PARAMETRIC LINK MODELS FOR KNOWLEDGE TRANSFER IN STATISTICAL LEARNING

Farid Beninel, Christophe Biernacki, Charles Bouveyron, Julien Jacques, Alexandre Lourme

Chapter 1

PARAMETRIC LINK MODELS FOR KNOWLEDGE TRANSFER IN STATISTICAL LEARNING

Beninel F.[1]**, *Biernacki C.***[2]**, *Bouveyron C.***[3]**, *Jacques J.***[*2] **and *Lourme A.***[4]
[1]CREST-ENSAI, Bruz, France
[2]Université Lille 1 & CNRS & INRIA, Lille, France
[3]Université Paris 1 Panthéon-Sorbonne, Paris, France
[4]Université de Pau et des Pays de l'Adour, Pau, France

Abstract

When a statistical model is designed in a prediction purpose, a major assumption is the absence of evolution in the modeled phenomenon between the training and the prediction stages. Thus, training and future data must be in the same feature space and must have the same distribution. Unfortunately, this assumption turns out to be often false in real-world applications. For instance, biological motivations could lead to classify individuals from a given species when only individuals from another species are available for training. In regression, we would sometimes use a predictive model for data having not exactly the same distribution that the training data used for estimating the model. This chapter presents techniques for transfering a statistical model estimated from a *source* population to a *target* population. Three tasks of statistical learning are considered: Probabilistic classification (parametric and semi-parametric), linear regression (including mixture of regressions) and model-based clustering (Gaussian and Student). In each situation, the knowledge transfer is carried out by introducing parametric links between both populations. The use of such transfer techniques would improve the performance of learning by avoiding much expensive data labeling efforts.

Key Words: Adaptive estimation, link between populations, transfer learning, classification, regression, clustering, EM algorithm, applications. **A**MS Subject Classification: 62H30, 62J99.

[*]E-mail address: julien.jacques@polytech-lille.fr

1. Introduction

Statistical learning [17] is a key tool for many science and application areas since it allows to explain and to predict diverse phenomena from the observation of related data. It leads to a wide variety of methods, depending on the particular problem at hand. Examples of such problems are numerous:

- Examples E_1: In *Credit Scoring*, predict the behavior of borrowers to pay back loan, on the basis of information known about these customers; In *Medicine*, predict the risk of lung cancer recurrence for a patient treated for a first cancer, on the basis of the type of treatment used for the first cancer and on clinical and demographic measurements for that patient.

- Examples E_2: In *Economics*, predict the housing price on the basis of several housing descriptive variables; In *Finance*, predict the profitability of a financial asset six months after purchase.

- Examples E_3: In *Marketing*, create customers groups according to their purchase history in order to target a marketing campaign; In *Biology*, identify groups in a sample of birds described by some biometric features which finally reveal the presence of different genders.

In a typical statistical learning problem, a response variable $y \in \mathcal{Y}$ has to be predicted from a set of d feature variables (or covariates) $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathcal{X}$. Spaces $\mathcal{X}$ and $\mathcal{Y}$ are usually quantitative or categorical. It is also possible to have heterogeneity in features variables (both quantitative and categorical for instance). The analysis always relies on a training dataset $\mathcal{S} = (\mathbf{x}, \mathbf{y})$, in which the response and feature variables are observed for a set of n individuals which are respectively denoted by $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$. Using $\mathcal{S}$, a predictive model is built in order to predict the response variable for a new individual, for which the covariates $\boldsymbol{x}$ are observed but not the response y. This typical situation is called *supervised* learning. In particular, if $\mathcal{Y}$ is a categorical space, it corresponds to a *discriminant analysis* situation; It aims to solve problems which look like Examples E_1. If $\mathcal{Y}$ is a quantitative space, it corresponds to a *regression* situation and aims to solve problems similar to Examples E_2. Note also that if $\mathbf{y}$ is only partially known in $\mathcal{S}$, it exhibits what is called *semi-supervised* learning.

Another typical statistical learning problem consists in predicting the whole responses $\mathbf{y}$ while having never observe them. In this case only the feature variables are known, thus $\mathcal{S} = \mathbf{x}$, and it corresponds to an *unsupervised* learning situation. If $\mathcal{Y}$ is restricted to a categorical space (the most frequent case), it consists in a *clustering* purpose, related problems being illustrated by Examples E_3.

In this chapter, we focus on statistical modeling for solving as well supervised and unsupervised learning. Many classical probabilistic methods exist and we will give useful references, when necessary, throughout the chapter. Thus, the reader interested for such references is invited to have a look in related sections below.

A main assumption in supervised learning is the absence of evolution in the modeled phenomenon between the training of the model and the prediction of the response for a new

individual. More precisely, the new individual is assumed to arise from the same statistical population than the training one. In unsupervised learning, it is also implicitly assumed that all individuals arise from the same population. Unfortunately, such classical hypotheses may not hold in many realistic situations as reflected by revisited Examples E_1 to E_3:

- Examples E_1^*: In *Credit Scoring*, the statistical scoring model has been trained on a dataset of customers but is used to predict behavior of non-customers; In *Medicine*, the risk of lung cancer recurrence is learned for an European patient but will be applied to an Asian patient.

- Examples E_2^*: In *Economics*, a real-estate agency implanted for a long time on the US East Coast aims to conquer new markets by opening several agencies on the West Coast but both markets are quite different; In *Finance*, expertise in financial asset of the past year is surely different from the current one.

- Examples E_3^*: In *Marketing*, customers to be classified correspond in fact to a pooled panel of new and older customers; In *Biology*, different subpecies of birds are pooled together and may consequently have highly different features for the same gender.

In the supervised setting, the question is "Q_1: Is it necessary to recollect new training data and to build a new statistical learning model or can the previous training data still be useful?" In the unsupervised setting, the question is "Q_2: Is it better to perform a unique clustering on the whole data set or to perform several independant clusterings on some identified subsets?".

Question Q_1 is addressed as *transfer* learning and a general overview is given in [28]. Transfer learning techniques aim to transfer the knowledge learned on a source population Ω to a target population Ω^*, in which this knowledge will be used in a prediction purpose. These techniques are divided into two important situations: The transfer of a model *does need* or *does not need* to observe some response variables in the target domain. The first case is quoted as *inductive transfer* learning whereas the second one is quoted as *transductive transfer* learning. Usually, the classification purpose as described in Examples E_1^* can be solved by either transductive or inductive transfer learning, this choice depending on the model at hand (generative or predictive models). Contrariwise, the regression purpose as described in Examples E_2^* can be only solved by inductive transfer learning since only predictive models are involved. Question Q_2 is adressed as *unsupervised transfer* learning. It corresponds to simultaneous clustering of several samples and, thus, it concerns Examples E_3^*.

A common expected advantage of all these transfer learning techniques is a real predictive benefit since knowledge learned on the source population is used in addition to the available information on the target population. However, the common challenge is to establish a "transfer function" between the source and the target populations. In this chapter, we focus on parametric statistical models. Besides being good competitors to nonparametric models in terms of prediction, these models have the advantage of being easily interpreted by practitioners. Since parametric models will be used, it will be natural to modelize the transfer function by some parametric links. Thus, in addition to a predictive benefit, the interpretability of the link parameters will give to practitioners useful information on the evolution and the differences between the source and target populations.

This chapter is organized as follows. Section 2. presents transfer learning for different discriminant analysis contexts: Gaussian model (continuous covariates), Bernoulli model (binary covariates) and logistic model (continuous or binary covariates). Section 3. considers the transfer of regression models for a quantitative response variable in two situations: Usual regression and mixture of regressions. Finally, Section 4. proposes models to cluster simultaneously a source and a target population in two situations again: Mixtures of Gaussian and Student distributions. Each section starts with a presentation of the classical statistical model before presenting the corresponding transfer techniques, and it concludes by an application on real data.

A useful notation In the following the notation "*" will refer to the target population.

2. Parametric transfer learning in discriminant analysis

Discriminant analysis is a large methodological field covering machine learning techniques dealing with data where individuals are described by the same set of d covariates or feature vector $\boldsymbol{x}$ and a response categorical variable $y \in \mathcal{Y} = \{1,\dots,K\}$ related to K classes, where $y = k$ if the individual described by $\boldsymbol{x}$ belongs to the kth class. In a statistical setting, the couple $(\boldsymbol{x},y)$ is assumed to be a realization of a random vector $(\boldsymbol{X},Y)$ where $\boldsymbol{X} = (X_1,\dots,X_d)$. Then the n-sample $\mathcal{S} = (\mathbf{x},\mathbf{y})$ is assumed to be n i.i.d. realizations of $(\boldsymbol{X},Y)$.

The purpose of discriminant analysis is to predict the group membership y, only on the basis of the covariates $\boldsymbol{x}$. The discriminant analysis proceeds as follows: Using $\mathcal{S}$, an allocation rule is built in order to classify non-labeled individuals. Many books explain in detail the numerous techniques related to discriminant analysis [16,17,25,29], among which the main are parametric ones, semi-parametric ones, non-parametric ones and borderline-based ones. In this section, we are interested only by parametric (Gaussian and Bernoulli distributions) and semi-parametric (logistic regression) methods.

