Adaptive Linear Models for Regression: improving prediction when population has changed
Charles Bouveyron, Julien Jacques

To cite this version:

HAL Id: hal-00305987
https://hal.archives-ouvertes.fr/hal-00305987v2
Submitted on 15 Dec 2008 (v2), last revised 30 Mar 2010 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Adaptive linear models for regression

Charles Bouveyron
SAMOS-MATISSE, CES, University Paris I (Panthéon-Sorbonne), Paris, France.

Julien Jacques
Laboratoire Paul Painlevé, UMR CNRS 8524, University Lille I, Lille, France.

Summary. 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 $X$ 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 a new sample by estimating a transformation between the original regression function $f(X)$ and the new one $f'(X)$. The main interest of this work is that a model for the new population can be build with only few observations. The efficiency of this strategy 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 (Searle (1971)) and generalized linear model (McCullagh and Nelder (1983)) which can be seen as parametric models, and non linear models which are mostly non-parametric models (Ratkowsky (1990)). See Draper and Smith (1998) 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, 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? 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 numerous housing descriptive variables. This regression model has been estimated by using a large sample of observations. To conquer new markets, this company plans to open several agency in the West coast, and would use its regression model without having to spend a lot of time and money in collecting or buying new housing market data for this region. Considering that the link between housing descriptive variables and housing price is probably not the same for the West and East coasts, but that it is also not probably completely different, we will consider a set of transformation models between both West
and East coasts regression models. The focus of this paper is to demonstrate how the knowledge on a reference population can be transferred to a new population, by connecting both regression models. Moreover, the exhibition of a link between both populations can 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 could be 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 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 never 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. Several recent works have contributed to analyze this context (Shimodaira (2000); Storkey and Sugiyama (2007); Sugiyama and Müller (2005, 2007); Sugiyama (2006)). However, most of them need to know the probability distribution of the new data or at least an estimation of this probability distribution and this 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 similar problem was studied in Biernacki et al. (2002) on quantitative variables and in Jacques and Biernacki (2007) 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 Beninel and Biernacki (2007). Finally, some works cover the problematic of knowledge transfer in specific industrial contexts. In the field of Chemometrics, Fed nale et al. (2002) 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, Wang et al. (1991), and the neural network based nonlinear transformation, Goodacre et al. (1997). Several works, Bertness et al. (1998); Tobin et al. (2002), have also considered this problem in the field of semiconductor industry.

The structure of the paper is 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. Usual estimation and model selection procedures are discussed in Section 4. Finally, Section 5 provides a simulation study in a spline regression context and two real applications including the modelling of the housing market in different U.S. cities.

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 \( S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \), arising from a population \( P \), are assumed to be independent and identically distributed samples of a couple of variables \((X, Y)\) with an unknown distribution. The observations \( x_i \) are values of the deterministic explanatory variable \( X = (X^{(1)}, \ldots, X^{(p)})^t \in \mathbb{R}^p \) 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 \( X \) (or covariate) and the response variable \( Y \) (or dependent variable). Both standard parametric and non-parametric regression approaches consider with the following model:

\[
Y = f(X, \beta) + \epsilon, \tag{1}
\]

where the residuals \( \epsilon \sim N(0, \sigma^2) \) are independent and where \( \beta \) is the vector of regression parameters. This model is equivalent to the distributional assumption that:

\[
Y|X \sim N(f(X, \beta), \sigma^2), \tag{2}
\]

where the regression function \( f(x, \beta) \) is defined as the conditional expectation \( E[Y|X = x] \). In particular, parametric regression achieves this connection by assuming a specific form for \( f(x, \beta) \). The most common model is the linear form (cf. Bishop (2006)):

\[
f(x, \beta) = \sum_{i=0}^{d} \beta_i \psi_i(x), \tag{3}
\]

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

\[
\psi_i : \mathbb{R}^p \rightarrow \mathbb{R},
\]

which can be for instance identity, polynomial, splines functions (Hastie et al. (2001)) or wavelets (Mallat (1999)). We refer to Draper and Smith (1998) for a general survey. Let notice that the usual linear regression occurs when \( d = p \) and \( \psi_i(x) = x^{(i)} \) for \( i = 1, \ldots, d \).

The regression function (2) can be also written in its matricial form as follows:

\[
f(x, \beta) = \beta^t \Psi(x), \tag{4}
\]

where \( \Psi(x) = (1, \psi_1(x), \ldots, \psi_d(x))^t \).

