Adaptive Linear Models for Regression

Charles Bouveyron, Julien Jacques

Technical Report SAMOS-MATISSE, CES, UMR CNRS 8174

ADAPTIVE LINEAR MODELS FOR REGRESSION

By Charles Bouveyron and Julien Jacques

University Paris I (Panthéon-Sorbonne) and University Lille I

The general setting of regression analysis is to identify a relationship between a response variable Y and one or several explanatory variables $\mathbf{X}$ by using a learning sample. In a prediction framework, the main assumption for predicting Y on a new sample of $\mathbf{X}$ observations is that the regression model $Y = f(\mathbf{X}) + \epsilon$ is still valid. Unfortunately, this assumption is not always true in practice and the model could have changed. We therefore propose to adapt the original regression model to the new sample by estimating a transformation between the original regression function $f(\mathbf{X})$ and the new one $f^*(\mathbf{X})$. The main interest of this work is that a model for the new population can be build with only few observations. This is illustrated by applications on artificial and real datasets, including the modelling of the housing market in different U.S. cities in which the regression model of a reference city is adapted to another city. A package for the R software dedicated to adaptive linear models is available on the author's webpage.

1. Introduction. The general setting of regression analysis is to identify a relationship (the regression model) between a response variable and one or several explanatory variables. Most of the works in regression analysis is focused on the nature of the regression model: linear model (16) and generalized linear model (13) which can be seen as parametric models, and non linear models which are mostly non-parametric models (14). See (7) for a general survey on regression analysis.

1.1. *The problem of adapting a knowledge to a new situation.* In this paper an alternative evolution is considered: how to adapt an existing regression model to a new situation, in which the variables are identical (with a possible different probability density distribution) but where the relationship between response and explanatory variables could have changed. Let consider the following example: a real-estate agency of the US East coast has to its disposal, through their long experience in this area, a regression model of the housing price versus numerous housing descriptive variables. Let assume that this company plans to conquer new markets on the West coast.

AMS 2000 subject classifications: 62J05

Keywords and phrases: regression, adaptive estimation, linear transformation models, knowledge transfer, housing market in different U.S. cities

The link between housing descriptive variables and housing price is probably not the same for the West and East coasts, but it is also not probably completely different. In this paper, we propose a way to transfer a knowledge on a reference population to a new population through its regression model. Therefore, it will be possible regarding the previous example to use the East coast experience to define cheaply a new regression model for the West coast market. The major challenge of this work consists in defining a link between both populations and in deducing a link between the associated regression models.

1.2. *Related works.* To our knowledge, only few scientists have dealt with this original problem altough it could be very interesting in practical application. In the machine learning community, a related problem is investigated under the keyword *Covariate Shift*. The covariate shift problem considers that the probability density for the new data is different from the one of the learning data and the regression model is assumed to be conserved. 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 rarely the case in practice and the regression model estimated with the learning data could be very disappointing applied to data with a different probability distribution. A lot of recent works have contributed to analyze this context (17–21), but most of them need to know (at least an estimation of) the probability distribution of the new data, which is in practice a very difficult problem. The focus of the present work wants to be more general by not assuming 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 there are only few available data for the new situation, which is not enough to correctly estimate their probability distribution. In supervised classification, a very similar problem was studied in (5) on quantitative variables and in (11) in the case of binary variables. The authors introduce a model-based discriminant rule for classifying individuals from a population which differs from the learning one. For this, they introduce a family of linear models modelling the transformation between the reference population and the new one. An extension of this work to logistic regression was recently proposed in (3). Finally, some works cover the problematic of knowledge transfer in specific industrial contexts. In the field of Chemometrics, (8) gives a good overview of solutions for model transfer specially developed for this application. Among the proposed transfer models, the most used are the piecewise direct standardization (24) and the neural network based nonlinear trans-

formation (9). Several works (4; 22) have also considered this problem in the field of semiconductor industry.

The present paper is organized as follows. Section 2 formulates the problem of adapting an existing regression model to a new population and Section 3 introduces a family of transformation models to solve this problem. Estimation and model selection procedures are proposed in Section 4. Finally, Section 5 provides a simulation study in a spline regression context and two real applications in biological and economical fields.

2. Problem formulation. In this section, after having reminded the general framework of regression analysis, the problem of adapting an existing regression model to another population is formulated.

2.1. *Linear models for regression.* In regression analysis, the data $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, arising from a population P, are assumed to be independent and identically distributed samples of a couple of variables $(\mathbf{X}, Y)$ with an unknown distribution. The observations $\mathbf{x}_i$ are values of the deterministic explanatory variable $\mathbf{X} = (X^{(1)}, \ldots, X^{(p)})^t \in \mathbb{R}^p$ and the corresponding y_i are realizations of the stochastic variable $Y \in \mathbb{R}$. A general data modelling problem consists in identifying the relationship between the explanatory variable $\mathbf{X}$ (or covariate) and the response variable Y (or dependent variable). Both standard parametric and non-parametric regression approaches consider with the following model:

$$Y = f(\mathbf{X}, \boldsymbol{\beta}) + \epsilon, \tag{2.1}$$

with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and where $\boldsymbol{\beta}$ is the vector of regression parameters. This model is equivalent to the distributional assumption that:

$$Y|\mathbf{X} \sim \mathcal{N}(f(\mathbf{X}, \boldsymbol{\beta}), \sigma^2),$$

where the regression function $f(\mathbf{x}, \boldsymbol{\beta})$ is defined as the conditional expectation $E[Y|\mathbf{X} = \mathbf{x}]$. Therefore, the only way to link the response variable Y and the covariate $\mathbf{X}$ is through the assumption on $f(\mathbf{x}, \boldsymbol{\beta})$. In particular, parametric regression achieves this connection by assuming a specific form for $f(\mathbf{x}, \boldsymbol{\beta})$. The most common model is the linear form (6):

$$f(\mathbf{x}, \boldsymbol{\beta}) = \sum_{i=0}^{d} \beta_i \psi_i(\mathbf{x}), \tag{2.2}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^t \in \mathbb{R}^{d+1}$ are the regression parameters, $\psi_0(\mathbf{x}) = 1$ and $(\psi_i)_{1 \leq i \leq d}$ is a basis of regression functions:

$$\psi_i : \mathbb{R}^p \to \mathbb{R},$$

which can be for instance identity, polynomial, splines functions (10) or wavelets (12). We refer to (6) for a general survey. Let notice that the usual linear regression occurs when $d = p$ and $\psi_i(\mathbf{x}) = x^{(i)}$ for $i = 1, \ldots, d$. The regression function (2.2) can be written in its matricial form as follows:

$$f(\mathbf{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^t \Psi(\mathbf{x}), \tag{2.3}$$

where $\Psi(\mathbf{x}) = (1, \psi_1(\mathbf{x}), \ldots, \psi_d(\mathbf{x}))^t$.

2.2. *How to adapt a regression model to another population.* Let assume that the regression function f has been estimated in a preliminary study by using the sample S, and that a new regression model has to be adjusted on a new sample $S^* = \{(\mathbf{x}_1^*, y_1^*), ..., (\mathbf{x}_{n^*}^*, y_{n^*}^*)\}$, measured on the same explanatory variables but arising from another population P^* (n^* is usually assumed to be small). The difference between P and P^* can be geographical (as in the introduction example), temporal or other but the nature of both populations have to be similar. The new regression model on P^* can be written:

$$Y|\mathbf{X}^* \sim \mathcal{N}(f(\mathbf{X}^*, \boldsymbol{\beta}^*), \sigma^2), \tag{2.4}$$

with

$$f(\mathbf{x}^*, \boldsymbol{\beta}^*) = \sum_{i=0}^{d} \beta_i^* \psi_i(\mathbf{x}^*).$$

For modelling the link between P and P^*, the following transformation model between both regression functions is assumed:

$$f(\mathbf{x}^*, \boldsymbol{\beta}^*) = \phi(f(\mathbf{x}, \boldsymbol{\beta})), \tag{2.5}$$

where ϕ is a transformation function.