2.1. Gaussian discriminant analysis

2.1.1. The statistical model

Gaussian discriminant analysis assumes that, conditionally to the group y, the feature variables $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^d$ arise from a random vector $\boldsymbol{X}$ distributed according to a d-variate Gaussian distribution

$$\boldsymbol{X}|Y = k \sim \mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$ are respectively the associated mean and covariance matrix. The probability density of $\boldsymbol{X}$ conditionally to $Y = k$ is

$$f_k(\bullet;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\bullet-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\bullet-\boldsymbol{\mu}_k)\right).$$

The marginal distribution of $\boldsymbol{X}$ is then a mixture of Gaussian distributions

$$\boldsymbol{X} \sim f(\bullet;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\bullet;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),$$

where $(\pi_1,\ldots,\pi_K)$ are the mixing proportions ($\pi_k > 0$ and $\sum_{k=1}^K \pi_k = 1$) and $\boldsymbol{\theta} = \{(\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) : k = 1,\ldots,K\}$ is the whole parameter. When the costs of bad classification are assumed to be symmetric, the *Maximum A Posteriori* (MAP) rule consists in assigning a new individual $\boldsymbol{x}$ to the group $\hat{y}$ maximizing the membership conditional probability $t_{\hat{y}}(\boldsymbol{x};\boldsymbol{\theta})$:

$$\hat{y} = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}}\ t_k(\boldsymbol{x};\boldsymbol{\theta}), \tag{1}$$

where

$$t_k(\boldsymbol{x};\boldsymbol{\theta}) = P(Y = k|\boldsymbol{X} = \boldsymbol{x};\boldsymbol{\theta}) = \frac{\pi_k f_k(\boldsymbol{x};\boldsymbol{\alpha}_k)}{f(\boldsymbol{x};\boldsymbol{\theta})}. \tag{2}$$

In the general heteroscedastic situation (quadratic discriminant analysis or QDA), $\boldsymbol{\theta}$ is estimated by its classical empirical estimates:

$$\hat{\pi}_k = \frac{n_k}{n}, \quad \hat{\boldsymbol{\mu}}_k = \frac{1}{n_k}\sum_{\{i:y_i=k\}} \boldsymbol{x}_i, \quad \hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k-1}\sum_{\{i:y_i=k\}} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)',$$

where $n_k = \text{card}\{i : y_i = k\}$ is the number of individuals of the training sample $\mathcal{S}$ belonging to the group k. In the restricted homoscedastic situation $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for all k (linear discriminant analysis or LDA), the covariance matrix is estimated by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-K}\sum_{k=1}^K \sum_{\{i:y_i=k\}} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)'.$$

2.1.2. The transfer learning and its estimation

Now we assume that the data consist of two samples: A first labeled n-sample $\mathcal{S} = (\mathbf{x},\mathbf{y})$, drawn from a source population Ω, and a second unlabeled n^*-sample $\mathcal{S}^* = \mathbf{x}^*$, drawn from a target population Ω^*. Our goal is to build a classification rule for the target population using both samples $\mathcal{S}$ and $\mathcal{S}^*$. An extension to a partially-labeled target sample $\mathcal{S}^*$ will be also presented later. The source labeled sample $\mathcal{S}$ is composed by n pairs $(\boldsymbol{x}_i, y_i)$, assumed to be i.i.d. realizations of the random couple $(\boldsymbol{X}, Y)$ of distribution

$$\boldsymbol{X}|Y = k \sim \mathcal{N}_d(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \qquad \text{and} \qquad Y \sim \mathcal{M}_1(\pi_1,...,\pi_K),$$

where $\mathcal{M}_1$ is the one-order multinomial distribution. The target unlabeled sample $\mathcal{S}^*$ is composed by n^* pairs $\boldsymbol{x}_i^*$ i.i.d. realizations of $\boldsymbol{X}^*$ with the following Gaussian mixture distribution

$$\boldsymbol{X}^* \sim f(\bullet;\boldsymbol{\theta}^*).$$

In order to use both samples $\mathcal{S}$ and $\mathcal{S}^*$ for the classification of $\mathcal{S}^*$ sample (or of any new individual $\boldsymbol{x}$ from Ω^*), the approach developped in [3] consists in establishing a stochastic relationship $\phi_k(\mathbb{R}^d \mapsto \mathbb{R}^d)$ between feature vectors of both populations conditionally to groups, *i.e.*

$$\boldsymbol{X}^*|Y = k \overset{\mathcal{D}}{=} \phi_k(\boldsymbol{X}|Y = k) = [\phi_k^1(\boldsymbol{X}|Y = k),\ldots,\phi_k^d(\boldsymbol{X}|Y = k)],$$

where $\mathcal{D}$ means that the equality is in distribution, and ϕ_k^j, $j = 1,\ldots,d$, is an application $(\mathbb{R}^d \mapsto \mathbb{R})$. Two natural assumptions are considered:

- $\mathcal{A}_1$: The jth component $\phi_k^j(\boldsymbol{X}|Y=k)$ only depends on the jth component of $\boldsymbol{X}|Y=k$,
- $\mathcal{A}_2$: Each ϕ_k^j is $\mathcal{C}^1$.

As a consequence of the previous assumptions [10] derive the K relations

$$\boldsymbol{X}^*|Y^*=k \stackrel{\mathcal{D}}{=} \mathbf{D}_k\boldsymbol{X}|Y=k+\mathbf{b}_k \qquad (k=1,\ldots,K) \tag{3}$$

with $\mathbf{D}_k$ a $d\times d$ real diagonal matrix and $\mathbf{b}_k$ a d dimensional real vector. Therefore, we establish the following relations between parameters of the Gaussian distributions related to populations Ω and Ω^*:

$$\boldsymbol{\mu}_k^*=\mathbf{D}_k\boldsymbol{\mu}_k+\mathbf{b}_k \qquad \text{and} \qquad \boldsymbol{\Sigma}_k^*=\mathbf{D}_k\boldsymbol{\Sigma}_k\mathbf{D}_k. \tag{4}$$

Such relations allow to determine the allocation rules for population Ω^* using parameters of feature vector distribution for individuals of Ω. Indeed, if the K pairs $(\mathbf{D}_k,\mathbf{b}_k)$ are known it is easy to derive pairs $(\boldsymbol{\mu}_k^*,\boldsymbol{\Sigma}_k^*)$ from $(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ by plug-in. In what follows we discuss issues where the pairs $(\mathbf{D}_k,\mathbf{b}_k)$ are unkwown and we propose several scenarios for estimating them.

Constrained models For identifiability reasons we impose that $\mathbf{b}_k=\mathbf{0}$ for all $k=1,\ldots,K$. This assumption is discussed in the seminal article on transfer learning in Gaussian discriminant analysis [3], and validated on the biological application analysed in this article. The case without constraints on $\mathbf{b}_k$ is treated in [22] which provides specific computation approach for avoiding identifiability problems (see also Section 4. of the present chapter).

In order to define parsimonious and meaningfull models, constraints are now imposed on the parameters of transfer $\mathbf{D}_k$ $(k=1,\ldots,K)$:

- Model M_1: $\mathbf{D}_k=\mathbf{I}_d$: The K distributions are the same ($\mathbf{I}_d$: identity matrix of $\mathbb{R}^{d\times d}$).
- Model M_2: $\mathbf{D}_k=\alpha\mathbf{I}_d$: Transformations are feature and group independent.
- Model M_3: $\mathbf{D}_k=\mathbf{D}$: Transformations are only group independent.
- Model M_4: $\mathbf{D}_k=\alpha_k\mathbf{I}_d$: Transformations are only feature independent.
- Model M_5: $\mathbf{D}_k$ is unconstrained, *i.e.* it is the most general situation.

Model M_1 consists in using allocation rules on Ω^* based only on $\mathcal{S}$, *i.e.* we deal here with classical discriminant analysis. Models M_2 and M_3 preserve homoscedasticity and consequently an eventual linearity of the rule: If $\boldsymbol{\Sigma}_1=\ldots=\boldsymbol{\Sigma}_K$ for Ω, then $\boldsymbol{\Sigma}_1^*=\ldots=\boldsymbol{\Sigma}_K^*$ for Ω^*. Last models M_4 and M_5 may transform linear allocation rules into quadratic ones on Ω^* with few parameters to estimate.

For each model, an additional assumption on the mixing proportions is done: They are the same in both populations or they have to be estimated in the target population. Corresponding models are quoted by M_j and respectively πM_j $(1\leq j\leq 5)$. The number of free parameters for each model are given in Table 1.

M_1	M_2	M_3	M_4	M_5	πM_1	πM_2	πM_3	πM_4	πM_5
0	1	d	K	dK	$K-1$	K	$d+K-1$	$2K-1$	$dK+K-1$

Table 1. Number of estimated parameters for each model.

