Adaptive Linear Models for Regression: improving prediction when population has changed

Charles Bouveyron, Julien Jacques

Adaptive linear models for regression: improving prediction when population has changed

Charles Bouveyron[a], Julien Jacques[b]

[a]*Laboratoire SAMM, Université Paris I Panthéon-Sorbonne, Paris, France.*
[b]*Laboratoire Paul Painlevé, UMR CNRS 8524, Université Lille I, Lille, France.*

Abstract

The general setting of regression analysis is to identify a relationship between a response variable Y and one or several explanatory variables X by using a learning sample. In a prediction framework, the main assumption for predicting Y on a new sample of observations is that the regression model $Y = f(X) + \epsilon$ 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 $f(X)$ and the new one $f^*(X)$. The main interest of the proposed adaptive models is to allow the build of a regression model for the new population with only a small number of observations using the knowledge on the reference population. The efficiency of this strategy is illustrated by applications on artificial and real datasets, including the modeling of the housing market in different U.S. cities. A package for the R software dedicated to the adaptive linear models is available on the author's web page.

Key words: regression adaptive, estimation, knowledge transfer, linear transformation models, housing market in different U.S. cities.

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 has focused on the nature of the regression model: linear model [1], generalized linear model [2] and non linear model [3]. We refer to [4] for a general survey on regression analysis.

1.1. The problem of adapting a knowledge to a new situation

In this paper, we are concerned with the following question: how to adapt an existing regression model to a new situation, for which the variables are identical (with a possible different probability density distribution) but where the relationship between response and explanatory variables could have changed? As a motivating example, our discussion will be centered on the following socio-economical application: 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 several housing descriptive variables estimated using a large learning sample. To conquer new markets, this company plans to open several agencies in the West coast, and would use its regression model without having to spend a lot of time and money in collecting new housing market data for this area. Considering that the link between housing descriptive variables and housing price for the West and East coasts is, on the one hand, probably not the same but, on the other hand, not completely different, this work will consider a set of transformation models between both West and East coast regression models. This paper will therefore focus on transferring the knowledge on a reference population to a

new population by inferring the relationship between both regression models. Moreover, the exhibition of a link between both populations could be helpful for the interpretation of the modeled phenomenon.

1.2. Related works

To our knowledge, there have been only few contributions dealing with this original problem although it is very interesting and very frequent in practical applications. In the machine learning community, a related problem is investigated under the keyword *Covariate Shift*. The covariate shift problem considers that the probability density of the new data is different from 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 never the case in practice and the regression model estimated with the learning data could be very disappointing when applied to data with a different probability distribution. Several recent works [5–9] have contributed to analyze this context. However, most of these works need to know the probability distribution of the new data or at least an estimation of this probability distribution. In practice, this is a difficult problem which requires a sufficiently large sample of observations. The focus of the present work is 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 [10] on quantitative

variables and in [11] in the case of binary variables. The authors considered a model-based discriminant rule for classifying individuals from a population which differs from the learning one. For this, they introduced a family of linear models modeling the transformation between the reference population and the new population. An extension of this work to logistic regression was recently proposed in [12]. Finally, some other works cover the problematic of knowledge transfer in specific industrial contexts. For instance, [13] 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 [14] and the neural network based nonlinear transformation [15]. Several works [16, 17] have also considered this problem in the field of semiconductor industry.

This paper is structured 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. Inference and model selection procedures are discussed in Section 4. Section 5 provides a simulation study in a spline regression context and two real applications including the modeling of the housing market in different U.S. cities. Finally, some concluding remarks and future directions are discussed in Section 6.

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 $\mathbf{x}_j$, $j = 1, ..., n$, are the values of the deterministic explanatory variable $\mathbf{X} = (X^{(1)}, \ldots, X^{(p)})^t \in \mathbb{R}^p$ and the corresponding y_j are the realizations of the stochastic variable $Y \in \mathbb{R}$. A general data modeling problem consists in identifying the relationship between the explanatory variable $\mathbf{X}$ (known as well as covariate) and the response variable Y (or dependent variable). Both standard parametric and non-parametric approaches consider the following regression model:

$$Y = f(\mathbf{X}, \boldsymbol{\beta}) + \epsilon, \tag{1}$$

where the residuals $\epsilon \sim \mathcal{N}(0, \sigma^2)$ are independent 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 specify the link between the response variable Y and the covariate $\mathbf{X}$ is through the assumptions 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 (*cf.* [18]):

$$f(\mathbf{x}, \boldsymbol{\beta}) = \sum_{i=0}^{d} \beta_i \psi_i(\mathbf{x}), \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^t \in \mathbb{R}^{d+1}$ 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 the identity, polynomial functions, splines functions [19] or wavelets [20]. We refer to [4] for a general survey. Let us 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) can be also written in a matrix form as follows:

$$f(\mathbf{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^t \Psi(\mathbf{x}), \tag{3}$$

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 us now assume that the regression function f has been estimated in a preliminary study by using a sample S of population P, 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 assumed to be small). The difference between P and P^* can be for instance geographical (as in the U.S. housing market application) or temporal. However, the nature of both populations has to be similar to match the purpose of this paper. The new regression model for P^* can be classically written:

$$Y^* | \mathbf{X}^* \sim \mathcal{N}(f^*(\mathbf{X}^*, \boldsymbol{\beta}^*), \sigma^{*2}), \tag{4}$$

with

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

The aim of this work is therefore to define a link between the regression functions f and f^*.