2.2. How to adapt a regression model to another population?

Let assume now 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 \( S^* = \{(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)\} \), 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 U.S. housing market application), temporal or other. However, the nature of both populations have to be similar to match the purpose of this paper. The new regression model on \( P^* \) can be written:

\[
Y^*|X^* \sim N(f^*(X^*, \beta^*), \sigma^*)^2, \tag{5}
\]
with
\[ f^*(x^*, \beta^*) = \sum_{i=0}^{d^*} \beta^*_i \psi^*_i(x^*) = \beta^* \Psi(x^*). \]

The aim of this work is therefore to define a link between both 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 then 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. The transformation model

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

Assumption (A1) First, we postulate that the number of regression basis functions and the regression basis functions themselves are the same for both regression models \((d^* = d)\) and \(\psi^*_i = \psi_i \) \), which is natural since the variables are identical in both populations. The regression function of the population \( P^* \) is then:
\[ f^*(x^*, \beta^*) = \beta^* \Psi(x^*) \]

Assumption (A2) Second, we assume that the transformation between \( f \) and \( f^* \) applies only on the regression parameters. We define by \( \Lambda \) the transformation matrix between the regression parameters \( \beta \) and \( \beta^* \):
\[ \beta^* = \Lambda \beta, \]
what leads to the following expression of \( f^* \):
\[ f^*(x^*, \Lambda \beta) = (\Lambda \beta)^t \Psi(x^*). \]

(5)

Given that the number of free parameters to estimate in the transformation matrix \( \Lambda \) is \((d+1) \times (d+1)\) and that the one for learning a new regression model directly from the sample \( S^* \) is \((d+1)\), the transformation model (5) 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 \( d \).

Assumption (A3) Third, we assume 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, for each \( i = 0, \ldots, d, \) the regression parameter \( \beta^*_i \) only depends on the regression parameter \( \beta_i \) and the matrix \( \Lambda \) is diagonal. The transformation can be finally written in term of the regression parameters of both models as follows:
\[ \beta^*_i = \lambda_i \beta_i \quad \forall i = 0, \ldots, d, \]

(6)

where \( \lambda_i \in \mathbb{R} \) is the \( i \)-th diagonal element of \( \Lambda \).
3.2. A family of transformation models

Since our aim is to be able 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 $\lambda_i$ to be equal to 1 (in this case the regression parameters $\beta_i^*$ are equal to $\beta_i$). Second, among the parameters $\lambda_i$ which are not equal to 1, we allow some of them to be equal ($\lambda_i = \lambda_j$ for given $0 \leq i < j \leq d$). The number of possible models obtained with such a strategy is then very large: $\sum_{m=0}^{d+1} \binom{d+1}{m} \times \left( 1 + \sum_{l=2}^{m} \binom{l}{m} \right)$. These models are declined below into two families: general transformation models and specific transformation models.

3.2.1. General transformation models

We propose in the present section a family of 7 transformation models, named further adaptive linear models, ranging from the most complex model (hereafter $M_0$) to the simplest one (hereafter $M_6$):

- Model $M_0$: $\beta_0^* = \lambda_0 \beta_0$ 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 $P^*$. It is equivalent to learning a new regression model from the sample $S^*$ since there is no constraint on the $d+1$ parameters $\beta_i^* (i = 0, ..., d)$ and the number of free parameters in $\Lambda$ is $d + 1$ as well.

- Model $M_1$: $\beta_0^* = \beta_0$ and $\beta_i^* = \lambda_i \beta_i$, for $i = 1, ..., d$. This model assumes that both regression models have the same intercept $\beta_0$.

- Model $M_2$: $\beta_0^* = \lambda_0 \beta_0$ and $\beta_i^* = \lambda \beta_i$, for $i = 1, ..., d$. This model assumes that the intercept of both regression models differ by the scalar $\lambda_0$ and all the other regression parameters differ by the same scalar $\lambda$.

- Model $M_3$: $\beta_0^* = \lambda \beta_0$ and $\beta_i^* = \lambda_i \beta_i$, for $i = 1, ..., d$. This model assumes that all the regression parameters of both regression models differ by the same scalar $\lambda$.

