM_{2a})	14.10	4.76	3.88
AMRp (pM_{2b})	**13.99**	**4.10**	**3.77**
AMRp (pM_{2c})	14.07	5.29	4.22
AMRp (pM_{2d})	17.82	4.45	4.66
AMRp (pM_{3a})	18.07	4.27	4.66
AMRp (pM_{3b})	18.00	5.62	4.44
AMRp (pM_{3c})	17.60	5.62	4.33
AMRp (pM_{3d})	26.61	6.12	4.55
AMRb	-	(†)	5.66
UR	20.87	7.95	7.21
MR	39.69	4.82	4.77

70% of the 1999's data ($n^* = 77$)

model	BIC	PRESS	MSE
AMRp (pM_{2a})	15.15	5.51	8.21
AMRp (pM_{2b})	14.82	**3.89**	3.77
AMRp (pM_{2c})	**14.71**	4.53	4.44
AMRp (pM_{2d})	19.00	5.83	4.99
AMRp (pM_{3a})	18.96	4.79	4.44
AMRp (pM_{3b})	19.06	4.34	4.22
AMRp (pM_{3c})	18.98	5.26	3.77
AMRp (pM_{3d})	27.57	5.55	4.88
AMRb	-	(†)	5.99
UR	22.08	8.00	7.10
MR	43.91	5.06	**3.33**

($n^* = 111$)

model	BIC	PRESS	MSE
AMRp (pM_{2a})	15.51	3.83	3.77
AMRp (pM_{2b})	15.54	3.87	4.77
AMRp (pM_{2c})	**15.34**	4.13	4.11
AMRp (pM_{2d})	20.14	4.41	4.33
AMRp (pM_{3a})	20.19	4.48	4.77
AMRp (pM_{3b})	20.03	4.41	4.33
AMRp (pM_{3c})	20.06	4.35	3.44
AMRp (pM_{3d})	29.55	4.76	5.44
AMRb	-	(†)	5.66
UR	23.62	7.53	6.99
MR	47.19	**3.66**	**2.89**

Table 2: MSE on the whole 1999's sample, PRESS and BIC criterion for the 8 parametric adaptive mixture models (AMRp pM_{2a} to pM_{3d}), 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.

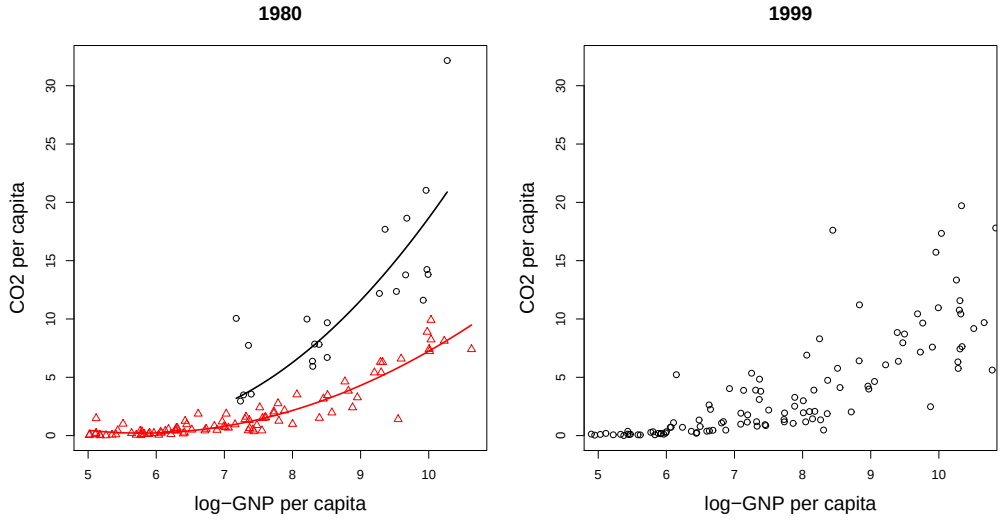

Figure 7: Emission of CO_2 per capita *versus* GNP per capita in 1980 (left) and 1999 (right).

and 70% of the whole sample. Furthermore, it should be noticed that ARMp 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 the other hand, the Bayesian approach AMRb appears to provide results as stable as the ones of AMRp but slightly less satisfying and this confirms the conclusions of the previous experiment on simulations.

This application illustrates well the interest of combining informations on both past (1980) and present (1999) situations in order to analyse the link between CO_2 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, contrary to what one might *prima facie* have thought, that all the countries have made an effort to reduce their CO_2 emissions and not only those which had the higher ones.

5 Conclusion

We propose 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 adopt 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 new population required for the estimation. As this 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. More-

over, as it has been showed in the real application, the competition between the parametric link models provides informations on the link between populations, which can be meaningful for the practician.
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 significantly 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 [21]. 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 [19] 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.

Acknowledgment

The authors would like to thank Christophe Biernacki, Gilles Celeux and Stéphane Girard for useful discussions and comments about this work.