3. Adaptive linear models for regression

In this section, a link between the regression function of P and P^* is exhibited, and a family of transformations is then introduced to solve the problem of adapting an existing regression model of a reference population P to a new population P^*.

3.1. The transformation model

In order to exhibit a link between both regression functions, we make the following important assumptions.

Assumption (A1). First, we postulate that the number of basis functions and the basis functions themselves are the same for both regression models ($d^* = d$ and $\psi_i^* = \psi_i$, $\forall i = 1, ..., d$), which is natural since the variables are identical in both populations. The regression function of the population P^* can be therefore written:

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

Assumption (A2). Second, we assume that the transformation between f and f^* applies only on the regression parameters. We therefore define $\boldsymbol{\Lambda}$, the transformation matrix between the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, as follows:

$$\boldsymbol{\beta}^* = \boldsymbol{\Lambda}\boldsymbol{\beta},$$

and this yields to the following expression of f^*:

$$f^*(\mathbf{x}^*, \mathbf{\Lambda\beta}) = (\mathbf{\Lambda\beta})^t \Psi(\mathbf{x}^*). \tag{5}$$

Given that the number of parameters to estimate in the transformation matrix $\mathbf{\Lambda}$ is $(d+1)\times(d+1)$ and that the number of free parameters for learning a new regression model directly from the sample S^* is $(d+1)$, the transformation model (5) is consequently highly over-parametrized. It is therefore necessary to introduce some constraints on the transformation model such that the number of free parameters to estimate is lower or equal to d.

Assumption (A3). Third, we assume that the relation between the response variable and a specific covariate in the new population P^* only depends on the relation between the response variable and the same covariate in the population P. Thus, for $i = 0, \ldots, d$, the regression parameter β_i^* only depends on the regression parameter β_i and the matrix $\mathbf{\Lambda}$ is consequently diagonal. The transformation can be finally written in term of the regression parameters of both models as follows:

$$\beta_i^* = \lambda_i \beta_i \qquad \forall i = 0, \ldots, d, \tag{6}$$

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

3.2. A family of transformation models

Since the aim of this study is to learn a regression model for P^* with only few observations, we define in this section parsimonious models by imposing some constraints on the transformation model (6). First, we allow some of the parameters λ_i to be equal to 1 (in this case the regression parameters β_i^* are

equal to β_i). Second, we allow as well some of the parameters λ_i to be equal to a common value, *i.e.* $\lambda_i = \lambda$ for given $0 \leq i \leq d$. The number of possible models obtained with such a strategy is consequently very large (formally $\sum_{m=0}^{d+1} \binom{d+1}{m} \times \left(1 + \sum_{l=2}^{m} \binom{l}{m}\right)$) and it obviously impossible to consider all these models in practice. These models, named adaptive linear models in the sequel, are declined below into two families: specific transformation models and prior-based transformation models.

3.2.1. Specific transformation models

We propose in this paragraph a family of 7 transformation models, selected on parsimony and interpretability criteria, 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 the populations P and P^*. It is equivalent to learning a new regression model from the sample S^*, since there is no constraint on the $d + 1$ parameters β_i^* $(i = 0, ..., d)$, and the number of free parameters in $\mathbf{\Lambda}$ is consequently $d + 1$ as well.

- 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

Table 1: Complexity (number of parameters ν) of the transformation models.

Model	$M0$	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$
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 ν	d+1	d	2	1	1	1	0

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 regression model.

The numbers of parameters to estimate for these transformation models are presented in Table 1. The choice of this family is arbitrary and motivated by the will of the authors to treat similarly all the covariates in this general discussion. However, in practical applications, we encourage the practician to consider some additional transformation models specifically designed to his application and motivated by his prior knowledge on the subject. This is discussed in the next section.

3.2.2. Prior-based transformation models

Although only seven pragmatic transformation models have been presented in the previous section, some other transformation models could be considered as well, for which the complexity (in number of parameters) will be intermediate between the $M1$ complexity (d) and the $M2$ complexity (2). Indeed, the practician could have in some specific cases to use intermediate transformation models suggested by some prior informations on the covariates, which leads to impose specific constraints on parameters λ_i for given $i \in \{1, \ldots, d\}$. For instance, let us 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, proposed in the previous section, is made of two main steps corresponding to the estimation of the regression parameters of the population P and to the estimation of the transformation parameters using samples of the population P^*. The regression parameters of P^* are then obtained by plug-in. The ordinary least square (OLS) method is used, but we present in this paper the equivalent maximum likelihood estimation method in order to compute penalized likelihood model selection criteria for model selection.