- Model $M_4$: $\beta_0^* = \beta_0$ and $\beta_i^* = \lambda \beta_i$, for $i = 1, ..., d$. This model assumes that both regression models have the same intercept $\beta_0$ and all the other regression parameters differ by the same scalar $\lambda$.

- Model $M_5$: $\beta_0^* = \lambda_0 \beta_0$ and $\beta_i^* = \beta_i$, for $i = 1, ..., d$. This model assumes that both regression models have the same parameters except the intercept.

- Model $M_6$: $\beta_0^* = \beta_0$ 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. 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.
Table 1. Complexity (number of parameters $\nu$) of the transformation models.

<table>
<thead>
<tr>
<th>Model</th>
<th>$M_0$</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_3$</th>
<th>$M_4$</th>
<th>$M_5$</th>
<th>$M_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0^*$ is assumed to be</td>
<td>$\lambda_0 \beta_0$</td>
<td>$\lambda_0 \beta_0$</td>
<td>$\lambda_0 \beta_0$</td>
<td>$\beta_0$</td>
<td>$\lambda_0 \beta_0$</td>
<td>$\beta_0$</td>
<td></td>
</tr>
<tr>
<td>$\beta_i^*$ is assumed to be</td>
<td>$\lambda_i \beta_i$</td>
<td>$\lambda_i \beta_i$</td>
<td>$\lambda_i \beta_i$</td>
<td>$\lambda_i \beta_i$</td>
<td>$\lambda_i \beta_i$</td>
<td>$\beta_i$</td>
<td></td>
</tr>
<tr>
<td>Nb. of parameters $\nu$</td>
<td>$d+1$</td>
<td>$d$</td>
<td>$2$</td>
<td>$1$</td>
<td>$1$</td>
<td>$1$</td>
<td>$0$</td>
</tr>
</tbody>
</table>

3.2.2. Specific transformation models

Although only seven pragmatic transformation models have been presented in the previous section, some other more specific transformation models could be considered as well, whose complexity (in number of parameters) will be intermediate between this one of $M_1$, $d$, and this one of $M_2$, 2. The practitioner 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 $\lambda_i$ for given $i \in \{1, \ldots, d\}$. For instance, let consider the specific transformation matrix $\Lambda = \text{diag}(\lambda_0, \lambda_1, \ldots, \lambda)$ where $\text{diag}(\lambda_0, \lambda_1, \lambda, \ldots, \lambda)$ 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 $\beta_i$, $i = 2, \ldots, d$ are transformed in the same manner whereas the intercept and $\beta_1$ 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 $P^*$. The regression parameters of $P^*$ are then obtained by plug-in. The Ordinary Least Square method is used, but we present in this paper the equivalent maximum likelihood estimation method for introducing penalized likelihood model selection criteria later.

4.1. Estimation of the regression parameters

Let consider a data set of inputs $x = \{x_1, \ldots, x_n\}$ with corresponding response values grouping in a column vector $y = (y_1, \ldots, y_n)^t$. Under the assumptions of the model (2), the log-likelihood of $y$ given $x$, $\beta$ and $\sigma^2$ is:

$$\ln l(y; x, \beta, \sigma^2) = -n \ln(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^t \Psi(x_i))^2. \quad (7)$$

Maximizing the log-likelihood according to $\beta$ is equivalent to minimizing $\sum_{i=1}^{n} (y_i - \beta^t \Psi(x_i))^2$ and thus the maximum likelihood estimator is equivalent to the ordinary least square estimator. The gradient of the log-likelihood function taken with respect to $\beta$ has the following form:

$$\nabla \ln l(y; x, \beta, \sigma^2) = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta^t \Psi(x_i)) \Psi(x_i)^t,$$

and setting this gradient to zero gives:

$$\sum_{i=1}^{n} y_i \Psi(x_i)^t = \beta^t \left( \sum_{i=1}^{n} \Psi(x_i) \Psi(x_i)^t \right).$$
Solving this equation according to $\beta$ leads to the well known ordinary least square (OLS) estimator for $\beta$:

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

where $\Psi$ is a $(n) \times (d + 1)$ matrix formed by the row vector $\Psi(x_i)^t (1 \leq i \leq n)$.

### 4.2. Estimation of the transformation parameters

For this step it is assumed that $\beta$ is known (in fact it is estimated by the previous step). 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 $x^* = \{x^*_1, \ldots, x^*_n\}$ drawn from $P^*$ with corresponding response values $y^* = (y^*_1, \ldots, y^*_n)^t$.