3. Adaptive linear models for regression. In this section, a family of transformations is introduced to solve the problem of adapting an existing regression model on a reference population P to a new population P^*.

3.1. *Assumptions on the transformation model.* Since the transformation model (2.5) is very general, it is necessary to make additional assumptions on the model to be able to characterize it. Therefore, we propose to assume that the transformation function ϕ has the following form:

$$\phi(f(\mathbf{x}, \boldsymbol{\beta})) = f(\mathbf{x}, \boldsymbol{\Lambda}\boldsymbol{\beta}) \tag{3.1}$$

where $\boldsymbol{\Lambda}$ is a $(d+1) \times (d+1)$ transformation matrix. By postulating that the regression functions ψ_i are the same for both regression models, which is natural since the variables are identical in both populations, this transformation is equivalent to the stochastic transformation between the expectations of Y conditionally to $\mathbf{X}$ and $\mathbf{X}^*$:

$$E[Y|\mathbf{X}^*] \sim \boldsymbol{\Lambda} E[Y|\mathbf{X}].$$

Given that the number of free parameters to estimate in the transformation matrix $\boldsymbol{\Lambda}$ being $(d+1)\times(d+1)$ and that the one for learning a new regression model directly from the sample S^* is $(d+1)$, the transformation model (3.1) is consequently highly over-parameterized. It is therefore necessary to introduce some constraints on the transformation model so that the number of free parameters to estimate is lower or equal to d.

3.2. *A family of transformation models.* A large class of model could be considered since the number of free parameters in the matrix $\boldsymbol{\Lambda}$ is lower or equal to d. First of all, it is assumed that the relation between the response variable and one given covariate in the new population P^* only depends on the relation between the response variable and the same covariate in the population P. Consequently, the regression parameter β_i^* only depends on the regression parameter β_i and the matrix $\boldsymbol{\Lambda}$ is diagonal. The transformation can be written in term of the regression parameters of both models as follows:

$$\beta_i^* = \lambda_i \beta_i \qquad \forall i = 1, \ldots, d, \tag{3.2}$$

where $\lambda_i \in \mathbb{R}$ is the i-th diagonal element of $\boldsymbol{\Lambda}$.

3.2.1. *Main transformation models.* We propose in this paper a family of 7 transformation models, named further adaptive linear models, ranging from the most complex model (hereafter $M0$) to the simplest one (hereafter $M6$):

- Model $M0$: $\beta_0^* = \lambda_0 \beta_0$ and $\beta_i^* = \lambda_i \beta_i$, for $i = 1, ..., d$. This model is the most complex model of transformation between both populations P and P^* and is equivalent to learning a new regression model from the sample S^*.

Model	M_0	M_1	M_2	M_3	M_4	M_5	M_6
β_0^* is assumed to be	$\lambda_0\beta_0$	β_0	$\lambda_0\beta_0$	$\lambda\beta_0$	β_0	$\lambda_0\beta_0$	β_0
β_i^* is assumed to be	$\lambda_i\beta_i$	$\lambda_i\beta_i$	$\lambda\beta_i$	$\lambda\beta_i$	$\lambda\beta_i$	β_i	β_i
Nb. of parameters	d+1	d	2	1	1	1	0

TABLE 1
Complexity (number of parameters) of the transformation models.

- Model $M1$: $\beta_0^* = \beta_0$ and $\beta_i^* = \lambda_i\beta_i$ for $i = 1, ..., d$. This model assumes that both regression models have the same intercept β_0.
- Model $M2$: $\beta_0^* = \lambda_0\beta_0$ and $\beta_i^* = \lambda\beta_i$ for $i = 1, ..., d$. This model assumes that the intercept of both regression models differ by the scalar λ_0 and all the other regression parameters differ by the same scalar λ.
- Model $M3$: $\beta_0^* = \lambda\beta_0$ and $\beta_i^* = \lambda\beta_i$ for $i = 1, ..., d$. This model assumes that all the regression parameters of both regression models differ by the same scalar λ.
- Model $M4$: $\beta_0^* = \beta_0$ and $\beta_i^* = \lambda\beta_i$ for $i = 1, ..., d$. This model assumes that both regression models have the same intercept β_0 and all the other regression parameters differ by the same scalar λ.
- Model $M5$: $\beta_0^* = \lambda_0\beta_0$ and $\beta_i^* = \beta_i$ for $i = 1, ..., d$. This model assumes that both regression models have the same parameters except the intercept.
- Model $M6$: $\beta_0^* = \beta_0$ and $\beta_i^* = \beta_i$ for $i = 1, ..., d$. This model assumes that both populations P and P^* have the same behaviour.

The numbers of parameters to estimate for these transformation models are presented in Table 1.

3.2.2. *Specific transformation models.* In some specific cases, the practician could have to use more specific transformation models suggested by some prior informations on the covariates. It is possible in such a context to consider intermediate transformation models by imposing specific constraints on parameters λ_i for given $i \in \{1, \ldots, d\}$. Thus, according to its experimental knowledge, the practician could assume that such explanatory variables have the same effect in the two regression models, and such other have not. For instance, it possible to consider the specific transformation matrix $\mathbf{\Lambda} = diag(\lambda_0, \lambda_1, \lambda, \ldots, \lambda)$ where $diag(\lambda_0, \lambda_1, \lambda, \ldots, \lambda)$ is the $(d+1) \times (d+1)$ diagonal matrix having $\{\lambda_0, \lambda_1, \lambda, \ldots, \lambda\}$ on its diagonal. This model assumes that the regression parameters β_i, $i = 2, \ldots, d$ are transformed in the same manner whereas the intercept and β_1 are not.

4. Estimation procedure and model selection. The estimation procedure associated with the adaptive linear models is made of two main steps corresponding to the estimation of the regression parameters on the population P and to the estimation of the transformation parameters using samples of the population P^*. Then, the regression parameters of P^* are obtained by plug-in. The maximum likelihood estimation method is retained.

4.1. *Estimation of the regression parameters.* Consider a data set of inputs $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with corresponding response values grouping in a column vector $\mathbf{y} = (y_1, \ldots, y_n)^t$. Under the assumptions of the model (2.2) the log-likelihood of $\mathbf{y}$ given the data $\mathbf{x}$ and the parameters $\boldsymbol{\beta}$ and σ^2 is:

$$\ln l(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}, \sigma^2) = -n \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}^t \Psi(\mathbf{x}_i^*)\right)^2. \tag{4.1}$$

Maximizing this log-likelihood according to $\boldsymbol{\beta}$ is equivalent to maximizing $\sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}^t \Psi(\mathbf{x}_i^*)\right)^2$ and thus the maximum likelihood estimator is equivalent to the least square estimator. The gradient of the log-likelihood function takes the following form:

$$\nabla \ln l(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}^t \Psi(\mathbf{x}_i^*)\right) \Psi(\mathbf{x}_i^*)^t,$$

and setting this gradient to zero gives:

$$\sum_{i=1}^{n} y_i \Psi(\mathbf{x}_i^*)^t = \boldsymbol{\beta}^t \left(\sum_{i=1}^{n} \Psi(\mathbf{x}_i^*) \Psi(\mathbf{x}_i^*)^t \right).$$

Solving this equation according to $\boldsymbol{\beta}$ leads to the well known ordinary least square (OLS) estimator (6) for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{\Psi}^t \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^t \mathbf{y},$$

where $\boldsymbol{\Psi}$ is a $(n) \times (d+1)$ matrix formed by the row vector $\Psi(\mathbf{x}_i)^t$ $(1 \leq i \leq n)$.