4.1. Estimation of the regression parameters

Let us consider a data set of inputs $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with corresponding response values grouped in a column vector $\mathbf{y} = (y_1, \ldots, y_n)^t$. Under the assumptions of the model (2), the log-likelihood of $\mathbf{y}$ given $\mathbf{x}$, $\boldsymbol{\beta}$ and σ^2 is:

$$\ln l(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}, \sigma^2) = -n \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{j=1}^{n} \left(y_j - \boldsymbol{\beta}^t \Psi(\mathbf{x}_j)\right)^2 . \tag{7}$$

Maximizing the log-likelihood according to $\boldsymbol{\beta}$ is equivalent to minimizing $\sum_{j=1}^{n} \left(y_j - \boldsymbol{\beta}^t \Psi(\mathbf{x}_j)\right)^2$ and thus the maximum likelihood estimator is equivalent to the ordinary least square estimator:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{OLS} &= (\boldsymbol{\Psi}^t \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^t \mathbf{y}, \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{j=1}^{n} \left(y_j - \boldsymbol{\beta}^t \Psi(\mathbf{x}_j)\right)^2 . \end{aligned}$$

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

4.2. Estimation of the transformation parameters

For this second step, it is assumed that $\boldsymbol{\beta}$ is known (in fact it is estimated in the previous step). As previously noticed, the full model $M0$ corresponds to a completely new regression model adjusted on the sample S^*. Similarly, the model $M6$, which assumes no transformation between P and P^*, does not require the estimation of any transformation parameters. Let us 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 (7), the log-likelihood of model (4) is:

$$\ln l(\mathbf{y}^*; \mathbf{x}^*, \boldsymbol{\Lambda}, \sigma^2) = -n^* \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{j=1}^{n^*} \left(y_j^* - \boldsymbol{\beta}^t \boldsymbol{\Lambda}^t \Psi(\mathbf{x}_j^*)\right)^2 . \tag{8}$$

This log-likelihood must be maximized according to the transformation matrix $\mathbf{\Lambda}$, what leads to the OLS estimator:

$$\hat{\mathbf{\Lambda}}^{OLS} = \underset{\mathbf{\Lambda}\in\mathcal{D}}{\operatorname{argmin}} \sum_{j=1}^{n^*} (y_j^* - (\mathbf{\Lambda}\boldsymbol{\beta})^t \Psi(\mathbf{x}_j^*))^2, \tag{9}$$

where $\mathcal{D}$ is a set of diagonal matrices depending on the model of transformation at hand. For instance, with the model $M3$, this set is $\mathcal{D} = \{\lambda I_{d+1}, \lambda \in \mathbb{R}\}$ where I_{d+1} is the identity matrix of size $d+1$.

4.2.1. Specific transformation models

Least square estimators of the specific models M1 to M5 are derived below.

Model M_1. As the transformation matrix is $\mathbf{\Lambda} = diag(1, \lambda_1, \ldots, \lambda_d)$, the log-likelihood (8) can be written

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

where $\mathbf{\Lambda}_{\sim k}$ and $\boldsymbol{\beta}_{\sim k}$ correspond respectively to $\mathbf{\Lambda}$ and $\boldsymbol{\beta}$ without the k-th row. This maximization is therefore similar to the maximization of (7) and leads to the following estimator of $\mathbf{\Lambda}_{\sim 1} = diag(\lambda_1, \ldots, \lambda_d)$:

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

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

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

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

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

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

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 $\mathbf{\Lambda} = diag(\lambda, \lambda, \ldots, \lambda)$. The maximization of the log-likelihood according to λ leads to the following estimator:

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

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:

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

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:

$$\hat{\lambda}_0^{OLS} = \frac{1}{n^*\beta_0}\sum_{j=1}^{n^*}[y_j^* - \sum_{i=1}^{d}\beta_i\psi_i(\mathbf{x}_j^*)].$$

4.2.2. Prior-based transformation models

As previously discussed, the practician may prefer in some cases to use some particular transformation models suggested by some prior informations.

A generic transformation model including all possible particular transformation models and the corresponding estimators is described below. In the sequel, the subscripts γ_j will be associated with 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 subscripts $\bar{\gamma}_j$ will be associated with regression parameters of the new population which are similar to the original population parameters, *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:

$$Y = \boldsymbol{Q}\boldsymbol{\Lambda}_q + \bar{\boldsymbol{Q}}\mathbf{1}_{p-q} + \epsilon,$$

where:

- $\boldsymbol{\Lambda}_q = \left(\lambda_{\gamma_1}, \ldots, \lambda_{\gamma_q}\right)^t$,
- $\boldsymbol{Q} = \begin{pmatrix} \beta_{\gamma_1}\psi_{\gamma_1}(x_1) & \cdots & \beta_{\gamma_q}\psi_{\gamma_q}(x_1) \\ \vdots & & \vdots \\ \beta_{\gamma_1}\psi_{\gamma_1}(x_n) & \cdots & \beta_{\gamma_q}\psi_{\gamma_q}(x_n) \end{pmatrix}$,
- $\bar{\boldsymbol{Q}} = \begin{pmatrix} \beta_{\bar{\gamma}_1}\psi_{\bar{\gamma}_1}(x_1) & \cdots & \beta_{\bar{\gamma}_q}\psi_{\bar{\gamma}_q}(x_1) \\ \vdots & & \vdots \\ \beta_{\bar{\gamma}_1}\psi_{\bar{\gamma}_1}(x_n) & \cdots & \beta_{\bar{\gamma}_q}\psi_{\bar{\gamma}_q}(x_n) \end{pmatrix}$,
- $\mathbf{1}_{p-q}$ is the unity vector of dimension $p-q$.