By replacing $\beta^* = \Lambda \beta$ in (7), the log-likelihood is:

$$\ln l(y^*; x^*, \Lambda, \sigma^2) = -n^* \ln(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n^*} (y^*_i - \beta^t \Lambda^t \Psi(x^*_i))^2. \tag{8}$$

This log-likelihood must be maximized according to the transformation matrix $\Lambda$, what leads to the Ordinary Least Square estimator:

$$\hat{\Lambda}^{OLS} = \arg\min_{\Lambda \in D} \sum_{i=1}^{n^*} (y^*_i - (\Lambda \beta)^t \Psi(x^*_i))^2. \tag{9}$$

where $D$ is a set of diagonal matrices depending on the model of transformation at hand. For instance, with the model $M_3$, this set is $D = \{\lambda I_{d+1}, \lambda \in \mathbb{R}\}$ where $I_{d+1}$ is the identity matrix of size $d + 1$. Exact formulation of the Least Square estimator $\hat{\Lambda}^{OLS}$ for transformation model $M_1$ to $M_5$ are derived in Appendix A.

**Specific transformation models**

As previously discussed, in some specific cases the practitioner could prefer 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 indices $\gamma_j$ will be associated to regression parameters of the new population to estimate using the relation $\beta^*_j = \lambda_{\gamma_j} \beta_{\gamma_j}$ with $j = 1, \ldots, q$ and $\gamma_j \in \{0, \ldots, d\}$. In the same manner, the indices $\bar{\gamma}_j$ will be associated to regression parameters of the new population which are similar to the original population ones, i.e. $\beta^*_j = \beta_{\bar{\gamma}_j}$ with $j = 1, \ldots, p - q$ and $\bar{\gamma}_j \in \{0, \ldots, d\}$. The regression model for the new population can be written as follows:

$$Y = QA_q + Q1_{p-q} + \epsilon,$$

where:

- $A_q = (\lambda_{\gamma_1}, \ldots, \lambda_{\gamma_q})^t$,

- $Q = \begin{pmatrix} \beta_{\gamma_1} \psi_{\gamma_1}(x_1) & \cdots & \beta_{\gamma_q} \psi_{\gamma_q}(x_1) \\ \vdots & \ddots & \vdots \\ \beta_{\gamma_1} \psi_{\gamma_1}(x_n) & \cdots & \beta_{\gamma_q} \psi_{\gamma_q}(x_n) \end{pmatrix}$,
• $Q = \begin{pmatrix}
\beta_1 \psi_{\gamma_1}(x_1) & \cdots & \beta_q \psi_{\gamma_q}(x_1) \\
\vdots & & \vdots \\
\beta_1 \psi_{\gamma_1}(x_n) & \cdots & \beta_q \psi_{\gamma_q}(x_n)
\end{pmatrix}$,

• $1_{p-q}$ is the unity vector of dimension $p - q$.

Consequently the maximum likelihood estimator of $\Lambda_q$ is

$$\hat{\Lambda}_q^{OLS} = (Q'Q)^{-1} Q' (y - Q1_{p-q}).$$

### 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 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 $\beta$ as well. As the parameters $\beta$ and $\Lambda$ appear as a product in the regression equation (5) for the sample $S^*$, it is not possible to jointly estimate it directly. An alternate iterative algorithm is proposed here to jointly estimate the regression parameter $\beta$ and the transformation matrix $\Lambda$. Starting from an intialization value $\hat{\beta}^{(0)}$ of $\beta$, the following two steps are alternate iteratively until the growth of the model likelihood is lower than a given threshold. At the iteration $q$:

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

(b) Compute the estimation $\hat{\beta}^{(q)}$ of $\beta$ given the estimation of $\hat{\Lambda}^{(q)}$ obtained in the previous step.