4.2. *Estimation of the transformation parameters.* As previously noticed, the full model $M0$ corresponds to a completely new regression model adjusted on the sample S^* and does not need the estimation of transformation parameters. Similarly, the model $M6$, which considers no transformation between P and P^*, does not require the estimation of transformation parameters. Consider now a sample $\mathbf{x}^* = \{\mathbf{x}_1^*, \ldots, \mathbf{x}_{n^*}^*\}$ drawn from P^* with

corresponding response values $\mathbf{y}^* = (y_1^*, \ldots, y_{n^*}^*)^t$. By replacing $\boldsymbol{\beta}^* = \boldsymbol{\Lambda}\boldsymbol{\beta}$ in (4.1), the log-likelihood is:

$$\ln l(\mathbf{y}^*; \mathbf{x}^*, \boldsymbol{\Lambda}, \sigma^2) = -n^* \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n^*} \left(y_i^* - \boldsymbol{\beta}^t \boldsymbol{\Lambda}^t \Psi(\mathbf{x}_i^*)\right)^2. \tag{4.2}$$

This log-likelihood must be maximized according to the transformation matrix $\boldsymbol{\Lambda} = diag(\lambda_0, \lambda_1, \ldots, \lambda_d)$.

Model M_1. As the transformation matrix is $\boldsymbol{\Lambda} = diag(1, \lambda_1, \ldots, \lambda_d)$, the log-likelihood (4.2) can be written

$$\ln l(\mathbf{y}^*; \mathbf{x}^*, \boldsymbol{\Lambda}, \sigma^2) = -n^* \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n^*} \left(y_i^* - \beta_0 - \boldsymbol{\beta}_{\sim 1}^t \boldsymbol{\Lambda}_{\sim 1}^t \Psi_{\sim 1}(\mathbf{x}_i^*)\right)^2$$

where $\boldsymbol{\Lambda}_{\sim k}$ and $\boldsymbol{\beta}_{\sim k}$ correspond respectively to $\boldsymbol{\Lambda}$ and $\boldsymbol{\beta}$ without the k-th row. This maximization is therefore similar to the maximization of (4.1) and leads to the following estimator of $\boldsymbol{\Lambda}_{\sim 1} = diag(\lambda_1, \ldots, \lambda_d)$:

$$\hat{\boldsymbol{\Lambda}}_{\sim 1} = (\boldsymbol{\Psi}_{\sim 1}^{*t} \boldsymbol{\beta}_{\sim 1} \boldsymbol{\beta}_{\sim 1}^t \boldsymbol{\Psi}_{\sim 1}^*)^{-1} \boldsymbol{\beta}_{\sim 1}^t \boldsymbol{\Psi}_{\sim 1}^* (\mathbf{y}^* - \beta_0)$$

where $\boldsymbol{\Psi}^*$ is a $(n^*) \times (d+1)$ matrix formed by the row vector $\Psi(\mathbf{x}_i^*)^t$ $(1 \leq i \leq n^*)$.

Model M_2. The transformation matrix has in this case the form $\boldsymbol{\Lambda} = diag(\lambda_0, \lambda, \ldots, \lambda)$. The maximization according to $\boldsymbol{\Lambda}$ of the following log-likelihood:

$$\ln l(\mathbf{y}^*; \mathbf{x}^*, \boldsymbol{\Lambda}, \sigma^2) = -n^* \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n^*} \left(y_i^* - \beta_0\lambda_0 - \boldsymbol{\beta}_{\sim 1}^t \boldsymbol{\Lambda}_{\sim 1}^t \Psi_{\sim 1}(\mathbf{x}_i^*)\right)^2$$

leads to the estimator of $\boldsymbol{\Lambda}_{M_2} = (\lambda_0, \lambda)^t$:

$$\hat{\boldsymbol{\Lambda}}_{M_2} = (\boldsymbol{Q}^t \boldsymbol{Q})^{-1} \boldsymbol{Q}^t \mathbf{y}^*,$$

where

$$\boldsymbol{Q} = \begin{pmatrix} \beta_0 & \sum_{i=1}^d \beta_i \psi_i(\mathbf{x}_1^*) \\ & \vdots \\ \beta_0 & \sum_{i=1}^d \beta_i \psi_i(\mathbf{x}_n^*) \end{pmatrix}.$$

Model M_3. For this model, the transformation matrix is formed by only one real parameter and $\boldsymbol{\Lambda} = diag(\lambda, \lambda, \ldots, \lambda)$. The maximization of the loglikelihood according to λ leads to the following estimator:

$$\hat{\lambda} = (\boldsymbol{\Psi}^{*t} \boldsymbol{\beta} \boldsymbol{\beta}^t \boldsymbol{\Psi}^*)^{-1} \boldsymbol{\beta}^t \boldsymbol{\Psi}^* \mathbf{y}^*.$$

Model M_4. In this case, the transformation matrix is formed by a constant and a unique transformation parameter λ. The transformation matrix has therefore the form $\mathbf{\Lambda} = diag(1, \lambda, \ldots, \lambda)$ and the corresponding estimator of λ is:

$$\hat{\lambda} = (\mathbf{\Psi}^{*t}_{\sim 1}\boldsymbol{\beta}_{\sim 1}\boldsymbol{\beta}^t_{\sim 1}\mathbf{\Psi}^*_{\sim 1})^{-1}\boldsymbol{\beta}^t_{\sim 1}\mathbf{\Psi}^*_{\sim 1}(\mathbf{y}^* - \beta_0).$$

Model M_5. For this model, the transformation matrix is $\mathbf{\Lambda} = diag(\lambda_0, 1, \ldots, 1)$ and the estimator of λ_0 is:

$$\hat{\lambda}_0 = \frac{1}{n^*\beta_0}\sum_{i=1}^{n^*}[y_i^* - \sum_{j=1}^{d}\beta_j\psi_j(\mathbf{x}_i^*)].$$

Specific transformation models. As said before, in some specific cases the practician could have to use more specific transformation models suggested by some prior informations. A generic transformation model including all possible specific transformation models and the corresponding estimator is described below. In the sequel, the indexes γ_j will be associated to regression parameters of the new population to estimate using the relation $\beta^*_{\gamma_j} = \lambda_{\gamma_j}\beta_{\gamma_j}$ with $j = 1, ..., q$ and $\gamma_j \in \{0, ..., d\}$. In the same manner, the indexes $\bar{\gamma}_j$ will be associated to regression parameters of the new population which are similar to the parameters of the original population, *i.e.* $\beta^*_{\bar{\gamma}_j} = \beta_{\bar{\gamma}_j}$ with $j = 1, ..., p-q$ and $\bar{\gamma}_j \in 0, ..., d$. The regression model for the new population can be written as follows:

$$Y = \boldsymbol{Q}\mathbf{\Lambda}_q + \bar{\boldsymbol{Q}} + \epsilon,$$

where:

- $\mathbf{\Lambda}_q = (\lambda_{\gamma_1}, \ldots, \lambda_{\gamma_q})^t$,
- $\boldsymbol{Q} = \begin{pmatrix} \beta_{\gamma_1}\psi_{\gamma_1}(x_1) & \cdots & \beta_{\gamma_q}\psi_{\gamma_q}(x_1) \\ \vdots & & \vdots \\ \beta_{\gamma_1}\psi_{\gamma_1}(x_n) & \cdots & \beta_{\gamma_q}\psi_{\gamma_q}(x_n) \end{pmatrix}$,
- $\bar{\boldsymbol{Q}} = \begin{pmatrix} \beta_{\bar{\gamma}_1}\psi_{\bar{\gamma}_1}(x_1) & \cdots & \beta_{\bar{\gamma}_q}\psi_{\bar{\gamma}_q}(x_1) \\ \vdots & & \vdots \\ \beta_{\bar{\gamma}_1}\psi_{\bar{\gamma}_1}(x_n) & \cdots & \beta_{\bar{\gamma}_q}\psi_{\bar{\gamma}_q}(x_n) \end{pmatrix}$.