Consequently the maximum likelihood estimator of $\boldsymbol{\Lambda}_q$ is

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

4.3. Full and profile likelihood estimation

In this work, a reference regression model on the population P is assumed to be 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 for inferring the new regression model could be disappointing in some cases, particularly when the size n of S is not large too. 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. But, as the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ appear as a product in the regression equation (5) for the sample S^*, the full likelihood estimation of $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ can not be achieved directly and is replaced by a profile likelihood estimation procedure. Starting from a initialization value $\boldsymbol{\beta}^{(0)}$ of $\boldsymbol{\beta}$, the following two steps iteratively alternate until the growth of the model likelihood is lower than a given threshold. At the iteration q:

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

For a given estimation $\hat{\boldsymbol{\Lambda}}^{(q)}$ 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) for the sample S and the log-likelihood of the same model in which the regression function ψ_i are multiplied by $\hat{\lambda}_i^{(q)}$ 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:

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

the estimator of $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\Lambda}}^{(q)}$ is:

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

4.4. Assumption validation and model selection

In regression analysis, there is two indispensable steps: validation of the model assumptions and selection of the regression model.

Assumption validation. An important step in regression analysis is the validation of the linear model assumptions: independence and homoscedasticity of the residuals, linearity of the regression. In this context, several statistical tests have been defined, see for instance [4], and the practician would have to validate the linear model assumptions for the selected regression model as usually. In this paper the regression model for the population P is known and the estimation of the regression model for another population P^* is investigated, and it would be natural to test the equality of both regression models [21]. Unfortunately, this can not be achieved easily since there are too few available data in S^* to efficiently estimate the regression model on

P^*. Nevertheless, the case of equality of the populations P and P^* is considered by the model $M6$, and a model selection procedure, described in the next section, is carried out in place of the regression equality test.

Model selection. The second important step is the selection of the most appropriate model of transformation between the populations P and P^*. We propose to use three well-known criteria. The reader interested in a comparison of the respective performance of these three criteria could refer for instance to [19]. The first criterion is the PRESS criterion [22], representing the sum of squared prediction errors computed on a cross-validation scheme, which is defined by:

$$PRESS = \frac{1}{n^*} \sum_{j=1}^{n^*} ||y_j^* - \widehat{y_j^*}^{-j}||^2$$

where $\widehat{y_j^*}^{-j}$ is the prediction of y_j^* obtained by the regression model estimated without using the j-th individual y_j^* 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 its computation is numerically feasible. Both following penalized likelihood criteria are less computationally heavy. They consist of selecting the models leading to the highest likelihood but penalizing those which have a large number of parameters. The Bayesian Information Criterion (BIC, [23]) 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, [24]) penalized the log-likelihood by 2ν. For all these three criteria, 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 criteria 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 difficult to report here numerical experiments for all existing transformation models, results are presented for only one transformation model: the model M2. Similar results could be obtained for the other transformation models. The true regression model for P is $y = \sin(x) + \sin(2x) + \log(1 + x)$, for $x \in [0, \pi]$, and the specific transformation matrix $\mathbf{\Lambda} = diag(1.5, 2, 2, 2, 2, 2)$ was chosen for generating the regression model of P^*. The size n of the sample S was fixed to 1000. In order to compare the performance of the different transformation models, some observations for population P^* were simulated from its regression

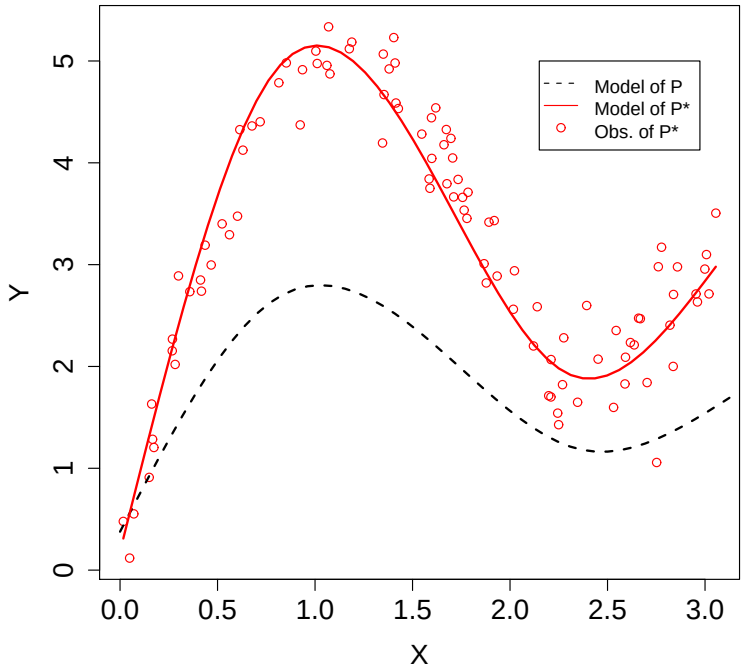

Figure 1: Regression models of the populations P and P^* and simulated observations of population P^*: the model of P was estimated on a basis of cubic Spline functions with 5 degrees of freedom and the model of P^* was obtained from the model of P by multiplying its parameters by (1.5,2,2,2,2,2).

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^*. The values of the three model selection criteria, presented in Section 4.4, were computed for each model to verify

their ability to find 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. The experiments were repeated 50 times in order to average the results.