For a given estimation $\hat{\Lambda}^{(q)}$ of $\Lambda = diag(\lambda_0, \lambda_1, \ldots, \lambda_d)$, the estimation of $\beta$ 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 $\psi_i$ are multiplied by $\hat{\lambda}_i^{(q)}$ for the sample $S^*$. By introducing $\tilde{y} = (y_1, \ldots, y_n, y_1^*, \ldots, y_n^*)^t$ and $\Psi$ the $(n + n^*) \times (d + 1)$ matrix defined as follows:

$$\tilde{\Psi} = \begin{pmatrix}
\psi_0(x_1) & \cdots & \psi_0(x_1) \\
\vdots & & \vdots \\
\psi_0(x_n) & \cdots & \psi_0(x_n) \\
\hat{\lambda}_0 \psi_0(x_1^*) & \cdots & \hat{\lambda}_d \psi_0(x_1^*) \\
\vdots & & \vdots \\
\hat{\lambda}_0 \psi_0(x_n^*) & \cdots & \hat{\lambda}_d \psi_0(x_n^*)
\end{pmatrix},$$

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

$$\hat{\beta}^{OLS} = (\tilde{\Psi}' \tilde{\Psi})^{-1} \tilde{\Psi}' \tilde{y}.$$
4.4. Assumption validation and model selection

In regression analysis, there are 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, statistical tests have been defined, see for instance Draper and Smith (1998), and the practitioner would have to validate the linear model assumptions for the selected regression model as in the usual case. In this paper the regression model for the population $P$ is well known and the estimation of this one for another population $P^*$ is investigated. It could be natural to firstly test the equality of both regression models (Chow (1960)). 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, as the identity transformation is included in the considered family of model, the situation of equality of both regression models is allowed.

**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 performances of these three criteria could refer for instance to Hastie et al. (2001). The first criterion is the PRESS criterion, Allen (1974), which represents the sum of squared prediction errors computed on a cross-validation scheme, and is defined by:

$$PRESS = \sum_{j=1}^{n} \| y^*_j - \hat{y}^*_j \|_2^2$$

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, and we encourage its use when its computation is numerically feasible. Both following penalized likelihood criteria are less computationally heavy. They 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, Schwarz (1978)) is defined by:

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

where $\ell$ is the maximum likelihood value and $\nu$ is the number of estimated parameters (see Table 1). With the same notations, the Akaike Information Criterion (AIC, Akaike (1974)) is defined by:

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

For 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.
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 $(1.5, 2, 2, 2, 2)$.

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 $\Lambda$. 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 $\Lambda = \text{diag}(1.5, 2, 2, 2, 2)$ 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 from its regression model. These observations were simulated with an additive Gaussian noise $\epsilon \sim N(0, 0.3)$. 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 $\hat{\Lambda}$) 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 criteria, presented in Section 4.4, were computed for each model 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 $M_0$, $M_1$, $M_2$, $M_3$, $M_4$ and $M_5$ 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 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 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 $M_2$ 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 criteria are well suited to select the transformation model in such a case. On the other hand, it can be noticed that the model $M_0$, 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 $M_1$ to $M_5$ are less sensitive. Furthermore, the model $M_0$ 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 $M_2$ 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 $M_0$ estimation is very far away from the actual model and the models $M_1$, $M_2$ and $M_3$ 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 $M_5$ 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 estimate correctly 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 demonstrated 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.
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 $P^*$. 