Consequently the maximum likelihood estimator of Λ_q is

$$\hat{\mathbf{\Lambda}}_q = \left(\boldsymbol{Q}^t\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}^t\left(\mathbf{y} - \bar{\boldsymbol{Q}}\right).$$

4.3. *Joint estimation.* In this work, a reference regression model on the population P is assumed to be well known and is transformed in a new regression model adapted to a new population P^* by estimating a transformation between both reference and new populations. However the regression parameters of the reference model are in practice never known but only estimated from a given sample S. Therefore, starting from this estimation to estimate the new regression model could be disappointing in some cases, particularly when the size n of S is not very large. As both populations P and P^* are assumed to be linked, it could be interesting to use both samples S and S^* for improving the estimation of the regression parameter $\boldsymbol{\beta}$ as well. An alternative algorithm is proposed here to jointly estimate the regression parameter $\boldsymbol{\beta}$ and the transformation matrix $\boldsymbol{\Lambda}$. The joint estimation procedure alternates the two following steps:

1. Estimate $\boldsymbol{\Lambda}$ given a current value of $\boldsymbol{\beta}$ (this step was the purpose of the previous section),
2. Estimate $\boldsymbol{\beta}$ given the estimation of $\boldsymbol{\Lambda}$ obtained in the previous step.

For a given estimation $\hat{\boldsymbol{\Lambda}}$ of $\boldsymbol{\Lambda} = diag(\lambda_0, \lambda_1, \ldots, \lambda_d)$, the estimation of $\boldsymbol{\beta}$ consists in maximizing the log-likelihood of the considered regression model (2.2) for the sample S and the log-likelihood of the same model in which the regression function ψ_i are multiplied by λ_i for the sample S^*. By introducing $\tilde{\mathbf{y}} = (y_1, \ldots, y_n, y_1^*, \ldots, y_{n^*}^*)^t$ and $\tilde{\boldsymbol{\Psi}}$ the $(n+n^*)\times(d+1)$ matrix defined as follows:

$$\tilde{\boldsymbol{\Psi}} = \begin{pmatrix} \psi_0(\mathbf{x}_1) & \cdots & \psi_d(\mathbf{x}_1) \\ \vdots & & \vdots \\ \psi_0(\mathbf{x}_n) & \cdots & \psi_d(\mathbf{x}_n) \\ \hat{\lambda}_0\psi_0(\mathbf{x}_1^*) & \cdots & \hat{\lambda}_d\psi_d(\mathbf{x}_1^*) \\ \vdots & & \vdots \\ \hat{\lambda}_0\psi_0(\mathbf{x}_n^*) & \cdots & \hat{\lambda}_d\psi_d(\mathbf{x}_n^*) \end{pmatrix},$$

the estimator of $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\Lambda}}$ is:

$$\hat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\Psi}}^t\tilde{\boldsymbol{\Psi}})^{-1}\tilde{\boldsymbol{\Psi}}^t\tilde{\mathbf{y}}.$$

4.4. *Model selection.* In order to select the most appropriate model of transformation between the populations P and P^*, we propose to use three well known criterions. The reader interested in a comparison of the respective performances of these three criterions could refer for instance to (10). The first criterion is the PRESS criterion (2), which represents the sum of squared

prediction errors computed on a cross-validation scheme, and is defined by:

$$PRESS = \sum_{j=1}^{n} ||y^*_{(j)} - \hat{y}^*_{(j)}||_2$$

where $y^*_{(j)}$ is the vector y^* without the j-th individual and $\hat{y}^*_{(j)}$ is the prediction of $y^*_{(j)}$ obtained by the regression model in which the parameters are estimated without using the j-th individual of the sample S^*. This criterion is one of the most often used for model selection in regression analysis. The two following criterions are penalized likelihood criterions. They both consist in selecting the models leading to the highest likelihood but penalizing those which have a large number of parameters. The Bayesian Information Criterion (BIC) (15) is defined by:

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

where ℓ is the maximum likelihood value and ν is the number of estimated parameters (see Table 1). With the same notations, the Akaike Information Criterion (AIC) (1) is defined by:

$$AIC = -2 \ln \ell + 2\nu.$$

For these three criterions, the most adapted model is the one with the smallest criterion value.

5. Experimental results. In this section, experimental results on artificial and real data illustrate the main features of the adaptive linear models.

5.1. *Simulation study.* This first experiment aims to evaluate the ability of the adaptive linear models, introduced in Section 3, to find the transformation between populations P and P^* as well as the ability of the model selection criterions to select the most appropriate transformation model.

Experimental setup. Firstly, a one-dimensional regression model was generated for the reference population P on a basis of natural cubic Splines with 5 degrees of freedom. Then, a regression model was built for the new population P^* from the model of P by multiplying the regression parameters of P by a given transformation matrix $\mathbf{\Lambda}$. Since it is impossible to report here numerical experiments for all existing transformation models, results are presented for only one transformation model: the model M2. The specific transformation matrix $\mathbf{\Lambda} = diag(1.5, 2, 2, 2, 2)$ was chosen for generating the regression model of P^*. In order to compare the performance of the different transformation models, some observations for population P^* were simulated

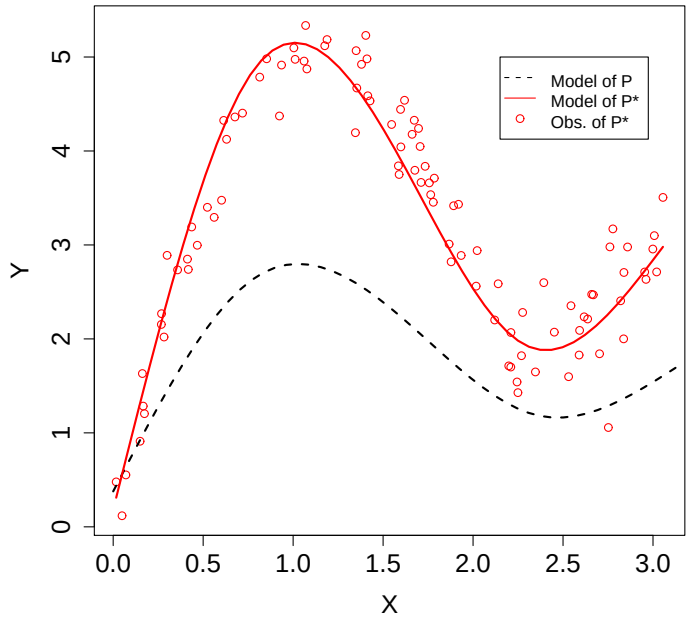

FIG 1. *Regression models of the populations P and P^* and simulated observations of population P^*: the model of P was generated on a basis of cubic Spline functions with 5 degrees of freedom and the model of P^* was obtained from the model of P by multiplying its parameters by (1.5,2,2,2,2).*

from its regression model. These observations were simulated with an additive Gaussian noise $\epsilon \sim \mathcal{N}(0, 0.3)$. Figure 1 shows the regression models for both populations P and P^* as well as 100 observations simulated from the regression model of P^*. The simulated observations of population P^* were used in the experiment by the different linear transformation models for estimating the transformation between P and P^*. Thus, it has been possible afterward to compare the estimated parameters of P^* (obtained by multiplying the regression parameters of P by the estimated transformation matrix $\hat{\mathbf{\Lambda}}$) with the actual regression parameters of P^*. This comparison was measured by the sum of squared differences between estimated and actual regression parameters of P^*. In addition, the values of the three model selection criterions, presented in Section 4.4, were computed for each model in order to empirically verify their ability for finding the most appropriate transformation model. Finally, the protocol described above was applied for different dataset sizes ranging from 25 to 1000 observations for studying the

effect of the learning dataset size on the prediction ability of the different models.