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 criteria 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 criteria PRESS, BIC and AIC associated to each model. Finally, the sixth column provides the mean square error (MSE) computed on a test dataset different from the learning set. The bold numbers of the table correspond to the "best values" of each column for a given dataset size (let us remind that for the three model selection criteria, 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 modeling the transformation between P and P^* and that corresponds to the truth. The first conclusion is that these three criteria 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

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, AIC and MSE values are per point, and the MSE value was computed on a test dataset.

n^*	Model	PRESS	BIC	AIC	MSE
25	M0	24283.92	16.326	16.033	199.827
	M1	0.131	0.902	0.658	0.109
	M2	**0.109**	**0.669**	**0.571**	**0.094**
	M3	0.128	0.796	0.748	0.119
	M4	0.192	1.241	1.192	0.162
	M5	0.597	2.340	2.291	0.584
50	M0	19196.07	16.209	15.979	51.884
	M1	0.098	0.669	0.478	0.103
	M2	**0.091**	**0.498**	**0.421**	**0.096**
	M3	0.111	0.661	0.623	0.119
	M4	0.157	1.042	1.004	0.163
	M5	0.525	2.220	2.182	0.545
100	M0	1754.953	8.800	8.644	41.239
	M1	0.096	0.614	0.484	0.091
	M2	**0.093**	**0.509**	**0.456**	**0.089**
	M3	0.115	0.699	0.673	0.109
	M4	0.172	1.128	1.102	0.157
	M5	0.455	2.072	2.046	0.511
250	M0	522.120	5.512	5.427	24.329
	M1	0.090	0.504	0.434	0.090
	M2	**0.089**	**0.450**	**0.422**	**0.089**
	M3	0.116	0.704	0.690	0.111
	M4	0.172	1.135	1.121	0.161
	M5	0.467	2.089	2.075	0.534
500	M0	270.574	5.034	5.004	6.633
	M1	0.092	0.495	0.453	0.091
	M2	**0.091**	**0.463**	**0.446**	**0.090**
	M3	0.116	0.698	0.689	0.113
	M4	0.167	1.090	1.082	0.155
	M5	0.463	2.075	2.067	0.501
1000	M0	184.00	4.669	4.618	3.519
	M1	0.089	0.450	0.425	0.091
	M2	**0.089**	**0.432**	**0.422**	**0.090**
	M3	0.113	0.669	0.665	0.112
	M4	0.168	1.093	1.088	0.156
	M5	0.453	2.051	2.046	0.501

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. 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 the present paper, are able to correctly estimate a transformation between two populations with non-linear regression models and that even in situations where the number of observations of P^* is limited. This study has also highlighted that either the cross-validated PRESS criterion and information criteria 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 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

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

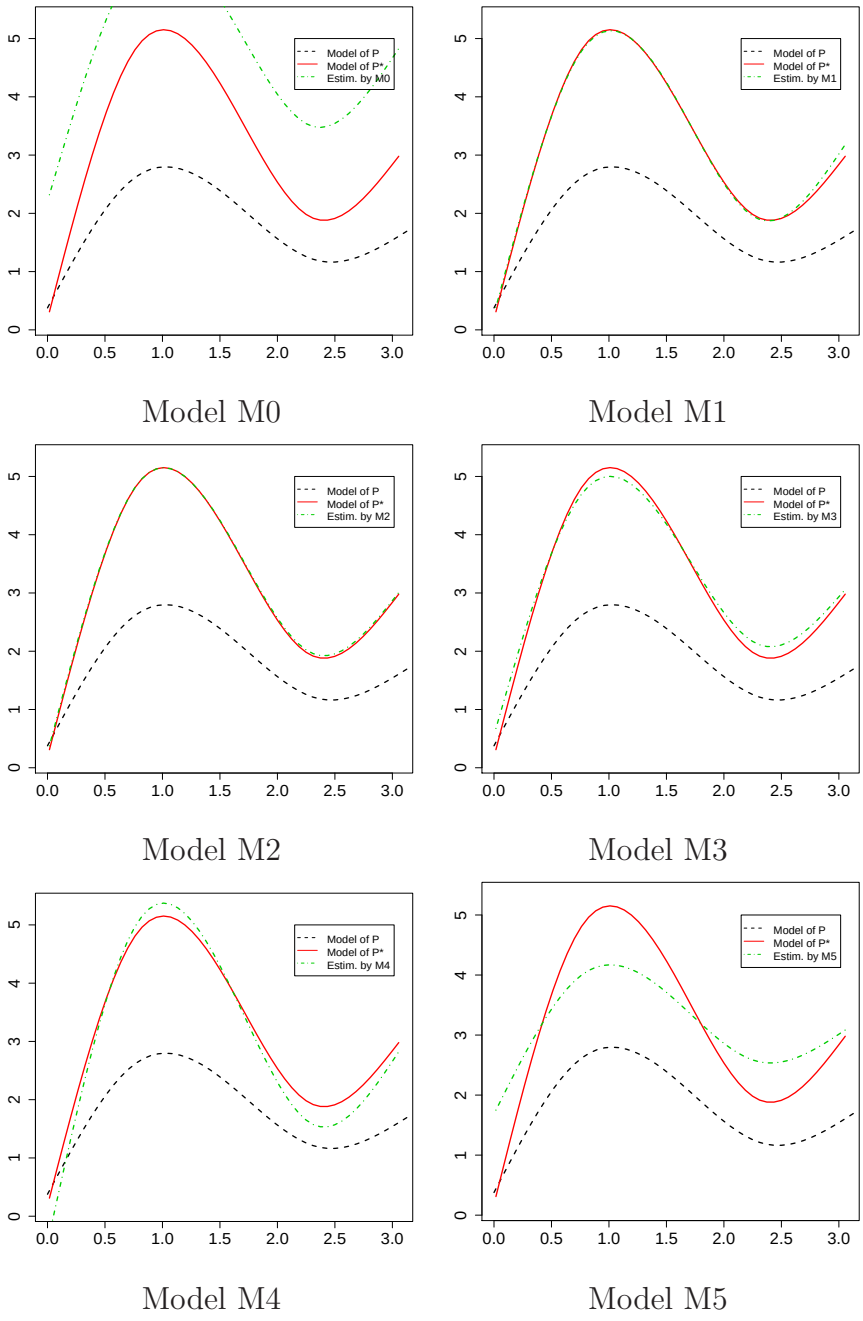

Figure 2: Parameter estimation with the different linear transformation models on data simulated according to the transformation model M2 on a basis of cubic Spline functions. These estimations were computed with a dataset of 100 observations.