<table>
<thead>
<tr>
<th>$n^*$</th>
<th>Model</th>
<th>PRESS</th>
<th>BIC</th>
<th>AIC</th>
<th>RSS</th>
<th>Prm. diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>M0</td>
<td>24283.92</td>
<td>16.326</td>
<td>16.033</td>
<td>199.827</td>
<td>1312.998</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>0.131</td>
<td>0.902</td>
<td>0.658</td>
<td>0.109</td>
<td>2.142</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>0.109</td>
<td>0.669</td>
<td>0.571</td>
<td>0.094</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>0.128</td>
<td>0.796</td>
<td>0.748</td>
<td>0.119</td>
<td>0.528</td>
</tr>
<tr>
<td></td>
<td>M4</td>
<td>0.192</td>
<td>1.241</td>
<td>1.192</td>
<td>0.162</td>
<td>1.255</td>
</tr>
<tr>
<td></td>
<td>M5</td>
<td>0.597</td>
<td>2.340</td>
<td>2.291</td>
<td>0.584</td>
<td>10.348</td>
</tr>
<tr>
<td>50</td>
<td>M0</td>
<td>19196.07</td>
<td>16.209</td>
<td>15.979</td>
<td>51.884</td>
<td>674.779</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>0.098</td>
<td>0.669</td>
<td>0.478</td>
<td>0.103</td>
<td>1.770</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>0.091</td>
<td>0.498</td>
<td>0.421</td>
<td>0.096</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>0.111</td>
<td>0.661</td>
<td>0.623</td>
<td>0.119</td>
<td>0.548</td>
</tr>
<tr>
<td></td>
<td>M4</td>
<td>0.157</td>
<td>1.042</td>
<td>1.004</td>
<td>0.163</td>
<td>1.211</td>
</tr>
<tr>
<td></td>
<td>M5</td>
<td>0.525</td>
<td>2.220</td>
<td>2.182</td>
<td>0.545</td>
<td>10.639</td>
</tr>
<tr>
<td>100</td>
<td>M0</td>
<td>1754.953</td>
<td>8.800</td>
<td>8.644</td>
<td>41.239</td>
<td>734.003</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>0.096</td>
<td>0.614</td>
<td>0.484</td>
<td>0.091</td>
<td>1.510</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>0.093</td>
<td>0.509</td>
<td>0.456</td>
<td>0.089</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>0.115</td>
<td>0.661</td>
<td>0.673</td>
<td>0.109</td>
<td>0.425</td>
</tr>
<tr>
<td></td>
<td>M4</td>
<td>0.172</td>
<td>1.128</td>
<td>1.102</td>
<td>0.157</td>
<td>1.002</td>
</tr>
<tr>
<td></td>
<td>M5</td>
<td>0.455</td>
<td>2.072</td>
<td>2.046</td>
<td>0.511</td>
<td>7.141</td>
</tr>
<tr>
<td>250</td>
<td>M0</td>
<td>522.120</td>
<td>5.512</td>
<td>5.427</td>
<td>24.329</td>
<td>466.621</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>0.090</td>
<td>0.504</td>
<td>0.434</td>
<td>0.090</td>
<td>1.404</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>0.089</td>
<td>0.450</td>
<td>0.422</td>
<td>0.089</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>0.116</td>
<td>0.704</td>
<td>0.690</td>
<td>0.111</td>
<td>0.474</td>
</tr>
<tr>
<td></td>
<td>M4</td>
<td>0.172</td>
<td>1.135</td>
<td>1.121</td>
<td>0.161</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>M5</td>
<td>0.467</td>
<td>2.089</td>
<td>2.075</td>
<td>0.534</td>
<td>7.33</td>
</tr>
<tr>
<td>500</td>
<td>M0</td>
<td>270.574</td>
<td>5.034</td>
<td>5.004</td>
<td>6.633</td>
<td>272.080</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>0.092</td>
<td>0.495</td>
<td>0.453</td>
<td>0.091</td>
<td>1.347</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>0.091</td>
<td>0.463</td>
<td>0.446</td>
<td>0.090</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>0.116</td>
<td>0.698</td>
<td>0.689</td>
<td>0.113</td>
<td>0.427</td>
</tr>
<tr>
<td></td>
<td>M4</td>
<td>0.167</td>
<td>1.090</td>
<td>1.082</td>
<td>0.155</td>
<td>0.926</td>
</tr>
<tr>
<td></td>
<td>M5</td>
<td>0.463</td>
<td>2.075</td>
<td>2.067</td>
<td>0.501</td>
<td>7.122</td>
</tr>
<tr>
<td>1000</td>
<td>M0</td>
<td>184.00</td>
<td>4.669</td>
<td>4.618</td>
<td>3.519</td>
<td>121.248</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>0.089</td>
<td>0.450</td>
<td>0.425</td>
<td>0.091</td>
<td>1.368</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td>0.089</td>
<td>0.432</td>
<td>0.422</td>
<td>0.090</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>M3</td>
<td>0.113</td>
<td>0.669</td>
<td>0.665</td>
<td>0.112</td>
<td>0.430</td>
</tr>
<tr>
<td></td>
<td>M4</td>
<td>0.168</td>
<td>1.093</td>
<td>1.088</td>
<td>0.156</td>
<td>0.947</td>
</tr>
<tr>
<td></td>
<td>M5</td>
<td>0.453</td>
<td>2.051</td>
<td>2.046</td>
<td>0.501</td>
<td>7.083</td>
</tr>
</tbody>
</table>
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.
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†, 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 \( P \) (32 observations) whereas cells without glucose will be considered as coming from population \( P^\ast \) (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.