Experimental results. Table 2 presents the numerical evaluation of the ability of the adaptive linear models M0, M1, M2, M3, M4 and M5 to estimate the transformation parameters and of the ability of the model selection criterions to find the most appropriate transformation model. The first and the second columns of Table 2 respectively indicate the size of the learning dataset and the name of the used transformation model. The third, fourth and fifth columns respectively give the values of the model selection criterions PRESS, BIC and AIC associated to each model. Finally, the sixth and the last columns respectively provide the Residual Sum of Squares (RSS), computed on a test dataset different from the learning set, and the sum of squared differences between estimated and actual parameters for population P^*. The bold numbers of the table correspond to the "best value" of each column for a given dataset size (let remind that for the three model selection criterions, the most appropriate model is the one associated with the smallest value). On the one hand, it appears clearly that both PRESS, BIC and AIC select the transformation model M2 as the most appropriate for modelling the transformation between P and P^* and that corresponds to the truth. The first conclusion is that these three criterions are well suited to select the transformation model in such a case. On the other hand, it can be noticed that the model M0, which corresponds to the usual OLS model on P^*, is very sensitive to the size of the dataset used for learning whereas the adaptive linear models M1 to M5 are less sensitive. Furthermore, the model M0 gives disappointing estimations for all dataset sizes whereas the other models, which are more parsimonious and which benefit from the knowledge on P, give satisfying results for a large range of dataset sizes. In particular the model M2 provides on average a very good estimation of the actual regression parameters, even with only 25 observations. Figure 2 shows the estimated regression model of the population P^* for the six studied models. These estimations were obtained with a learning dataset of 100 observations. As it could be expected, the M0 estimation is very far away from the actual model and the models M1, M2 and M3 give very good estimations of the regression model. The effect of the constraints on the models can also be observed on this figure. For instance, the model M5 is not flexible enough to correctly estimate the transformation and this is due to the fact that it assumes that only the intercept is modified.

To summarize, this experiment has shown that the adaptive linear models, proposed in this paper, are able to estimate correctly a transformation between two populations with non-linear regression models and that even in

n^*	Model	PRESS	BIC	AIC	RSS	Prm. diff.
25	M0	24283.92	16.326	16.033	199.827	1312.998
	M1	0.131	0.902	0.658	0.109	2.142
	M2	**0.109**	**0.669**	**0.571**	**0.094**	**0.118**
	M3	0.128	0.796	0.748	0.119	0.528
	M4	0.192	1.241	1.192	0.162	1.255
	M5	0.597	2.340	2.291	0.584	10.348
50	M0	19196.07	16.209	15.979	51.884	674.779
	M1	0.098	0.669	0.478	0.103	1.770
	M2	**0.091**	**0.498**	**0.421**	**0.096**	**0.056**
	M3	0.111	0.661	0.623	0.119	0.548
	M4	0.157	1.042	1.004	0.163	1.211
	M5	0.525	2.220	2.182	0.545	10.639
100	M0	1754.953	8.800	8.644	41.239	734.003
	M1	0.096	0.614	0.484	0.091	1.510
	M2	**0.093**	**0.509**	**0.456**	**0.089**	**0.014**
	M3	0.115	0.699	0.673	0.109	0.425
	M4	0.172	1.128	1.102	0.157	1.002
	M5	0.455	2.072	2.046	0.511	7.141
250	M0	522.120	5.512	5.427	24.329	466.621
	M1	0.090	0.504	0.434	0.090	1.404
	M2	**0.089**	**0.450**	**0.422**	**0.089**	**0.005**
	M3	0.116	0.704	0.690	0.111	0.474
	M4	0.172	1.135	1.121	0.161	0.993
	M5	0.467	2.089	2.075	0.534	7.33
500	M0	270.574	5.034	5.004	6.633	272.080
	M1	0.092	0.495	0.453	0.091	1.347
	M2	**0.091**	**0.463**	**0.446**	**0.090**	**0.004**
	M3	0.116	0.698	0.689	0.113	0.427
	M4	0.167	1.090	1.082	0.155	0.926
	M5	0.463	2.075	2.067	0.501	7.122
1000	M0	184.00	4.669	4.618	3.519	121.248
	M1	0.089	0.450	0.425	0.091	1.368
	M2	**0.089**	**0.432**	**0.422**	**0.090**	**0.001**
	M3	0.113	0.669	0.665	0.112	0.430
	M4	0.168	1.093	1.088	0.156	0.947
	M5	0.453	2.051	2.046	0.501	7.083

TABLE 2

Evaluation of the model selection and of the parameter estimation on data simulated according to the model M2 on a basis of cubic Spline functions for different dataset sizes: PRESS, BIC and AIC values are per point, the RSS value was computed on a test dataset and "Prm. diff" is the sum of squared differences between estimated and actual parameters for population P^*.

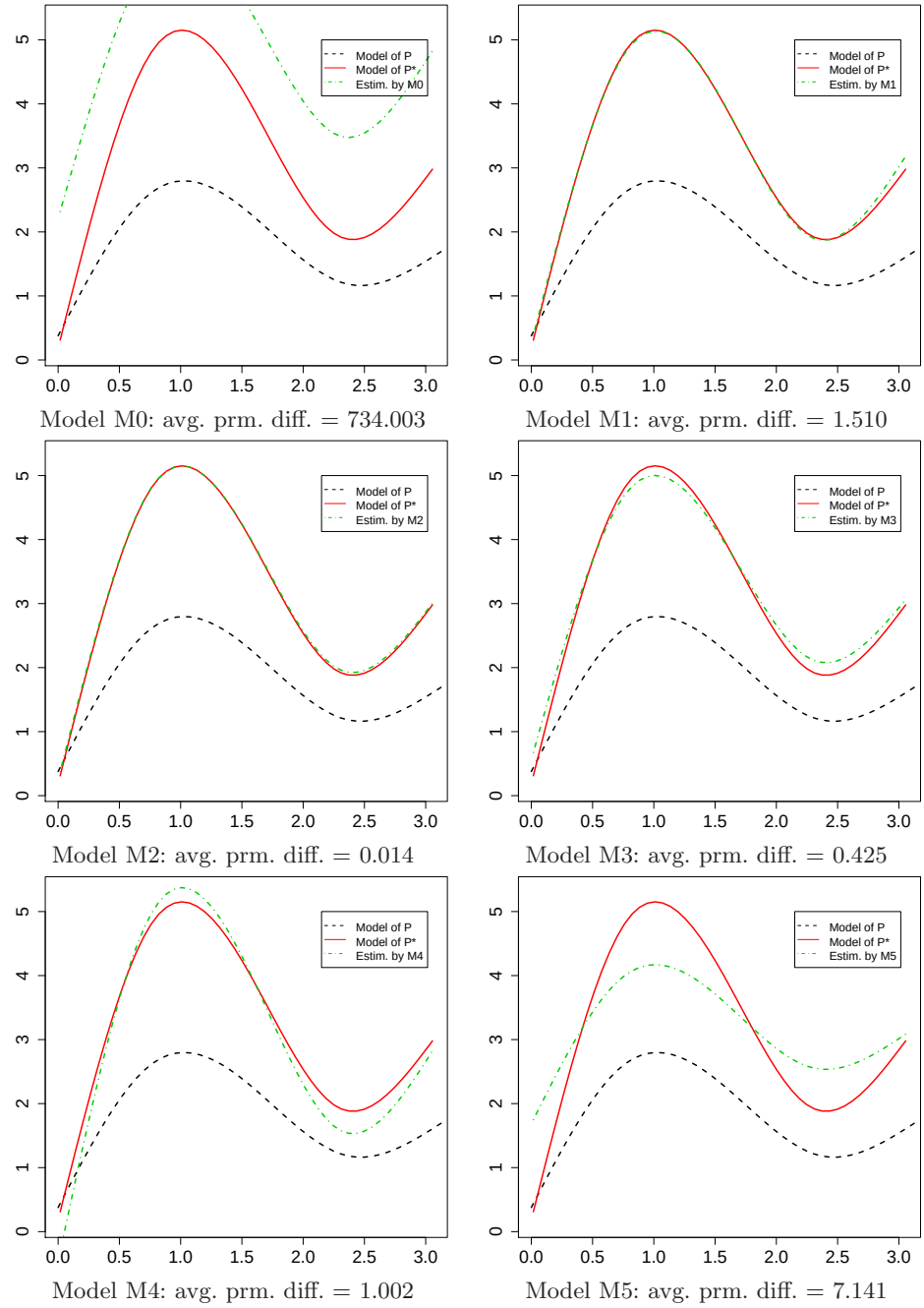

FIG 2. *Parameter estimation with the different linear transformation models on data simulated according to the transformation model M2 on a basis of cubic Spline functions. These estimations were computed with a dataset of 100 observations. The difference between estimated and actual regression parameters is measured by the sum of squared differences.*