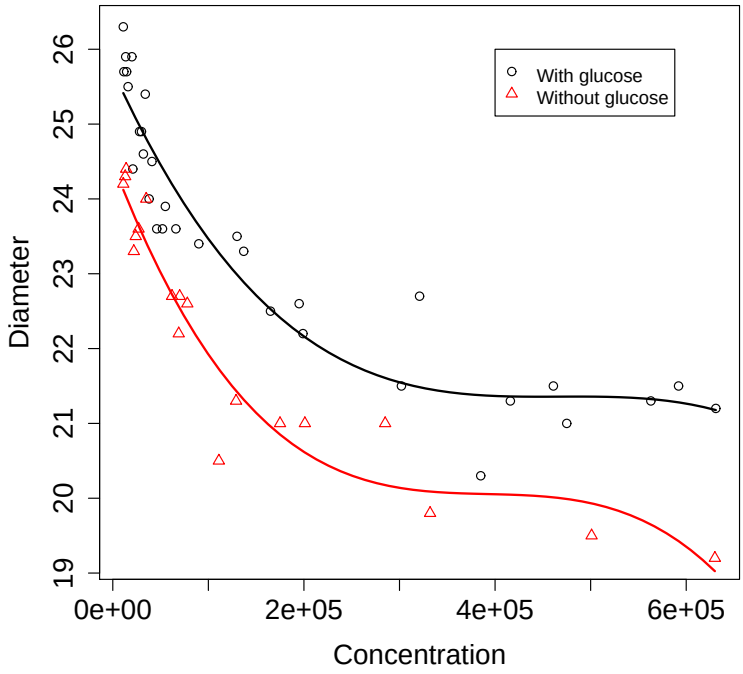

Figure 3: The *hellung* dataset: diameter *vs.* concentration for *Tetrahymena* cells.

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 (32 observations) whereas cells without glucose will be considered as coming from population P^* (between 11 to 19 observations).

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. 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 MSE value on the whole P^* dataset are computed for these 4 different sizes of learning dataset.