Parameter estimation A sequential *plug-in* procedure is used to estimate matrices $\mathbf{D}_1,\ldots,\mathbf{D}_K$ (and eventually $\pi_1^*,\ldots,\pi_K^*$). The corresponding estimators will depend on parameter $\boldsymbol{\theta}$ of population Ω. When this last is unknown, it is simply replaced by its estimate. Estimating all π_k^* and all $\mathbf{D}_k$ is performed by maximizing the following likelihood, under the constraints given in (4) and under the constraint of one of the previous parsimonious models M_j or πM_j ($j=1,\ldots,5$),

$$L(\boldsymbol{\theta}^*)=\prod_{i=1}^{n^*} f(\boldsymbol{x}_i^*;\boldsymbol{\theta}^*). \tag{5}$$

A usual way to maximize the likelihood when the group membership y_i^* are unknown is to use an EM algorithm [11] which consists in iterating the two following steps:

- **E step**: Estimation of the group membership y_i^* by its expectation conditional to the observed data:

$$\hat{y}_i^* = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}}\ t_k(\boldsymbol{x}_i^*;\boldsymbol{\theta}^*).$$

- **M step**: Computation of the parameter $\boldsymbol{\theta}^*$ maximizing, under the constraints given in (4) and under the constraint of a given parsimonious models (M_j or πM_j), the following completed log-likelihood:

$$\ell_c(\boldsymbol{\theta}^*)=\sum_{k=1}^{K}\sum_{\{i:y_i^*=k\}}\ln\left[\pi_k^* f_k(\boldsymbol{x}_i^*;\boldsymbol{\alpha}_k^*)\right].$$

The EM algorithm stops when the growth of the likelihood is lower than a fixed threshold.

In order to choose between several constrained models, the BIC criterion (*Bayesian Information Criterion*, [31]) is used:

$$\mathrm{BIC}=-2\ln\ell+|\boldsymbol{\theta}^*|\ln n^* \tag{6}$$

where ℓ is the maximum log-likelihood value and $|\boldsymbol{\theta}^*|$ denotes the number of continuous model parameters in $\boldsymbol{\theta}^*$. The model leading to the minimum BIC value is retained. Note that the BIC criterion is faster to compute that any cross-validation criterion.

2.1.3. A biological application

Data Data are related to seabirds from Cory's Shearwater *Calanectris diomedea* species breeding in the Mediterranean and North Atlantic, where presumably contrasted oceanographic conditions have led to the existence of marked subspecies differing in size as well

as coloration and behavior [32]. Subspecies are *borealis*, living in the Atlantic islands (the Azores, Canaries, etc.), *diomedea*, living in the Mediterranean islands (Balearics, Corsica, etc.), and *edwardsii*, from the Cape Verde Islands.

A sample of *borealis* ($n = 206$, 45% females) was measured using skins in several National Museums. Five morphological variables are measured: Culmen (bill length), tarsus, wing and tail lengths, and culmen depth. Similarly, a sample of subspecies *diomedea* ($n = 38$, 58% females) was measured using the same set of variables. Figure 1 plots culmen depth and tarsus length for *borealis* and *diomedea* samples.

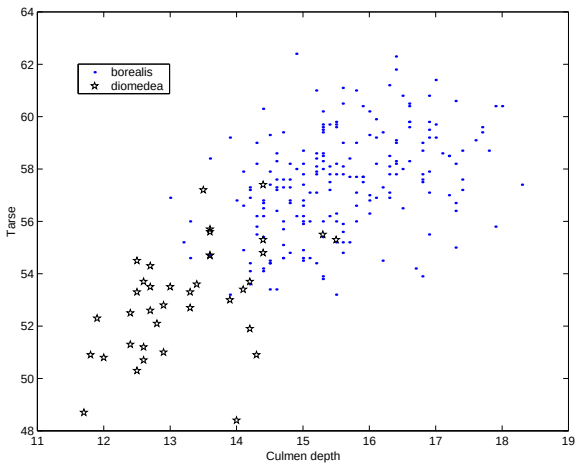


Figure 1. *Borealis* and *diomedea* for variables culmen depth and tarsus length.

In the following, we consider the *borealis* sample as being the source labeled population, and the *diomedea* as being the target population (non-labeled of partially-labeled). In reality, in our data, both samples are sexed but sex of *diomedea* will be only used to measure quality of results provided by the proposed method.

Results in the non-sexed case We consider in this section that all *diomedea* specimen are non-sexed. Linear discriminant analysis model is selected for the *borealis* population. We apply parameters estimated by the *borealis* sample using the 10 models to the non-sexed *diomedea* sample. Results, empirical error rate (deduced from the true partition of *diomedea*) and BIC value, are given for each model in Table 2. Moreover, empirical error rate of the cluster analysis situation is reported at the last column of Table 2. The clustering procedure (see for instance [9]) consists in estimating the Gaussian mixture parameters of the non-sexed sample *diomedea* without using the *borealis* sample.
High error rates are generally obtained with standard discriminant analysis (models M_1 and πM_1) and with standard cluster analysis, as compared to the other transfer learning models. The best model selected by the empirical error rate is πM_3 . This model preserves homoscedasticity, a relevant property since both discriminant rules selected by cross-validation criterion separately on each sample $\mathcal{S}$ and $\mathcal{S}^*$ were homoscedastic (LDA). Moreover it indicates that the proportion of females is not the same in the two samples. Model selected by the BIC criterion is M_3 and the error rate is the second best value. So,

model	M_1	M_2	M_3	M_4	M_5
error	42.11	31.58	18.43	28.95	21.06
BIC	-753.49	-502.13	**-451.51**	-503.74	-457.69

model	πM_1	πM_2	πM_3	πM_4	πM_5	clustering
error	42.11	42.11	**15.79**	42.11	21.06	44.73
BIC	-725.24	-489.43	-453.20	-491.23	-459.51	–

Table 2. Empirical error rate (error) and BIC value (BIC) in the non-sexed case.

transformation from *borealis* to *diomedea* seems to be sex-independent but not variable-independent. It should be noted also that BIC's value for πM_3 is very close to the one for M_3.

Results in the partially-sexed case We consider in this section that two labels (therefore 5.26% of the data set) are known in the *diomedea* sample, thus a part $\tilde{\mathbf{y}}^*$ of $\mathbf{y}^*$ is known and $\mathcal{S}^* = (\mathbf{x}^*, \tilde{\mathbf{y}}^*)$. Empirical error rate is obtained for the 36 *a priori* non-sexed birds. The two labels are choosen at random 30 times and, so, it leads to 30 partially-sexed samples. The 10 models and cluster analysis (using also this new sex information, what leads to a semi-supervised situation) are applied successively to the 30 partially-sexed *diomedea* samples. Mean of the error rate and the BIC criterion are displayed in Table 3.

model	M_1	M_2	M_3	M_4	M_5
error	42.41	31.94	18.70	29.91	18.98
BIC	-753.49	-502.13	**-451.56**	-503.92	-457.95

model	πM_1	πM_2	πM_3	πM_4	πM_5	clustering
error	42.41	42.69	**15.37**	42.69	20.93	21.13
BIC	-725.99	-489.95	-453.32	-491.77	-460.74	–

Table 3. Mean on the 30 samples of the empirical error rate (error) and the BIC value (BIC) in the partially-sexed case.

Partial information on sex provides lower error rates in models πM_3, πM_5, M_5 and the clustering method, with the model πM_3 still being the best. The BIC criterion still selects the model M_3 (with a low error rate) and then πM_3. We note that, except model M_5, only adapted models improve thanks to this new label knowledge. Moreover, the more complex the model is, the more the error of classification strongly decreases. This is the case for clustering: It has a good improvement in this example, coming from the last rank to a level close to πM_5.

2.2. Discriminant analysis for binary data

2.2.1. The statistical model

We now consider discriminant analysis for binary feature variables, so $\mathcal{X} = \{0,1\}^d$. If the Gaussian assumption is common for quantitative feature variables, binary feature x_j is commonly assumed to arise from a random variable X_j having, conditionally on Y, a Bernoulli distribution $\mathcal{B}(\alpha_{kj})$ of parameter α_{kj} ($0 < \alpha_{kj} < 1$):

$$X_j|Y = k \sim \mathcal{B}(\alpha_{kj}) \quad (j = 1,\ldots,d). \tag{7}$$

Using the assumption of conditional independence of the explanatory variables [8, 13], the probability density function of $\boldsymbol{X}$, conditionally on Y, is:

$$f_k(\boldsymbol{x};\boldsymbol{\alpha}_k) = \prod_{j=1}^{d} {\alpha_{kj}}^{x_j}(1-\alpha_{kj})^{1-x_j}, \tag{8}$$

where $\boldsymbol{\alpha}_k = (\alpha_{k1},\ldots,\alpha_{kd})$. The mixing proportions π_k and the whole parameter $\boldsymbol{\theta} = \{(\pi_k,\boldsymbol{\alpha}_k), k = 1,\ldots,K\}$ are then defined similarly to the previous Gaussian situation. Maximum likelihood (ML) estimates of all α_{kj} are simply given by the following relative empirical frequencies:

$$\hat{\alpha}_{kj} = \frac{\text{card}\{i : y_i = k, x_{ij} = 1\}}{n_k}.$$

The estimation of any y is then obtain by the MAP principle given in (1), $\boldsymbol{\theta}$ being plug-in by its estimate (estimate of the mixing proportion π_k are the same as in the Gaussian situation).