situations where the number of observations of P^* is limited. This study has also demonstrated that either the cross-validated PRESS criterion and information criterions BIC and AIC are adapted to select the most appropriate model among the 7 adaptive linear models.

5.2. *Real data study: growth of Tetrahymena cells.* A biological dataset is considered here in order to highlight the ability of our approach to deal with real data.

The data. The *hellung* dataset [1], collected by P. Hellung-Larsen, reports the growth conditions of *Tetrahymena* cells. The data arise from two groups of cell cultures: cells with and without glucose added to the growth medium. For each group, the average cell diameter (in μm) and the cell concentration (count per ml) were recorded. The cell concentrations of both groups were set to the same value at the beginning of the experiment and it is expected that the presence of glucose in the medium affects the growth of the cell diameter. In the sequel, cells with glucose will be considered as coming from population P whereas cells without glucose will be considered as coming from population P^*.

Experimental setup. In order to fit a regression model on the cell group with glucose, the PRESS criterion was used to select the most appropriate basis function. It results that a 3rd degree polynomial function is the most adapted model for these data and this specific basis function will be used for all methods in this experiment. The Figure 3 shows the ordinary least square (OLS) estimates of the 3rd degree polynomial regression model respectively for the cell population P (with glucose) and the cell population P^* (without glucose). The first remark suggested by this figure is that the right extremity of the OLS regression curve of population P^* (bottom red line) is very influenced by the last observation. This highlights the non-robustness of this regression model learned on only 19 points. The goal of this experiment is to compare the stability and the effectiveness of the usual OLS regression method with our adaptive linear regression models according to the size of the P^* learning dataset. For this, 4 different learning datasets are used: all P^* observations (19 obs.), all P^* observations for which the concentration is smaller than 4×10^5 (17 obs.), smaller than 2×10^5 (14 obs.) and smaller than 1×10^5 (11 obs.). In order to evaluate the prediction ability of the different methods, the PRESS criterion as well as the RSS value on the whole P^* dataset are computed for these 4 different sizes of learning dataset.

[1] The hellung dataset is available in the ISwR package for R.

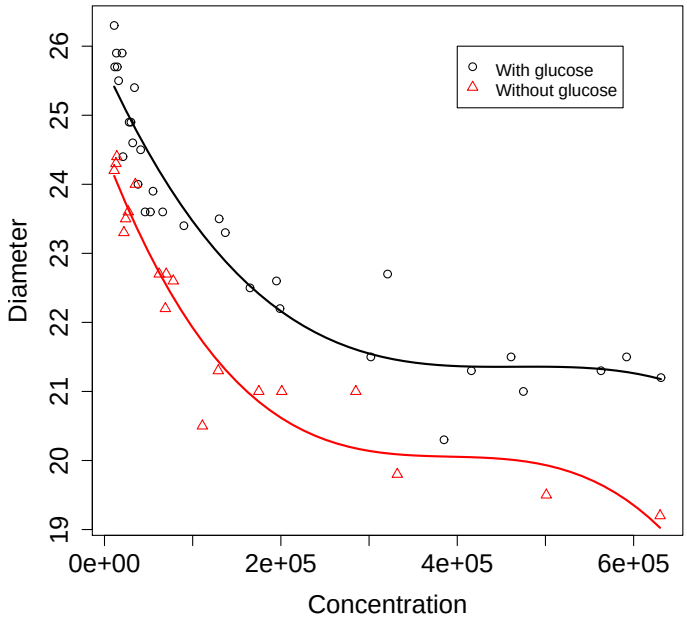

FIG 3. *The* hellung *dataset: diameter* vs. *concentration for* Tetrahymena *cells.*

Experimental results. Figure 4 illustrates the effect of the learning set size on the prediction ability of the studied regression methods. The panels of Figure 4 displays the curve of the usual OLS regression method (M0) in addition to the curves of the 5 adaptive linear models (models M1 to M5) for different sizes of the learning set (the blue zones indicate the ranges of the observations of P^* used for learning the models). The model M6 which is equivalent to the usual OLS regression method on the population P is also displayed. The first remark suggested by these results is that the most complex models, OLS (M0) and M1, appear to be very instable in such a situation where the number of learning observations is small. Secondly, the model M4 is more stable but its main assumption (same intercept as the regression model of P) seems to be an overly strong constraint and stops it from fitting correctly the data. Finally, the models M2, M3 and M5 turn out to be very stable and flexible enough to correctly model the new population P^* even with very few observations. This visual interpretation of the experiment is confirmed by the numerical results presented in Tables 3 and 4. These tables respectively report the value of the PRESS criterion and

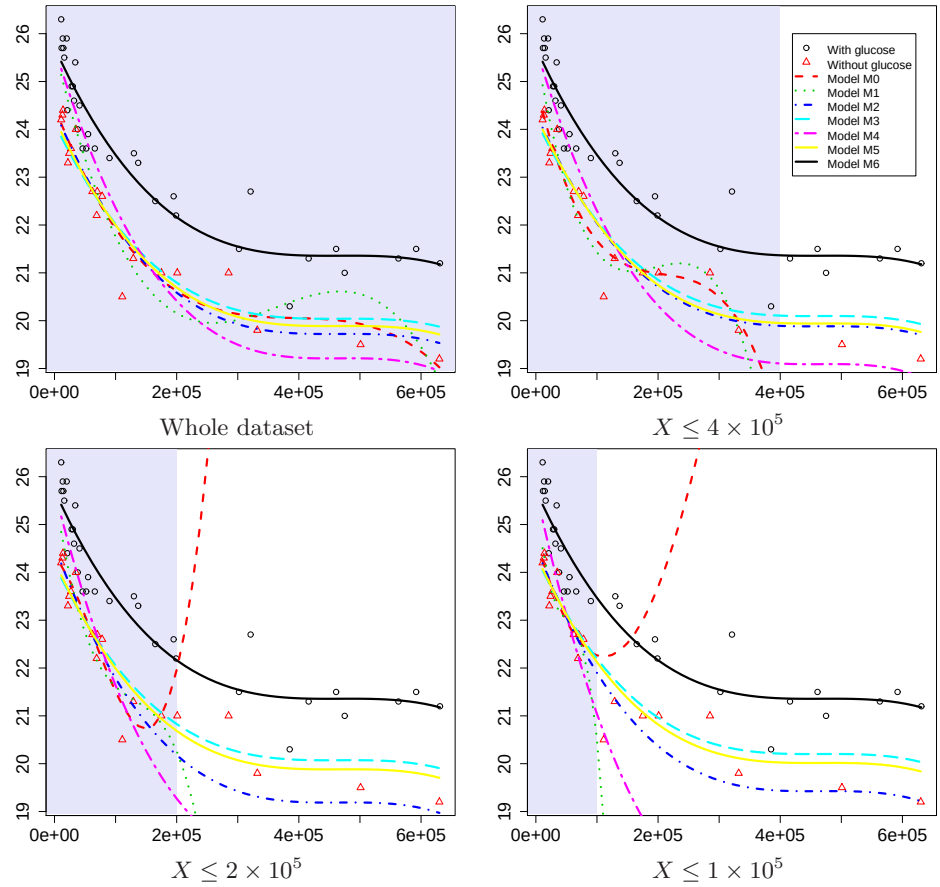

FIG 4. *Effect of the learning set size on the prediction ability of the studied regression methods for the* hellung *dataset. The blue zones correspond to the parts of the observations of* P^* *used for learning the models.*