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

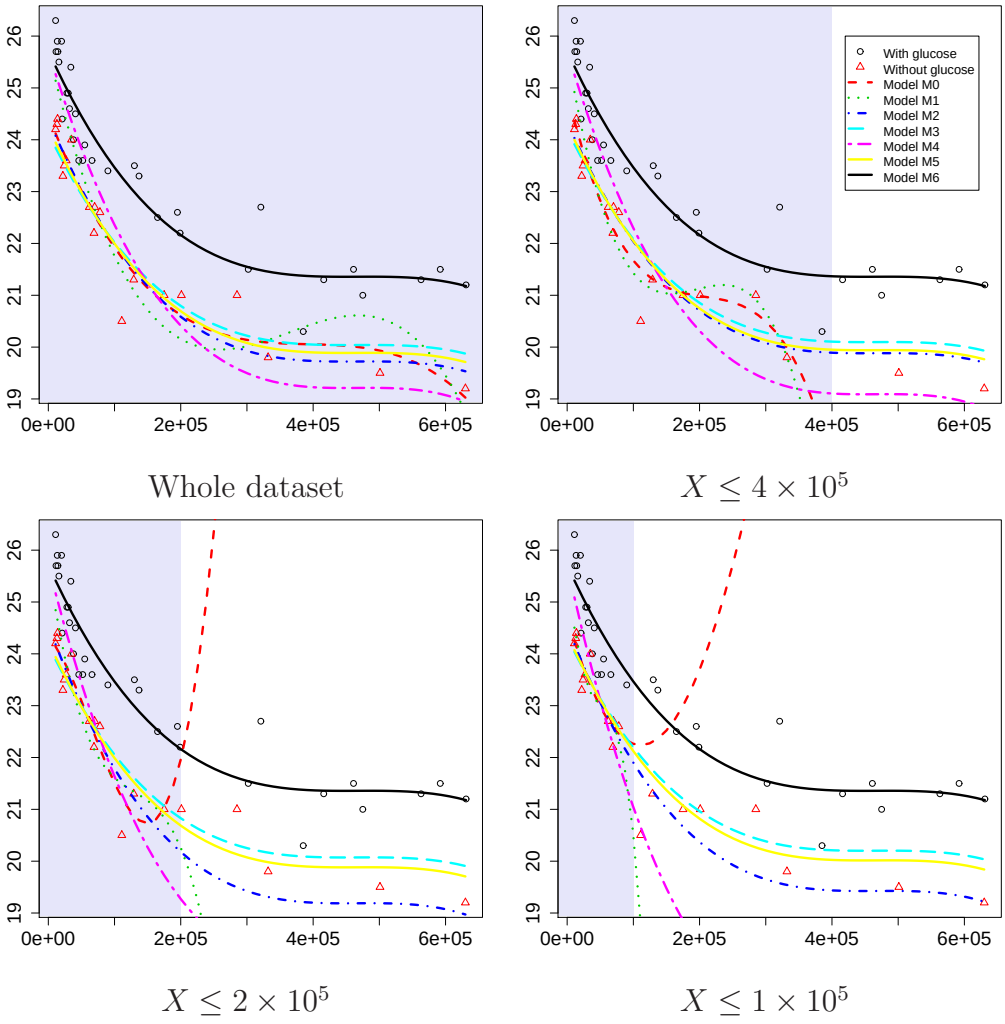

Figure 4: Effect of the learning set size on the prediction ability of the studied regression methods for the *hellung* dataset. The blue zones correspond to the parts of the observations of P^* used for learning the models.

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 unstable 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 the MSE 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. These two models consider a shift of the intercept, which confirms the guess that we can have by examining graphically the dataset, and moreover by quantifying this shift. To conclude, this study has shown that the adaptive linear models can be successfully applied to real data for transferring a knowledge on a reference population (here the cells with glucose) to a new population (here the

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)	0.897	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 4: Effect of the learning set size on the MSE value 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.

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

cells without 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,...).

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 modeling 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 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 [25] 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 modeling 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.

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

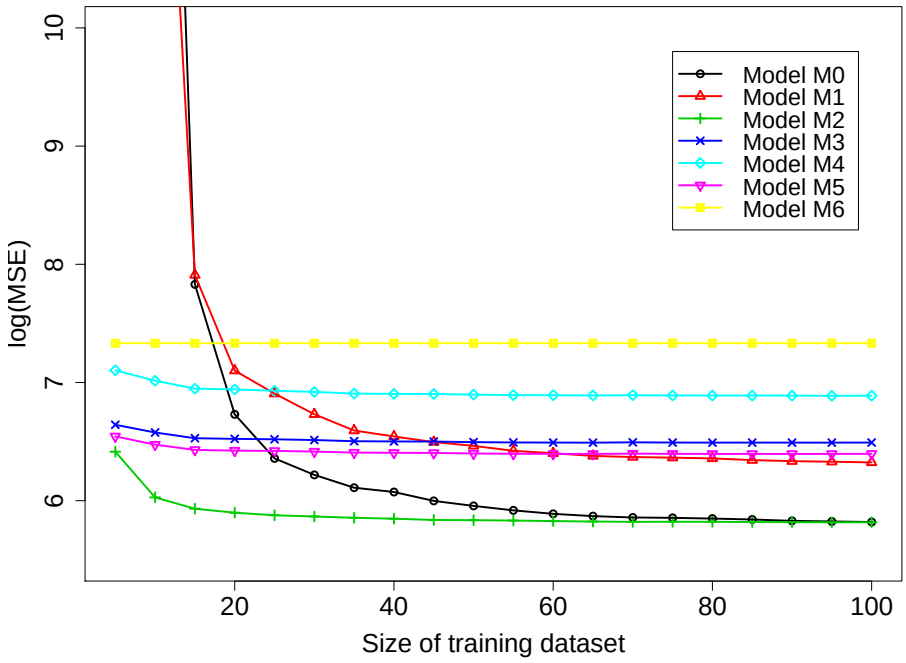

Figure 5: MSE results for the Birmingham-San Jose data.

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 for averaging the results. For each adaptive linear model, the PRESS criterion and the MSE were computed, by using the selected sample for PRESS and the whole San Jose dataset for MSE.

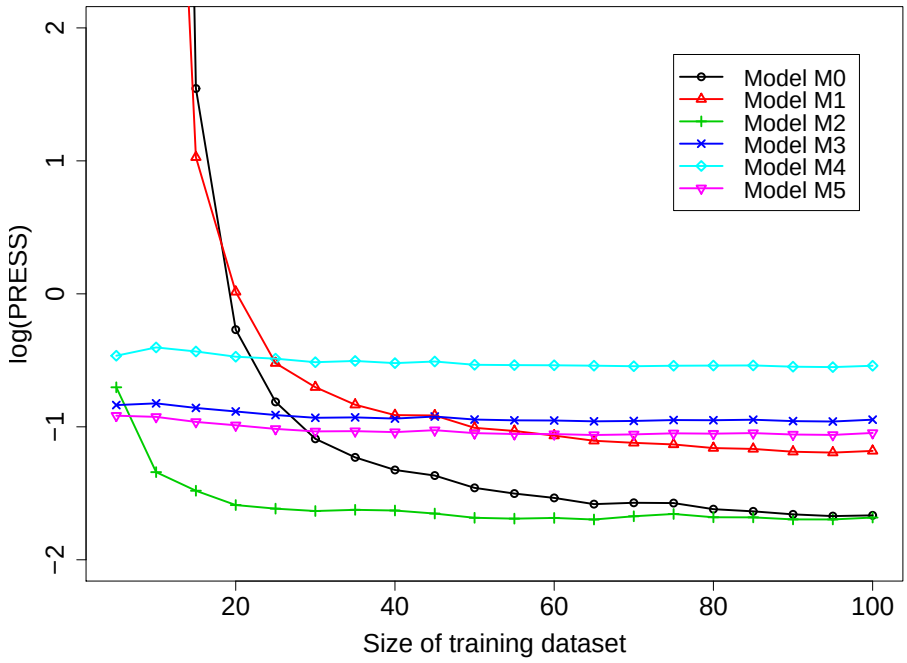

Figure 6: PRESS criterion for the Birmingham-San Jose data.