2.2.2. The transfer learning

Defining a transfer function Feature variables in the target population Ω^* are assumed to have the same distribution as (7) but with possibly different parameters α^*_{kj}

$$X^*_j|Y^* = k \sim \mathcal{B}(\alpha^*_{kj}).$$

In a multinormal context, the transfer learning challenge has been reached by considering a linear stochastic relationship between the source Ω and the target Ω^*. This link was not only justified (under very few assumptions) but also intuitive [3]. In the binary context, such an intuitive relationship seems more difficult to exhibit. The idea developed in [21] is to assume that the binary variables result from the discretization of some latent Gaussian variables. From a stochastic link between the latent variables analogous to (3), the following link between the parameters α^*_{kj} of Ω^* and α_{kj} of Ω is obtained:

$$\alpha^*_{kj} = \Phi\Big(\delta_{kj}\,\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_{kj}\Big), \tag{9}$$

where Φ is the cumulative density function of $\mathcal{N}(0,1)$, $\delta_{kj} \in \mathbb{R}^+ \setminus \{0\}$, $\lambda_j \in \{-1,1\}$ and $\gamma_{kj} \in \mathbb{R}$. Note that this relationship corresponds to a linear link between the *probit* functions of both α_{kj} and α^*_{kj}.

Conditionally to the fact that α_{kj} are known (they will be estimated in practice), estimation of the Kd continuous parameters α^*_{kj} is thus obtained from estimates of the link

parameters δ_{kj}, γ_{kj} and λ_j between Ω and Ω^* (plug-in method). However, estimating the number of parameters for the link map is $2Kd$ and one thus obtains that the model is overparameterized. This fact should not be surprising since the underlying Gaussian model is by far more complex (in terms of the number of parameters) than the Bernoulli model. Hence there is a need to reduce the number of free continuous parameters in (9), and [21] propose constrained models imposing natural additional constraints on the transformation between both populations Ω and Ω^*.

Constrained models The parameters δ_{kj} ($1 \leq k \leq K$ and $1 \leq j \leq d$) will be successively constrained to be equal to 1 (denoted by 1), to be class- and dimension-independent (δ), to be only class-dependent (δ_k) or only dimension-dependent (δ_j). In the same way, γ_{kj} can be constrained to be equal to 0, γ (constant w.r.t. k and j), γ_k (constant w.r.t. j) or γ_j (constant w.r.t. k). Thus, 16 models can be defined and indexed using the *ad hoc* notation summarized in Table 4. For these 16 models, as for the Gaussian case, the assumption on

	0	γ	γ_j	γ_k
1	10	1γ	$1\gamma_j$	$1\gamma_k$
δ	10	$\delta\gamma$	$\delta\gamma_j$	$\delta\gamma_k$
δ_j	$\delta_j 0$	$\delta_j\gamma$	$\delta_j\gamma_j$	$\delta_j\gamma_k$
δ_k	$\delta_k 0$	$\delta_k\gamma$	$\delta_k\gamma_j$	$\delta_k\gamma_k$

Table 4. Constrained models for binary discriminant analysis transfer learning.

the group proportions is also taken into account: For instance, an equal proportion model is quoted $\delta_k\gamma_k$ and a free proportion model is quoted $\pi\delta_k\gamma_k$. The number of constrained models is thus growing to 32. The ML estimation of the parameter $\boldsymbol{\theta}^*$ is carried out by the EM algorithm under the link constraint (9) and also w.r.t. the considered model on the link parameters. Even if some of these models can be non-identifiable, [21] show that identifiability will occur in practical situations. Then, the choice between these 32 models can be performed by the BIC criterion given in (6).

2.2.3. Biological application

In this application birds from the species *puffins* are considered, and the goal is to predict their sex [7]. Two groups of subspecies are considered: The first one is composed of subspecies living in Pacific Islands – *subalaris* (Galapagos Island), *polynesial*, *dichrous* (Enderbury and Palau Islands) and *gunax* – and the second one is composed of subspecies living in Atlantic Islands – *boydi* (Cap Verde Islands). Here, the difference between populations is the geographical range (Pacific *vs.* Atlantic Islands). A sample of Pacific birds ($n = 171$) was measured using skins in several National Museums. Four variables are measured on these birds: Coller, stripe and piping (absence or presence for these three variables) and under-caudal (self couloured or not). Similarly, a sample of Atlantic birds ($n = 19$) was measured using the same set of variables. Like in the previous example, two groups are present (males and females) and the sex of all the birds is known.

Pacific birds are chosen as the source population and Atlantic ones as the target population. Choosing Atlantic birds as the target population corresponds to a realistic situation because it could be hazardous to perform a clustering process on a sample of such a small size. This is a typical situation where the proposed methodology could be expected to provide a parsimonious and meaningful alternative. According to the biologist who provided the data, the morphological variables which are used in this application are not very discriminative, and then one can not expect that the error rate will be better than $40-45\%$.

The 32 transfer learnings models for binary discrimination, among which standard discriminant analysis 10, are applied on these data and the results are presented in Table 5. Clustering is also applied, and the obtained error rate is 49.05%. The best transfer learning

model	10	1γ	$1\gamma_k$	$1\gamma_j$	$\delta 0$	$\delta\gamma$	$\delta\gamma_k$	$\delta\gamma_j$
error	50.94	**43.39**	45.28	43.39	50.94	**43.39**	45.28	45.28
BIC	212	**209**	216	224	212	**209**	216	224
model	$\delta_k 0$	$\delta_k\gamma$	$\delta_k\gamma_k$	$\delta_k\gamma_j$	$\delta_j 0$	$\delta_j\gamma$	$\delta_j\gamma_k$	$\delta_j\gamma_j$
error	45.28	45.28	52.83	45.28	45.28	52.83	50.94	50.94
BIC	210	210	215	226	225	224	227	239
model	$\pi 10$	$\pi 1\gamma$	$\pi 1\gamma_k$	$\pi 1\gamma_j$	$\pi\delta 0$	$\pi\delta\gamma$	$\pi\delta\gamma_k$	$\pi\delta\gamma_j$
error	45.28	50.94	50.94	45.28	45.28	50.94	50.94	45.28
BIC	213	213	220	228	213	213	220	228
model	$\pi\delta_k 0$	$\pi\delta_k\gamma$	$\pi\delta_k\gamma_k$	$\pi\delta_k\gamma_j$	$\pi\delta_j 0$	$\pi\delta_j\gamma$	$\pi\delta_j\gamma_k$	$\pi\delta_j\gamma_j$
error	45.28	45.28	47.16	45.28	45.28	52.83	45.28	52.83
BIC	214	213	213	229	228	227	224	243

Table 5. Classification error rates (%) and value of the BIC criterion for target population of Atlantic birds with source on Pacific birds population.

model gives an error rate lower than standard discriminant analysis (50.94%) or clustering (49.05%) to classify birds according to their sex. Moreover the BIC criterion leads to choose the model with the smallest error rate. The relatively poor classification results (the minimal error rate is 43%) confirms the assumption of the biologist.

2.3. Logistic regression

2.3.1. The statistical model

Contrary to both previous approaches, logistic regression (see for instance [19, 26]) can be viewed as a partially parametric method since it models only the ratio $f_k(\boldsymbol{x})/f_{k'}(\boldsymbol{x})$ $(k \neq k')$ instead of modeling each single group distribution $f_k(\boldsymbol{x})$. Here, we study the case where covariates $\boldsymbol{x}$ include d components which are continuous and/or binary. The general categorical case (more than two levels for some components of $\boldsymbol{x}$) is easily taken into account by replacing each $r \geq 2$ levels covariate by $r-1$ binary covariates. We consider also that K groups have to be discriminated. Following the conventional but arbitrary choice of the

Kth group as a base group, the logistic model fundamentally assumes that

$$\ln \frac{f_k(\boldsymbol{x})}{f_K(\boldsymbol{x})} = \beta_{0k}^0 + \boldsymbol{\beta}_k' \boldsymbol{x}$$

where $\beta_{0k}^0 \in \mathbb{R}$ and $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{dk})' \in \mathbb{R}^d$ $(k = 1, \ldots, K-1)$. Equivalently, it can be written by using the conditional membership probabilities $t_k(\boldsymbol{x}; \boldsymbol{\beta})$

$$\ln \frac{t_k(\boldsymbol{x}; \boldsymbol{\beta})}{t_K(\boldsymbol{x}; \boldsymbol{\beta})} = \beta_{0k} + \boldsymbol{\beta}_k' \boldsymbol{x}$$

where $\beta_{0k} = \beta_{0k}^0 + \ln(\pi_k/\pi_K)$ and $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{0K-1}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1})'$. It leads to the following "logistic-like" expression of conditional membership probabilities

$$t_k(\boldsymbol{x}; \boldsymbol{\beta}) = \frac{\exp(\beta_{0k} + \boldsymbol{\beta}_k' \boldsymbol{x})}{1 + \sum_{h=1}^{K-1} \exp(\beta_{0h} + \boldsymbol{\beta}_h' \boldsymbol{x})}.$$