Method	whole dataset	$X \leq 4 \times 10^5$	$X \leq 2 \times 10^5$	$X \leq 1 \times 10^5$
OLS on P^* (M0)	0.897	0.364	0.432	0.303
Model M1	3.332	0.283	2.245	0.344
Model M2	0.269	0.294	0.261	0.130
Model M3	0.287	0.271	0.289	0.133
Model M4	0.859	1.003	0.756	0.517
Model M5	**0.256**	**0.259**	**0.255**	**0.124**

TABLE 3

Effect of the learning set size on the PRESS criterion of the studied regression methods for the hellung *dataset. The best values of each column are in bold.*

Method	whole dataset	$X \leq 4 \times 10^5$	$X \leq 2 \times 10^5$	$X \leq 1 \times 10^5$
OLS on P^* (M0)	0.195	47.718	4.5×10^3	145.846
Model M1	0.524	164.301	2.3×10^3	5.9×10^5
Model M2	**0.218**	**0.226**	0.304	**0.245**
Model M3	0.258	0.262	0.259	0.290
Model M4	0.791	0.796	1.472	3.046
Model M5	*0.230	*0.233	***0.230**	*0.246
OLS on P (M6)	2.388	2.388	2.388	2.388

TABLE 4

Effect of the learning set size on the PRESS criterion of the studied regression methods for the hellung *dataset. Best values of each column are in bold and the stars indicate the selected models by the PRESS criterion.*

the RSS associated to the studied regression methods for the different sizes of learning dataset. Table 3 confirms clearly that the most stable, and therefore appropriate, model for estimating the transformation between populations P and P^* is the model M5. Another interesting conclusion is that both models M2 and M3 obtained very low PRESS values as well. These predictions of the model stability appear to be satisfying since the comparison of Tables 3 and 4 shows that the model selected by the PRESS criterion is always an efficient model for prediction. Indeed, the Table 4 show that the most efficient models in practice are the models M2 and M5 which are the "preferred" models by PRESS.

To conclude, this study demonstrates that the adaptive linear models can be successfully applied to real data in order to transfer a knowledge on a reference population (here the cells without glucose) to a new population (here the cells with glucose). As it could be expected, the advantage of adaptive linear models makes particularly sense when the number of observations of the new population is limited and this happens frequently in real situations due to censorship or to technical constraints (experimental cost, scarcity, ...).

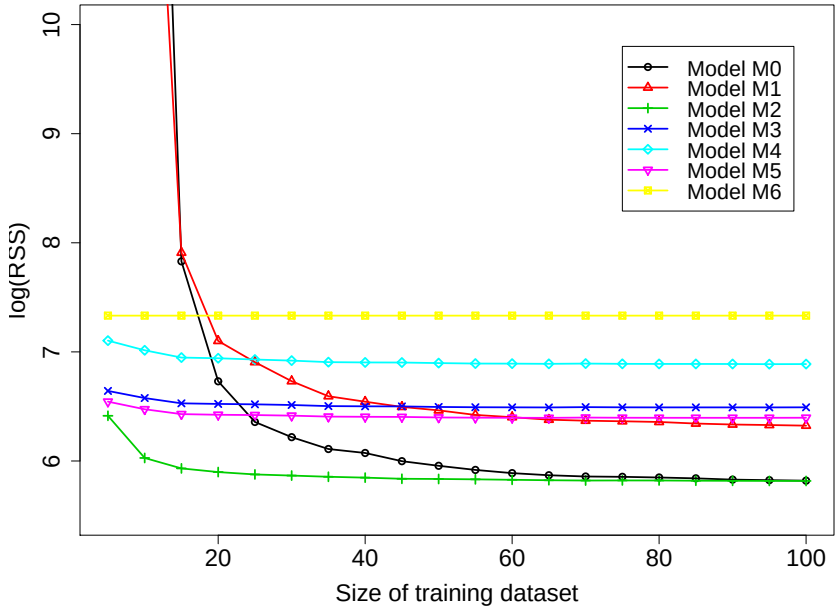

FIG 5. *RSS results for the Birmingham-San Jose data.*

5.3. *Real data study: modelling of housing market in different U.S. cities.* In this section, the interest of the adaptive linear models is illustrated by an application to the modelling of housing market in different U.S. cities. This application aims to demonstrate that it is possible to adapt a regression model learned on a reference city to another one *via* the adaptive linear models by using only few samples from the new city and thus to save up an expensive collect of new data.

The data. For this experiment, the 1984 American Housing Survey of the U.S. Department of Commerce is used. The data collection (23) contains information from samples of housing units in 11 Metropolitan Statistical Areas, among which the cities of Birmingham, Alabama (East coast) and of San Jose, California (West coast). Fourteen relevant features have been selected among more than 500 available features for modelling the housing market of Birmingham. The selected features include the number of rooms, the area, the monthly cost of the housing as well as other informations about the unit and the tenants. Finally, based on these 14 features, the response variable to predict is the value of the housing.

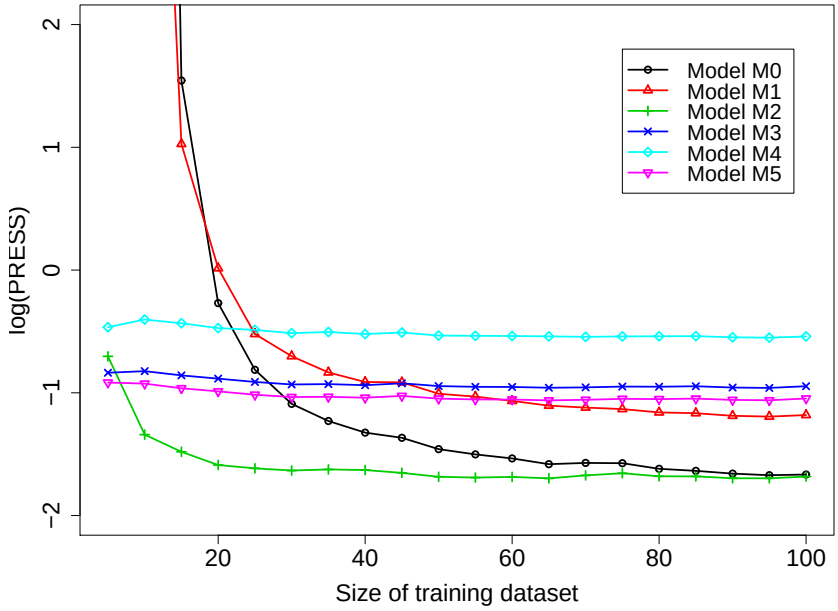

FIG 6. *PRESS criterion for the Birmingham-San Jose data.*

Experimental setup. A semi-log regression model for the housing market of Birmingham was learned using all the 1541 available samples and, then, the 7 adaptive linear models were used to transfer the regression model of Birmingham to the housing market of San Jose. In order to evaluate the ability of the adaptive linear models to transfer the Birmingham knowledge to San Jose in different situations, the experiment protocol was applied for different sizes of San Jose samples ranging from 5 to 921 observations. For each dataset size, the San Jose samples were randomly selected among all available samples and the experiment was repeated 50 times in order to average the results. For each adaptive linear model, the PRESS criterion and the residual sum of squares (RSS) were computed, by using the selected sample for PRESS and the whole San Jose dataset for RSS.