†The hellung dataset is available in the ISwR package for \( R \).
Adaptive linear models for regression

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 \( 4 \times 10^5 \) (17 obs.), smaller than \( 2 \times 10^5 \) (14 obs.) and smaller than \( 1 \times 10^5 \) (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.

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 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. These two models consider a shift of the intercept, what confirms the guess that we can have by examining graphically the dataset, and moreover by quantifying this shift. To conclude, this study demonstrates that the adaptive linear models can be successfully applied to real data 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. 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.

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>whole dataset</th>
<th>$X \leq 4 \times 10^5$</th>
<th>$X \leq 2 \times 10^5$</th>
<th>$X \leq 1 \times 10^5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS on $P^*$ (M0)</td>
<td>0.897</td>
<td>0.364</td>
<td>0.432</td>
<td>0.303</td>
</tr>
<tr>
<td>Model M1</td>
<td>3.332</td>
<td>0.283</td>
<td>2.245</td>
<td>0.344</td>
</tr>
<tr>
<td>Model M2</td>
<td>0.269</td>
<td>0.294</td>
<td>0.261</td>
<td>0.130</td>
</tr>
<tr>
<td>Model M3</td>
<td>0.287</td>
<td>0.271</td>
<td>0.289</td>
<td>0.133</td>
</tr>
<tr>
<td>Model M4</td>
<td>0.859</td>
<td>1.003</td>
<td>0.756</td>
<td>0.517</td>
</tr>
<tr>
<td>Model M5</td>
<td><strong>0.256</strong></td>
<td><strong>0.259</strong></td>
<td><strong>0.255</strong></td>
<td><strong>0.124</strong></td>
</tr>
</tbody>
</table>
Table 4. Effect of the learning set size on the RSS 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.

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

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 spare 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 United States Department of Commerce (1989) 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.

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 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
Adaptive linear models for regression

Fig. 6. PRESS criterion for the Birmingham-San Jose data.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>10 obs.</th>
<th>25 obs.</th>
<th>50 obs.</th>
<th>100 obs.</th>
<th>250 obs.</th>
<th>all obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model M0</td>
<td>3.5×10^7</td>
<td>386.1</td>
<td>336.8</td>
<td>310.7</td>
<td>297.5</td>
<td>297.5</td>
</tr>
<tr>
<td>Model M2</td>
<td>414.8</td>
<td>356.7</td>
<td>342.1</td>
<td>336.0</td>
<td>332.5</td>
<td>330.1</td>
</tr>
<tr>
<td>Model M6</td>
<td>1528.9</td>
<td>1528.9</td>
<td>1528.9</td>
<td>1528.9</td>
<td>1528.9</td>
<td>1528.9</td>
</tr>
</tbody>
</table>

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. To sum up, 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. 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 shows how it is possible to adapt a regression model from a given situation to another new one, and thus to spare a new expensive collect of data. In this perspective, a non-exhaustive 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 a satisfying estimation of the new regression model. All the interest of this work arises then when the sample size for the new population is too small to efficiently estimate a regression model by usual OLS procedure without using information known for the reference population. 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 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 needed to obtain a similar estimation quality with the adaptive linear models. In addition, the estimated transformation parameters could help practitioners to analyse the differences between both populations.

A. Least square estimators of the transformation parameters

In the following the least square estimators of the parameters of the transformation model $M_1$ to $M_5$ are derived.

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

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

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

$$
\hat{\Lambda}_{OLS}^{OL} = (\Psi_{-1}^t \beta_{-1} \beta_{-1}^t \Psi_{-1})^{-1} \beta_{-1}^t \Psi_{-1}^t (y^* - \beta_0)
$$

where $\Psi^*$ is a $(n^*) \times (d + 1)$ matrix formed by the row vector $\Psi(x_i^*)^t$ $(1 \leq i \leq n^*)$.

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

$$
\ln l(y^*; x^*, \Lambda, \sigma^2) = -n^* \ln(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n^*} \left( y_i^* - \beta_0 \lambda_0 - \beta^t \Lambda_{-1} \Psi_{-1}(x_i^*) \right)^2
$$
Adaptive linear models for regression

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

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

where

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

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

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

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

$$\hat{\lambda}^{OLS} = (\Psi_{-1}^* \beta_{-1} \beta_{-1}^t \Psi_{-1}^*)^{-1} \beta_{-1}^t \Psi_{-1}^* (y^* - \beta_0).$$

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

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

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