Appendix

A Link parameters maximization for the M step of the EM algorithm

The maximums in the M step of the EM algorithm are

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

For the models pM_{3a}, pM_{3b} and pM_{3c} the formulas are the same by omitting the sum under k. For the model pM_{2d}, the maximization is done on the same scheme as in Section 2.2, with the following update formulas:

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

and

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

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 the link parameters. For these latter, another descending algorithm maximizing at each step a scalar parameter according to the current value of the others has to be used as well. Update formulas are the following:

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

$$\lambda_J^{(q+1)} = \left(\sum_{i=1}^{n} \sum_{k=1}^{K} \frac{t_{ik}^{(q)}}{\sigma_k^{*2}} x_{i\sim 0}^{*t} \beta_{k\sim 0} x_{iJ}^* \beta_{kJ} \right)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{t_{ik}^{(q)}}{\sigma_k^{*2}} x_{i\sim 1}^{*t} \beta_{k\sim 1} \left(y_i^* - \beta_{k0} - \sum_{j=1, j\neq J}^{p} \lambda_j^{(q+1)} x_{ij}^* \beta_{kj} \right),$$

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

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

References

[1] David M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

[2] F. Beninel and C. Biernacki. Modèles d'extension de la régression logistique. *Revue des Nouvelles Technologies de l'Information, Data Mining et apprentissage statistique : application en assurance, banque et marketing*, (A1):207–218, 2007.

[3] K. Bertness, R. Hickernell, S. Hays, and D. Christensen. Nose reduction in optical in situ measurements for molecular bean epitaxy by substrate woble normalization. *Journal of Vacuum Science and Technology B*, 16(3):1492–1497, 1998.

[4] C. Biernacki, F. Beninel, and V. Bretagnolle. A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics*, 58(2):387–397, 2002.

[5] C. Bouveyron and J. Jacques. Adaptive linear models for regression: improving prediction when population has changed. *Pattern Recognition Letters*, in press, 2010.

[6] G. Celeux, M. Hurn, and C. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95:957–970, 2000.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data (with discussion). *Journal of the Royal Statistical Society. Series B*, 39:1–38, 1977.

[8] J. Diebolt and C. Robert. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B*, 56(2):363–375, 1994.

[9] N. Feudale, N. Woody, H. Tan, D. Kell, J. Maddock, Heginbothom, and J. M., Magee. Transfer of multivariate calibration models: a review. *Chemometrics and Intelligent Laboratory System*, 64:181–192, 2002.

[10] M. Goldfeld and R.E. Quandt. A markov model for switching regressions. *Journal of Econometrics*, 1:3–16, 1973.

[11] R. Goodacre, E. Timmins, A. Jones, D. Kell, J. Maddock, Heginbothom M., and J. Magee. On mass spectrometer instrument standardization and interlaboratory calibration transfer using neural networks. *Analytica Chimica Acta*, 348(1):511–532, 1997.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[13] C. Henning. *Classification in the Information Age*. Springer-Verlag, Heidelberg, 1999.

[14] C. Henning. Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17:273–296, 2000.

[15] Merrilee Hurn, Ana Justel, and Christian P. Robert. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79, 2003.

[16] J. Jacques and C. Biernacki. Extension of model-based classification for binary data when training and test populations differ. *Journal of Applied Statistics*, 37(5):749–766, 2010.

[17] Abbas Khalili and Jiahua Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479):1025–1038, 2007.

[18] F. Leisch. Flexmix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8):3–16, 2004.

[19] A. Lourme and C. Biernacki. 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*, 2008.

[20] A. Lourme and C. Biernacki. Classification simultanée à base de mélanges gaussiens pour des échantillons d'origines multiples. In *41èmes Journées de Statistique de la SFdS*, 2009.

[21] Adrian E. Raftery, Michael A. Newton, Jaya M. Satagopan, and Pavel N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In *Bayesian statistics 8*, Oxford Sci. Publ., pages 371–416. Oxford Univ. Press, Oxford, 2007.

[22] S. Richardson and P. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Serie B*, 59:731–792, 1997.

[23] C. Robert. *The Bayesian Choice*. Springer, 2007.

[24] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[25] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[26] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Serie B*, 62(4):795–809, 2000.

[27] A. Storkey and M. Sugiyama. *Mixture regression for covariate shift*, pages 1337–1344. Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, 2007.

[28] M. Sugiyama. Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, 7:141–166, 2006.

[29] M. Sugiyama and K-R. Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23, 2005.

[30] M. Sugiyama and Krauledat M. Müller, K-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

[31] K. Tobin, T. Karnowski, L. Arrowood, R. Ferrell, J. Goddard, and F. Lakhani. Content-based image retrieval for semiconductor process characterization. *Journal on Applied Signal Processing*, 1:704–713, 2002.

[32] Y. Wang, D. Veltkamp, and B. Kowalski. Multivariate instrument standardization. *Analytical chemistry*, 63(23):2750–2756, 1991.

[33] Hong-Tu Zhu and Heping Zhang. Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society. Series B.*, 66(1):3–16, 2004.