This expression highlights the predictive focus of this model: Only useful terms for predicting the group membership (the $t_k(\boldsymbol{x}; \boldsymbol{\beta})$'s) are modeled whatever be the way the covariates $\boldsymbol{x}$ are generated. In particular, the logistic assumption includes a wide variety of families of distributions: Multivariate homoscedastic normal distributions, multivariate discrete distributions following the log-linear model with equal interaction terms, joint distributions of continuous and discrete variables of both but not necessarily independent, truncated versions of them. It implies some high flexibility which is highly appreciated by lots of practitioners in many different fields such Credit Scoring, Medicine, *etc.*

The whole parameter $\boldsymbol{\beta}$ has to be estimated from a n sample $\mathcal{S} = (\mathbf{x}, \mathbf{y})$. As previously noticed, covariates $\mathbf{x}$ arise from an unspecified mixture distribution of K groups. Then, conditionally to $\boldsymbol{x}_i$, each y_i is independently drawn from a random vector Y following the K-modal conditional multinomial distribution of order one

$$Y | \boldsymbol{X} = \boldsymbol{x} \sim \mathcal{M}_1(t_1(\boldsymbol{x}; \boldsymbol{\beta}), \ldots, t_K(\boldsymbol{x}; \boldsymbol{\beta})).$$

With this sampling scheme, the (conditional) log-likelihood to be maximized is given by

$$\ell(\boldsymbol{\beta}) = \sum_{k=1}^{K} \sum_{\{i; y_i = k\}} \ln t_k(\boldsymbol{x}_i; \boldsymbol{\beta}).$$

No closed-formed solution exists but the log-likelihood being globally concave it can have at most one maximum. A numerical optimization algorithm has to be used and generally a Newton-Raphson procedure is retained with starting parameter $\boldsymbol{\beta} = \mathbf{0}$. Note that the ML estimates do not exist when complete or quasi-complete separation occurs.

2.3.2. Transfer learning and its estimation

In case where the sample size is small, the ML estimates of the model parameters may be of poor accuracy. It may not even exist since complete or quasi-complete separation is expected to be more frequent for small sample size. In such situations, two standard solutions occur: Either some restrictions have to be made on the model by constraining the

model parameters, or the sample size has to be increased what may be difficult in many real situations (unavailable or too expensive new labeled data). An original intermediate solution is to transfer some information from another logistic regression model, estimated on a source population for which the available data are more numerous, to the previous logistic regression model. Let be Ω the source population with associated sample $\mathcal{S} = (\mathbf{x}, \mathbf{y})$ and parameter $\boldsymbol{\beta}$. Let be Ω^* the target population with associated sample $\mathcal{S}^* = (\mathbf{x}^*, \mathbf{y}^*)$ and parameter $\boldsymbol{\beta}^*$. Note that, in this regression case, all $\mathbf{y}^*$ have to be known since some $\boldsymbol{x}^*$ without its corresponding y^* value is useless in such predictive models.

Three questions naturally arise from this general idea (in italic font) with the proposed associated answers (in normal font):

1. *Why both populations Ω and Ω^* are not unrelated?* A good indicator of possible relationship between both populations is that (i) covariates $\boldsymbol{x}$ and $\boldsymbol{x}^*$ are equal or at least of same meaning and (ii) response variables y and y^* are equal or at least of same meaning.

2. *Which information is already available on $\boldsymbol{\beta}$?* An accurate estimate $\hat{\boldsymbol{\beta}}$ on $\boldsymbol{\beta}$ is easily available, typically from a sample $\mathcal{S}$ of size n greater than n^*.

3. *How both parameters $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ are linked?* A collection of simple, realistic, parsimonious and meaningful parametric links $\{\phi\}$ between both model parameters has to be proposed:

$$\boldsymbol{\beta}^* = \phi(\boldsymbol{\beta}).$$

One of the simplest link ϕ is affine. At this step it is fundamental to remind that $\boldsymbol{\beta}$ includes two kinds of parameters: The intercept parameters β_{0k} which have a translation effect on covariates $\boldsymbol{x}$ and the scale parameters $\boldsymbol{\beta}_k$ which have a scaling effect on covariates $\boldsymbol{x}$ ($k = 1, \ldots, K-1$). Thus it is meaningful to constraint the affine link ϕ to model a translation between β_{0k} and β^*_{0k} and to model a scaling between $\boldsymbol{\beta}_k$ and ${\boldsymbol{\beta}_k}^*$. Such a mapping is written

$$\beta^*_{0k} = \beta_{0k} + \delta_k \quad \text{and} \quad \boldsymbol{\beta}^*_k = \Lambda_k \boldsymbol{\beta}_k$$

where $\delta_k \in \mathbb{R}$ and Λ_k is a $d \times d$ diagonal matrix. Obviously, this model corresponds only to a reparameterization of the initial *logistic parameter* $\boldsymbol{\beta}$ into a *link parameter* (δ_k, Λ_k) ($k = 1, \ldots, K-1$). However, it is now possible to propose some meaningful and parsimonious restrictions on this link. In this aim, we propose the following constraints on (δ_k, Λ_k) inspired by the work of [2]:

- Three constraints on the translation δ_k:
  - $\delta_k = 0$: Both logistic models share a common intercept (simplest case);
  - $\delta_k = \delta$: Translation between both regressions is group independent;
  - δ_k free: Translation between both regressions is free (most general case).
- Five constraints on the scaling Λ_k:
  - $\Lambda_k = I$: Both logistic models share a common scaling (simplest case);

– $\Lambda_k = \lambda I$: Scaling between both regressions is covariate and group independent;
– $\Lambda_k = \lambda_k I$: Scaling between both regressions is covariate independent;
– $\Lambda_k = \Lambda I$: Scaling between both regressions is group independent;
– Λ_k free: Scaling between both regressions is free (most general case).

All the previous constraints on δ_k and Λ_k can be combined and it leads to 15 models of constraints on the whole link parameter (δ_k, Λ_k). These models are noted $0I$, $\delta\lambda I$, $\delta_k\Lambda_k$,... Table 6 displays the number of parameters for each of them.

Number of parameters		Λ_k				
		I	λI	$\lambda_k I$	Λ	Λ_k
δ_k	0	0	1	$K-1$	d	$d(K-1)$
	δ	1	2	K	$d+1$	$d(K-1)+1$
	δ_k	$K-1$	K	$2(K-1)$	$K+d-1$	$(d+1)(K-1)$

Table 6. Number of free parameters for each of the 15 models linking two logistic models.

We can notice some particular models:

- The simplest model $0I$ corresponds to set $\beta = \beta^*$;
- The most compex model $\delta_k\Lambda_k$ corresponds to unlinked parameters β and β^*;
- If $K = 2$, models indexed by k are equivalent to group independent models, so only 6 different models exist in this case: $0I$, $0\lambda I$, 0Λ, δI, $\delta\lambda I$, $\delta\Lambda$;
- If $d = 1$, models including $\Lambda_{(k)}$ are equivalent to models including $\lambda_{(k)}$, so only 9 different models exist in this case.

Conditionally to β, estimating β^* for a given model is easy. Indeed, since the traditional log-likelihood $\ell(\beta^*)$ is concave and since all models correspond to linear constraints on β^*, the resulting log-likelihood function to be maximized is also concave. As a consequence, there exists again at most one maximum and any optimisation algorithm subject to linear constraint can be involved.

Selecting a model can be performed either by some standard cross-validation methods or by using the BIC criterion [31].

2.3.3. Biological and Marketing applications

The last question that has to be raised about the proposed models is "are the link models realistic?" Applications are now useful for assessing this essential property. All applications will share the same following design experiment process:

1. β is estimated by maximum likehood from $\mathcal{S}$;
2. We draw from another sample $\mathcal{S}^*$, and with replacement, R samples $\mathcal{S}^*_{\tilde{n}^*,r}$ ($r = 1,\ldots,R$) of size $\tilde{n}^* \in N$ where N denotes a set of sample sizes;

3. β^* is then estimated by the ML estimate $\hat{\beta}^*_{\tilde{n}^*,r}$ for each sample $\mathcal{S}^*_{\tilde{n}^*,r}$ and for each available model;

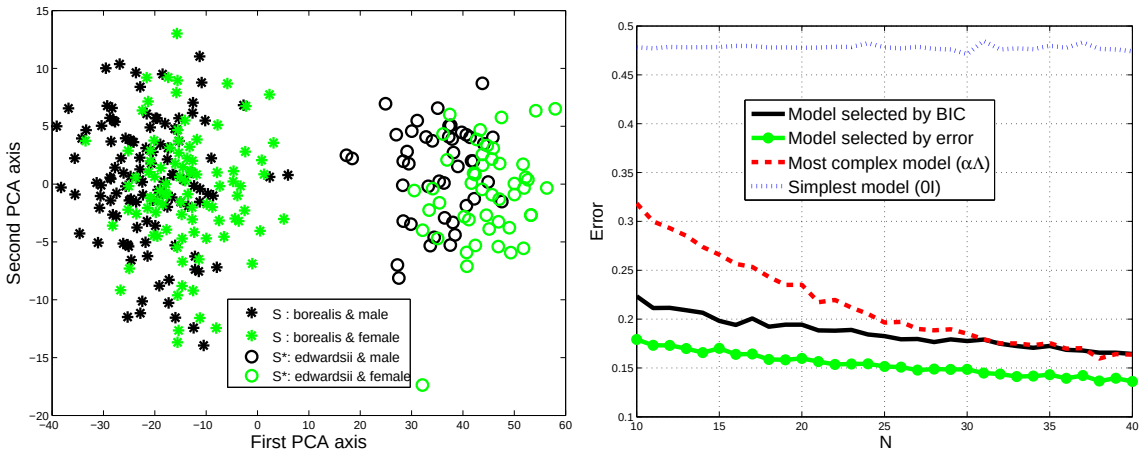
4. For each estimate $\hat{\beta}^*_{\tilde{n}^*,r}$, the error rate $e_{\tilde{n}^*,r}$ is estimated from the corresponding unused remaining sample $\mathcal{S}^* \backslash \mathcal{S}^*_{\tilde{n}^*,r}$;

5. Finally, the mean error rate $\bar{e}_{\tilde{n}^*} = R^{-1} \sum_{r=1}^{R} e_{\tilde{n}^*,r}$ is displayed for all $\tilde{n}^*$ values of N and for four models of particular interest: Model selected by BIC, model of lowest error rate, most complex model ($\delta_k \Lambda_k$), simplest model ($0I$).