Experimental results. Figure 5 shows the logarithm of the RSS for the different adaptive linear models regarding to the size of the used San Jose samples. Similarly, Figure 6 shows the logarithm of the PRESS criterion. Firstly, Figure 5 indicates that the model M6, which corresponds to the Birmingham's model, is actually not adapted for modelling the housing market of

San Jose since it obtains a not satisfying RSS value. Let notice that the curve corresponding to the RSS of the model M6 is constant since the regression model has been learned on the Birmingham's data and consequently does not depend on the size of the San Jose's dataset selected for learning. Secondly, the model M0, which is equivalent to OLS on the San Jose samples, is particularly disappointing (large values of RSS) if it is learned with a very small number of observations and becomes more efficient for learning datasets larger than 50 observations. The model M1 has a similar behaviour for small learning datasets but turns out to be less interesting than M0 when the size of the learning dataset is larger. These behaviours are not surprising since both models M0 and M1 are very complex models and then need large datasets to be correctly learned. Conversely, the models M2 to M5 appear not to be sensitive to the size of the dataset used for adapting the Birmingham model. Particularly, the model M2 obtains very low RSS values for a learning dataset size as low as 20 observations. This indicates that the model M2 is able to adapt the Birmingham model to San Jose with only 20 observations. Moreover Table 5 indicates that the model M2 provides better prevision results than the model M0 for the housing market of San Jose for learning dataset sizes less than 100 observations. Naturally, since the model M0 is more complex, it becomes more efficient than the model M2 for larger datasets even if the difference is not so big for large learning datasets. Figure 6 demonstrates that the PRESS criterion, which will be used in practice since it is computed without a validation dataset, allows the practician to successfully select the most appropriated transfer model. Indeed, it appears clearly that the PRESS curves are very similar to the RSS curves computed on the whole dataset. Finally, in such a context, the transformation parameters obtained by the different adaptive linear models can be interpreted in an economic way and this could be interesting for economists. In particular, the estimated transformation parameters by the model M2 with the whole San Jose dataset are $\lambda_0 = 1.439$ and $\lambda = 0.447$. The fact that the San Jose's intercept is almost 50% larger than the one of Birmingham suggests that the minimal basis price of an housing is more expensive in San Jose than in Birmingham. However, the fact that the regression coefficients associated to the explanatory variables of San Jose are on average 50% smaller than the one of Birmingham could mean that the growing of the price according to the housing features is more moderated.

This experiment has shown that the adaptive linear models are able to transfer the knowledge on the housing market of a reference city to the market of a different city with a small number of observations. Furthermore, the interpretation of the estimated transformation parameters could help the

Model	10 obs.	25 obs.	50 obs.	100 obs.	250 obs.	all obs.
Model M0	3.5×10^7	576.9	386.1	336.8	**310.7**	**297.5**
Model M2	**414.8**	**356.7**	**342.1**	**336.0**	332.5	330.1
Model M6	1528.9	1528.9	1528.9	1528.9	1528.9	1528.9

TABLE 5
RSS results for the Birmingham-San Jose data.

practician to analyse in an economic way the differences between the studied populations.

6. Conclusion. Before each statistical analysis, the indispensable collect of data is often an expensive step. Even if the same analysis has been achieved in a relatively similar situation, a new collect of data is needed since the situation is usually not exactly similar. In a regression framework, this paper provides a way to adapt a regression model from a given situation to another new one, and thus to save up a new expensive collect of data. In this perspective, a family of adaptive linear models has been defined and, since they are more parsimonious than a complete regression model, they need only few samples for providing a satisfying estimation of the new regression model. The conducted experiments have demonstrated that the adaptive linear models are able to successfully transfer a knowledge on a well known reference population to another population even with a very few observation. In particular, the efficiency of the proposed models has been illustrated on a economic application by adapting the regression of the housing price versus housing features from the city of Birmingham to the city of San Jose. While a sample size of at least 100 observations is needed to estimate directly the San Jose's regression model, only 20 data are needed to obtain a similar estimation quality with the adaptive linear models. In addition, the estimated transformation parameters could help practicians to analyse the differences between both populations.

Acknowledgments. The authors would like to thank Professor Patrice Gaubert (University Paris XII) for providing the preprocessed economical data and for his very useful advices and Professor Christophe Biernacki (University Lille I) for comments and discussions.

References.

[1] AKAIKE, H. (1974). *A new look at the statistical model identification*, IEEE Transactions on Automatic Control 19(6):716–723.

[2] ALLEN, D.M. (1974). *The relationship between variable selection and data augmentation and a method for prediction*, Technometrics, 16:125–127.

[3] Beninel, F. and Biernacki, C. (2007). *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.

[4] Bertness, K., Hickernell, R., Hays, S. and Christensen, D. (1998). *Noise Reduction in Optical in situ Measurements for Molecular Beam Epitaxy by Substrate Wobble Normalization*, Journal of Vacuum Science and Technology B, 16(3):1492-1497.

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

[6] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*, Springer.

[7] Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis*, Third ed., Wiley, New-York.

[8] Feudale, N., Woody, N., Tan, H., Myles, A., Brown, S. and Ferre J. (2002). *Transfer of multivariate calibration models: a review*, Chemometrics and Intelligent Laboratory Systems, 64:181:192.

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

[10] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, Springer, New-York.

[11] Jacques, J. and Biernacki, C. (2007). *Classement de données binaires lorsque les populations d'apprentissage et de test sont différentes.*, Revue des Nouvelles Technologies de l'Information, Data Mining et apprentissage statistique : application en assurance, banque et marketing, A1, 109-130.

[12] Mallat, S. (1999). *A Wavelet Tour of Signal Processing*, Second ed., Academic Press.

[13] McCullagh, P. and Nelder, J. (1990). *Generalized Linear Models*, Chapman and Hall, London.

[14] Ratkowsky, D.A. (1990). *Handbooks of Nonlinear Regression Models*, Marcel Dekker, New-York.

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

[16] Searle, S.R. (1971). *Linear Models*, Wiley, New-York.

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

[18] Storkey, A. and Sugiyama, M. (2007). *Mixture regression for covariate shift*, in Advances in Neural Information Processing Systems 19, Schölkopf, J. C. Platt and T. Hoffmann editors, Cambridge, MIT Press, 1337–1344.

[19] Sugiyama, M. and Müller, K-R. (2005). *Input-dependent estimation of generalization error under covariate shift*, Statistics & Decisions, 4:249–279.

[20] Sugiyama, M., Müller, K-R. and Krauledat, M. (2007). *Covariate Shift Adaptation by Importance Weighted Cross Validation*, Journal of Machine Learning Research, 8:985–1005.

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

[22] Tobin, K., Karnowski, T., Arrowood L., Ferrell, R., Goddard, J. and Lakhani, F. (2002). *Content-based image retrieval for semiconductor process char-*

acterization, Journal on Applied Signal Processing, 1:704–713.
[23] United States Department of Commerce, Bureau of the Census, Washington, DC (1989). *American Housing Survey, 1984: MSA File*, Ann Arbor, MI: Inter-university Consortium for Political and Social Research.
[24] Wang, Y., Veltkamp D. and Kowalski B. (1991). *Multivariate instrument standardization*, Analytical chemistry, 63(23):2750-2756.

Charles Bouveyron
SAMOS-MATISSE, CES,
UMR CNRS 8174,
University Paris I (Panthéon-Sorbonne),
Paris, France
charles.bouveyron@univ-paris1.fr
URL: http://samos.univ-paris1.fr/spip/-Charles-Bouveyron

Julien Jacques
Laboratoire Paul Painlevé,
UMR CNRS 8524,
University Lille I,
Lille, France
julien.jacques@polytech-lille.fr
URL: http://math.univ-lille1.fr/∼jacques