Experimental results. Figure 5 shows the logarithm of the MSE 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 modeling the housing market of San Jose since it obtains a not satisfying MSE value. Let us notice that the curve corresponding to the MSE 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 MSE) 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 MSE 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 prediction 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 though the difference is not so big for large learning datasets. Figure 6 shows 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 MSE 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

Table 5: MSE results for the Birmingham-San Jose data.

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

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

6. Discussion

Before each statistical analysis, the indispensable collect of data is often an expensive step. Even though 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 shows how it is possible to adapt a regression model from a given situation to another new one, and thus to save an expensive new collect of data.

In this perspective, a family of adaptive linear models has been introduced and, since they are more parsimonious than a complete regression model, they need only few samples for providing satisfying estimation of the new regression model. To summarize, the main interest of this work arises when the sample size for the new population is too small to efficiently estimate a regression model by the usual OLS procedure without using information known for the reference population. The conducted experiments have shown that the proposed adaptive linear models are able to successfully transfer a knowledge on a reference population to another population even with very few observations. 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 necessary 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. This could be the subject of a further study and of a collaboration with the economists who provided these data. Another interesting perspective of this work concerns the presence of correlation between the covariates. Indeed, if the correlation between variables is different from one population to the other, it will be necessary to consider different transformation parameters for these variables.

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] S. R. Searle, Linear models, John Wiley & Sons Inc., New York, 1971.

[2] P. McCullagh, J. A. Nelder, Generalized linear models, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1983.

[3] D. Ratkowsky, Handbooks of nonlinear regression models, Chapman & Hall, London, 1990.

[4] N. Draper, H. Smith, Applied regression analysis, 3rd Edition, John Wiley & Sons Inc., New York, 1998.

[5] H. Shimodaira, Improving predictive inference under covariate shift by weighting the log-likelihood function, J. Statist. Plann. Inference 90 (2) (2000) 227–244.

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

[7] M. Sugiyama, K.-R. Müller, Input-dependent estimation of generalization error under covariate shift, Statistics & Decisions 23 (2005) 249–279.

589 [8] M. Sugiyama, K. M. Müller, K-R., Covariate shift adaptation by impor-
590 tance weighted cross validation, Journal of Machine Learning Research
591 8 (2007) 985–1005.

592 [9] M. Sugiyama, Active learning in approximately linear regression based
593 on conditional expectation of generalization error, Journal of Machine
594 Learning Research 7 (2006) 141–166.

595 [10] C. Biernacki, F. Beninel, V. Bretagnolle, A generalized discriminant rule
596 when training population and test population differ on their descriptive
597 parameters, Biometrics 58 (2) (2002) 387–397.

598 [11] J. Jacques, C. Biernacki, Extension of model-based classification for
599 binary data when training and test populations differ, Journal of Applied
600 Statistics (2010) in press.

601 [12] F. Beninel, C. Biernacki, Modèles d'extension de la régression logis-
602 tique, Revue des Nouvelles Technologies de l'Information, Data Mining
603 et apprentissage statistique : application en assurance, banque et mar-
604 keting (A1) (2007) 207–218.

605 [13] N. Feudale, N. Woody, H. Tan, D. Kell, J. Maddock, Heginbothom,
606 J. M., Magee, Transfer of multivariate calibration models: a review,
607 Chemometrics and Intelligent Laboratory System 64 (2002) 181–192.

608 [14] Y. Wang, V. D., K. B., Multivariate instrument standardization, Ana-
609 lytical chemistry 63 (23) (1991) 2750–2756.

610 [15] R. Goodacre, E. Timmins, A. Jones, D. Kell, J. Maddock, H. M.,

611 J. Magee, On mass spectrometer instrument standardization and inter-
612 laboratory calibration transfer using neural networks, Analytica Chim-
613 ica Acta 348 (1) (1997) 511–532.

614 [16] K. Bertness, R. Hickernell, S. Hays, D. Christensen, Noise reduction in
615 optical in situ measurements for molecular beam epitaxy by substrate
616 wobble normalization, Journal of Vacuum Science and Technology B
617 16 (3) (1998) 1492–1497.

618 [17] K. Tobin, T. Karnowski, A. L., R. Ferrell, J. Goddard, F. Lakhani,
619 Content-based image retrieval for semiconductor process characteriza-
620 tion, Journal on Applied Signal Processing 1 (2002) 704–713.

621 [18] C. Bishop, Pattern recognition and machine learning, Information Sci-
622 ence and Statistics, Springer, New York, 2006.

623 [19] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learn-
624 ing, Springer Series in Statistics, Springer-Verlag, New York, 2001.

625 [20] S. Mallat, A wavelet tour of signal processing, 2nd Edition, Academic
626 Press, 1999.

627 [21] G. Chow, Tests of equality between sets of coefficients in two linear
628 regressions, Econometrica 28 (1960) 591–605.

629 [22] D. M. Allen, The relationship between variable selection and data aug-
630 mentation and a method for prediction, Technometrics 16 (1974) 125–
631 127.

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

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

[25] W. D. United States Department of Commerce, Bureau of the Census, American housing survey, 1984: Msa file, Ann Arbor, MI: Inter- university Consortium for Political and Social Research.