Thus samples $\mathcal{S}^*_{\tilde{n}^*,r}$ will act as many samples $\mathcal{S}^*$ of different sizes, allowing to study the effect of $\tilde{n}^*$ on each model error rate.

Biology: Continuous covariates and two groups The data set has been already described in Subsection 2.1. We retain *borealis* subspecies as the data set $\mathcal{S}$ ($n = 206$) and *edwardsii* subspecies as the data set $\mathcal{S}^*$ (size $n^* = 92$). Both samples share both common biometrical features ($d = 5$) and common group meaning males/females ($K = 2$). All data are displayed in Figure 2(a) through the first two PCA axes and result of the previous design experiment process, with $N = \{10, 11, 12, \ldots, 40\}$ and $R = 300$, is displayed in Figure 2(b).

In this example, both samples $\mathcal{S}$ and $\mathcal{S}^*$ are so different that the simplest model leads to



(a) (b)

Figure 2. Birds: (a) Data on the first two PCA axes, (b) mean of the error rate $\bar{e}_{\tilde{n}^*}$ for four models of particular interest.

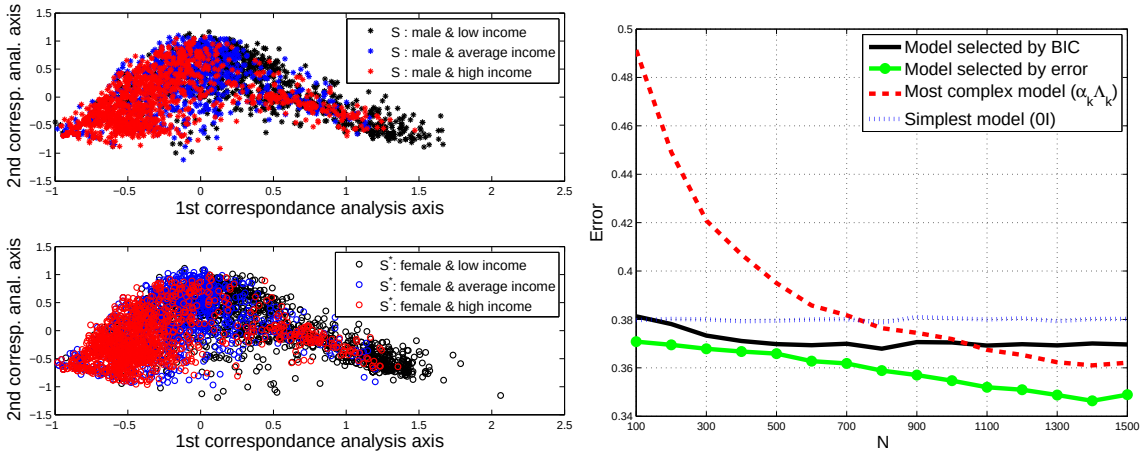
a very high error rate with subsamples $\mathcal{S}^*_{\tilde{n}^*,r}$ whatever be the sample size $\tilde{n}^*$. As expected, the more complex model is poorly efficient for sample values of $\tilde{n}^*$ but it improves when $\tilde{n}^*$ signicantly increases. The intermediate model retained by the BIC criterion allows to obtain not only a lower error rate that these too extreme models but also shows a better error rate stability through the values of $\tilde{n}^*$.

Marketing: Categorical covariates and three groups The IncomeESL data set originates from an example in the book [17]. The data set is an extract from this survey. It

consists of 8993 instances (obtained from the original data set with 9409 instances, by removing those observations with the annual income missing) with 14 categorical (factors and ordered factors) demographic attributes. It provides from questionnaires containing 502 questions which were filled out by shopping mall customers in the San Francisco Bay area in 1987.

We removed cases with missing values and divide the whole data set into two data sets according to the gender: $\mathcal{S}$ and $\mathcal{S}^*$ respectively correspond to males ($n = 3067$) and females ($n^* = 3809$). In each sample, three groups ($K = 3$) of annual income of household are considered as the response variable: Low income (Less than \$19,999), average income (\$20,000 to \$39,999) and high income (\$40,000 or more). The goal is to predict the annual income of household from the $d = 12$ remaining categorical covariates of two or more modalities and corresponding to demographic attributes. Figure 3(a) displays the first two MCA axes and result of the previous design experiment process, with $N = \{100, 200, 300, \ldots, 1500\}$ and $R = 50$, is displayed in Figure 3(b).

Again, we can see both a nice error rate stability and a low error rate of the intermediate



(a) (b)

Figure 3. Birds: (a) Data on the first two MCA axes, (b) mean of the error rate $\vec{e}_n^h$ for four models of particular interest.

model retained by the BIC criterion through the values of $\tilde{n}^*$. The simplest model is here more challenging than in the previous biological example since income difference between males and females exists but is quite moderate. A large sample size $\tilde{n}^*$ is required for obtaining good results for the complex model. Thus, the new models act as powerful adaptive challengers for standard models.

3. Parametric transfer learning in regression

Linear regression and mixture of regressions are two very popular techniques to establish a relationship between a quantitative response $y \in \mathcal{Y} = \mathbb{R}$ variable and one or several explanatory variables $\boldsymbol{x}$. However, as in the classification context, most of the regression methods assume the absence of evolution in the modeled phenomenon between the training and the

prediction stages. This section presents parametric transformation models which allows both regression models to deal with evolving populations.

3.1. Linear regression

Linear regression assumes that the response variable $Y \in \mathbb{R}$ can be linked to the explanatory variables $\boldsymbol{x} \in \mathbb{R}^d$ though the relation:

$$Y = \sum_{j=0}^{p} \beta_j \psi_j(\boldsymbol{x}) + \varepsilon,$$

where the residuals $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ are independent, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)' \in \mathbb{R}^{p+1}$ are the regression parameters, $\psi_0(\boldsymbol{x}) = 1$ and $(\psi_j)_{1 \leq j \leq p} : \mathbb{R}^d \to \mathbb{R}$ is a basis of regression functions. The regression functions can be for instance the identity, polynomial functions or splines functions [17]. Let us notice that the usual linear regression occurs when $d = p$ and $\psi_j(\boldsymbol{x}) = x_j$ for $j = 1, \ldots, d$. This model is equivalent to the distributional assumption:

$$Y | \boldsymbol{X} = \boldsymbol{x} \sim \mathcal{N}(g(\boldsymbol{x}, \boldsymbol{\beta}), \sigma^2),$$

where the regression function $g(\boldsymbol{x}, \boldsymbol{\beta}) = \sum_{j=0}^{p} \beta_j \psi_j(\boldsymbol{x})$ is defined as the conditional expectation $E[Y|\boldsymbol{x}]$. Notice that the regression function can be also written in a matrix form as follows:

$$Y = \boldsymbol{\beta}' \Psi(\boldsymbol{x}) + \varepsilon, \tag{10}$$

where $\Psi(\boldsymbol{x}) = (1, \psi_1(\boldsymbol{x}), \ldots, \psi_p(\boldsymbol{x}))'$.

Learning such a model from a training sample $\mathcal{S} = (\mathbf{x}, \mathbf{y})$ is usually straightforward and relies on ordinary least square (OLS) estimation.

3.1.1. Transfer learning for linear regression

Let us now assume that the regression parameters $\boldsymbol{\beta}$ have been estimated in a preliminary study by using a sample $\mathcal{S}$ of the source population Ω, and that a new regression model has to be adjusted on a new sample $\mathcal{S}^* = (\mathbf{x}^*, \mathbf{y}^*)$, measured on the same explanatory variables but arising from another population Ω^* and for which n^* is assumed to be quite small. The difference between Ω and Ω^* can be for instance geographical or temporal as described in examples of this chapter introduction. The new regression model for Ω^* can be classically written:

$$Y^* | \boldsymbol{X}^* = \boldsymbol{x}^* \sim \mathcal{N}(\boldsymbol{\beta}^{*\prime} \Psi^*(\boldsymbol{x}^*), \sigma^{2*}), \tag{11}$$

The statistical transformation model aims therefore to define a link between the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. In order to exhibit a link between both regression functions, we make the following important assumptions:

- $\mathcal{A}_1$: We first postulate that the number of basis functions and the basis functions themselves are the same for both regression models ($p^* = p$ and $\psi_j^* = \psi_j$, $\forall j = 1, ..., p$).

- $\mathcal{A}_2$: We also assume that the transformation between $g(\bullet,\boldsymbol{\beta})$ and $g(\bullet,\boldsymbol{\beta}^*)$ applies only on the regression parameters. We therefore define the transformation matrix $\boldsymbol{\Lambda}$ between the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ such that $\boldsymbol{\beta}^* = \boldsymbol{\Lambda}\boldsymbol{\beta}$.

- $\mathcal{A}_3$: We finally assume that the relation between the response variable and a specific covariate in the new population $\boldsymbol{\Omega}^*$ only depends on the relation between the response variable and the same covariate in the population $\boldsymbol{\Omega}$. Thus, for $j = 0,\ldots,p$, the regression parameter β_j^* only depends on the regression parameter β_j (*i.e.* Λ is diagonal).

The transformation can be finally written in term of the regression parameters of both models as follows:

$$\beta_j^* = \lambda_j \beta_j \qquad \forall j = 0,\ldots,p, \tag{12}$$

where $\lambda_j \in \mathbb{R}$ is the j-th diagonal element of Λ. As previously, it is possible to make further assumptions on the transformation model to makes it more parsimonious. For instance, we allow some of the parameters λ_j to be equal to 1 (in this case the regression parameters β_j^* are equal to β_j). We also allow some of the parameters λ_j to be equal to a common value, *i.e.* $\lambda_j = \lambda$ for given $0 \leq j \leq d$. We list below some of the possible models as declined in [5].

- Model M_0: $\beta_0^* = \lambda_0\beta_0$ and $\beta_j^* = \lambda_j\beta_j$, for $j = 1,...,p$. This model is the most complex model of transformation between the populations $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^*$. It is equivalent to learn a new regression model from the sample $\mathcal{S}^*$, since there is no constraint on the $p+1$ parameters β_j^* ($j = 0,...,p$), and the number of free parameters in $\boldsymbol{\Lambda}$ is consequently $p+1$ as well.

- Model M_1: $\beta_0^* = \beta_0$ and $\beta_j^* = \lambda_j\beta_j$ for $j = 1,...,p$. This model assumes that both regression models have the same intercept β_0.

- Model M_2: $\beta_0^* = \lambda_0\beta_0$ and $\beta_j^* = \lambda\beta_j$ for $j = 1,...,p$. 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 M_3: $\beta_0^* = \lambda\beta_0$ and $\beta_j^* = \lambda\beta_j$ for $j = 1,...,p$. This model assumes that all the regression parameters of both regression models differ by the same scalar λ.

- Model M_4: $\beta_0^* = \beta_0$ and $\beta_j^* = \lambda\beta_j$ for $j = 1,...,p$. 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 M_5: $\beta_0^* = \lambda_0\beta_0$ and $\beta_j^* = \beta_j$ for $j = 1,...,p$. This model assumes that both regression models have the same parameters except the intercept.

- Model M_6: $\beta_0^* = \beta_0$ and $\beta_j^* = \beta_j$ for $j = 1,...,p$. This model assumes that both populations $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^*$ have the same regression model.

The numbers of parameters to estimate for these transformation models are presented in Table 7. 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 practicionners to consider some additional transformation models specifically designed to his application and motivated by his prior knowledge on the subject.

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

Table 7. Number of parameters to estimate for the transfer learning models in the linear regression case.

The estimation of the parameters β^* can be deduced from the estimation of the link parameters Λ. Least square estimators for the models M_1 to M_5 are derived in [5].

3.1.2. Biological application

A biological dataset is considered here to highlight the ability of our approach to deal with real data. The *hellung* dataset[1], collected by P. Hellung-Larsen, reports the growth conditions of *Tetrahymena* cells. The data arise from two groups of cell cultures: Cells with and without glucose added to the growth medium. For each group, the average cell diameter (in μm) and the cell concentration (count per ml) were recorded. The cell concentrations of both groups were set to the same value at the beginning of the experiment and it is expected that the presence of glucose in the medium affects the growth of the cell diameter. In the sequel, cells with glucose will be considered as the source population Ω ($n = 32$ observations) whereas cells without glucose will be considered as the target population Ω^* (between $n^* = 11$ to 19 observations).

In order to fit a regression model on the cell group with glucose, the PRESS criterion was used to select the most appropriate basis function. It results that a 3rd degree polynomial function is the most adapted model for these data and this specific basis function will be used for all methods in this experiment. The 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 Ω^* training dataset. For this, 4 different training datasets are used: All Ω^* observations (19 obs.), all Ω^* 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, we compute for these 4 different sizes of training dataset the PRESS criterion [1], which represents the mean squared prediction error computed on a cross-validation scheme. This criterion is one of the most often used for model selection in regression analysis, and we encourage its use when it is computationally feasible. In addition, the MSE value (Mean Square Error) on the whole Ω^* dataset is also computed.

Figure 4 illustrates the effect of the training 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 (M_0) in addition to the curves of the 5 transfer learning models (models

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

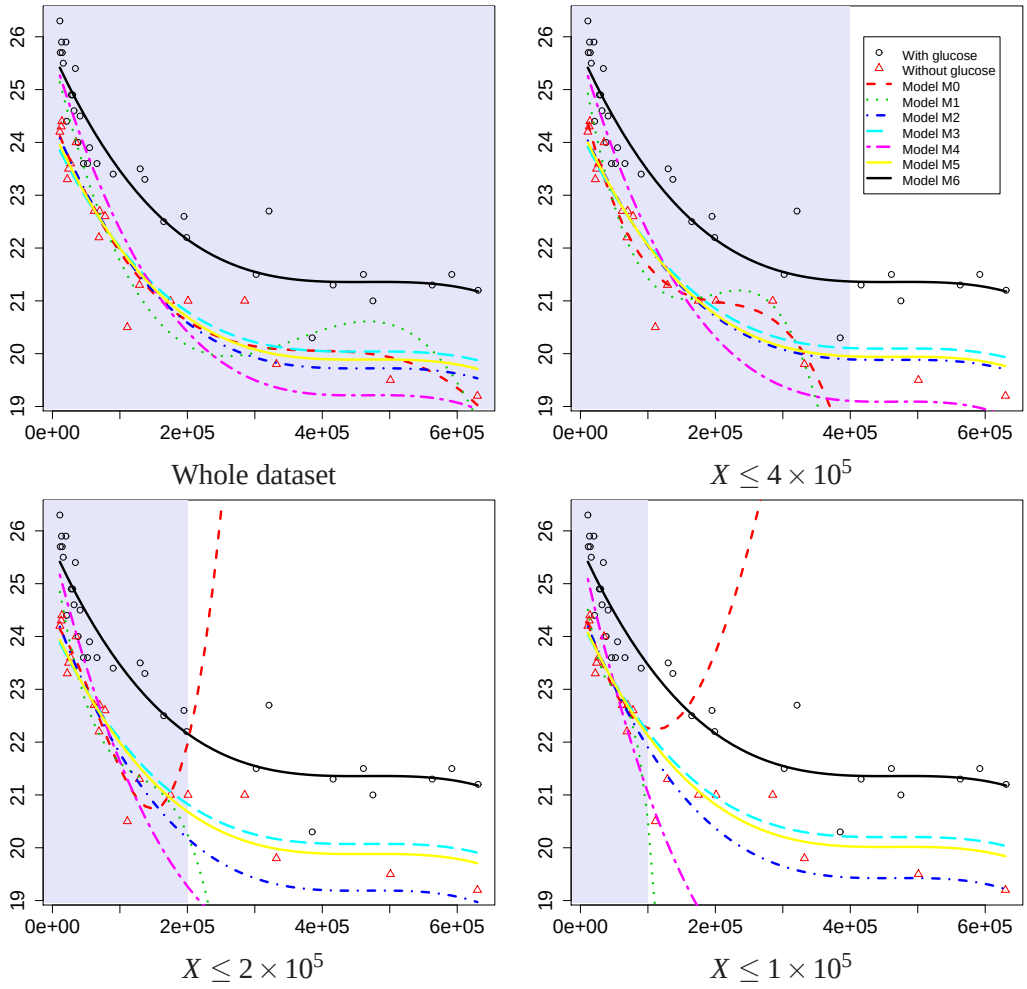


Figure 4. Effect of the learning set size on the prediction ability of the studied regression methods for the *hellung* dataset. The blue zones correspond to the parts of the observations of Ω^* used for learning the models.

M_1 to M_5) for different sizes of the training set (the blue zones indicate the ranges of the observations of Ω^* used for training the models). The model M_6 which is equivalent to the usual OLS regression method on the population Ω is also displayed. The first remark suggested by these results is that the most complex models, OLS (M_0) and M_1, appear to be very unstable in such a situation where the number of training observations is small. Secondly, the model M_4 is more stable but its main assumption (same intercept as the regression model of Ω) seems to be an overly strong constraint and stops it from fitting correctly the data. Finally, the models M_2, M_3 and M_5 turn out to be very stable and flexible enough to correctly model the target population Ω^* even with very few observations. This visual interpretation of the experiment is confirmed by the numerical results presented in Tables 8 and 9. These tables respectively report the value of the PRESS criterion and the MSE associated to the studied regression methods for the different sizes of training dataset. Table 8

confirms clearly that the most stable, and therefore appropriate, model for estimating the transformation between populations Ω and Ω^* is the model M_5. Another interesting conclusion is that both models M_2 and M_3 obtained very low PRESS values as well. These predictions of the model stability appear to be satisfying since the comparison of Tables 8 and 9 shows that the model selected by the PRESS criterion is always an efficient model for prediction. Indeed, Table 9 shows that the most efficient models in practice are the models M_2 and M_5 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.

Method	whole dataset	$X \leq 4\times10^5$	$X \leq 2\times10^5$	$X \leq 1\times10^5$
OLS on Ω^* (M_0)	0.897	0.364	0.432	0.303
Model M_1	3.332	0.283	2.245	0.344
Model M_2	0.269	0.294	0.261	0.130
Model M_3	0.287	0.271	0.289	0.133
Model M_4	0.859	1.003	0.756	0.517
Model M_5	**0.256**	**0.259**	**0.255**	**0.124**

Table 8. 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	$X \leq 4\times10^5$	$X \leq 2\times10^5$	$X \leq 1\times10^5$
OLS on Ω^* (M_0)	0.195	47.718	4.5×10^3	145.846
Model M_1	0.524	164.301	2.3×10^3	5.9×10^5
Model M_2	**0.218**	**0.226**	0.304	**0.245**
Model M_3	0.258	0.262	0.259	0.290
Model M_4	0.791	0.796	1.472	3.046
Model M_5	*0.230	*0.233	***0.230**	*0.246
OLS on Ω (M_6)	2.388	2.388	2.388	2.388

Table 9. 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.

As it could be expected, the advantage of adaptive linear models makes particularly sense when the number of observations of the target population is limited and this happens frequently in real situations due to censorship or to technical constraints (experimental cost, scarcity,...).

3.2. Mixture of regressions

The mixture of regressions, introduced by [15] as the switching regression model and also named clusterwise linear regression model in [18], is a popular regression model for modeling complex systems for which the linear regression model is not flexible enough. In particular, the switching regression model is often used in Economics for modeling phenomena with different phases. Figure 5 illustrates such a situation.

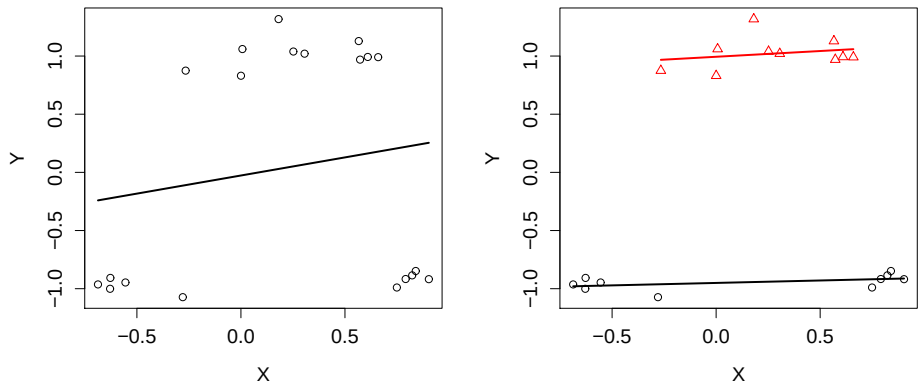


Figure 5. Modelling of a two-state phenomena with the linear regression model (left) and the regression mixture model (right).

This model assumes that the dependent variable $Y \in \mathcal{Y} = \mathbb{R}$ can be linked to a covariate $\tilde{\boldsymbol{x}} = (1, \boldsymbol{x}) \in \mathbb{R}^{d+1}$ by one of K possible regression models:

$$Y = \boldsymbol{\beta}_k' \Psi(\tilde{\boldsymbol{x}}) + \sigma_k \varepsilon, \; k = 1, ..., K$$

with mixing proportions $\pi_1, \ldots, \pi_K$, where $\varepsilon \sim \mathcal{N}(0,1)$, $\boldsymbol{\beta}_k = (\beta_{k0}, ..., \beta_{kd}) \in \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$ is the regression parameter vector in $\mathbb{R}^{d+1}$ and where $\sigma_k^2 \in \{\sigma_1^2, \ldots, \sigma_K^2\}$ is the residual variance. The conditional density distribution of Y given $\tilde{\boldsymbol{x}}$ is therefore:

$$f(y|\tilde{\boldsymbol{x}}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, f_k(y; \boldsymbol{\beta}_k^t \Psi(\tilde{\boldsymbol{x}}), \sigma_k^2),$$

where $f_k(\bullet; \boldsymbol{\beta}_k^t \Psi(\tilde{\boldsymbol{x}}), \sigma_k^2)$ is the univariate Gaussian density of mean $\boldsymbol{\beta}_k' \Psi(\boldsymbol{x})$ and variance σ_k^2. We have also noted $\boldsymbol{\theta} = ((\pi_k, \boldsymbol{\beta}_k, \sigma_k^2) : k = 1, \ldots, K)$. For such a model, the prediction of y for a new observed covariate $\tilde{\boldsymbol{x}}$ is usually carried out in two steps: First the component membership of the data is estimated by the MAP rule following the same principle as in (1) and then y is predicted using the selected regression model.

3.2.1. Transfer learning for mixture of regressions

We make the same assumptions $\mathcal{A}_1$ to $\mathcal{A}_3$ as in the linear regression case and we assume in addition that each mixture is assumed to have the same number of components (*i.e.* $K^* = K$). Conditionally to an observation $\boldsymbol{x}$ of the covariates, we would like to exhibit a distributional relationship between the dependent variables of the same mixture component from available samples $\mathcal{S} = (\mathbf{x}, \mathbf{y})$ and $\mathcal{S}^* = (\mathbf{x}^*, \mathbf{y}^*)$. Let $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}_k^*$ $(1 \leq k \leq K)$ be respectively the parameters of the mixture regression models in the source and the target populations Ω and Ω^* respectively. [5] assume that the distributional relationship consists of a the following parametric link between the regression parameters of both populations:

$$\boldsymbol{\beta}_k^* = \Lambda_k \boldsymbol{\beta}_k, \text{ where } \Lambda_k = \text{diag}(\lambda_{k0}, \lambda_{k1}, \ldots, \lambda_{kd}) \quad \text{and} \quad \sigma_k^* \text{ is free,}$$

where $\text{diag}(\lambda_{k0},\lambda_{k1},\ldots,\lambda_{kd})$ is the diagonal matrix containing $(\lambda_{k0},\lambda_{k1},\ldots,\lambda_{kd})$ on its diagonal. In order to introduce parsimony some constraints are put on Λ_k and σ_k^*. Such defined parsimonious models include many of the situations that may be encountered in practice:

- MM_1 assumes both populations are the same: $\Lambda_k = I_d$ is the identity matrix ($\sigma_k^* = \sigma_k$),
- MM_2 models assume the link between both populations is covariate and mixture component independent:
  - MM_{2a} : $\lambda_{k0} = 1$, $\lambda_{kj} = \lambda$ and $\sigma_k^* = \lambda\sigma_k \quad \forall 1 \leq j \leq d$,
  - MM_{2b} : $\lambda_{k0} = \lambda$, $\lambda_{kj} = 1$ and $\sigma_k^* = \sigma_k \quad \forall 1 \leq j \leq d$,
  - MM_{2c} : $\Lambda_k = \lambda I_d$ and $\sigma_k^* = \lambda\sigma_k$,
  - MM_{2d} : $\lambda_{k0} = \lambda_0$, $\lambda_{kj} = \lambda_1$ and $\sigma_k^* = \lambda_1\sigma_k \quad \forall 1 \leq j \leq d$,
- MM_3 models assume the link between both populations is covariate independent:
  - MM_{3a} : $\lambda_{k0} = 1$, $\lambda_{kj} = \lambda_k$ and $\sigma_k^* = \lambda_k\sigma_k \quad \forall 1 \leq j \leq d$,
  - MM_{3b} : $\lambda_{k0} = \lambda_k$, $\lambda_{kj} = 1$ and $\sigma_k^* = \sigma_k \quad \forall 1 \leq j \leq d$,
  - MM_{3c} : $\Lambda_k = \lambda_k I_d$ and $\sigma_k^* = \lambda_k\sigma_k$,
  - MM_{3d} : $\lambda_{k0} = \lambda_{k0}$, $\lambda_{kj} = \lambda_{k1}$ and $\sigma_k^* = \lambda_{k1}\sigma_k \quad \forall 1 \leq j \leq d$,
- MM_4 models assume the link between both populations is mixture component independent (σ_k^* free):
  - MM_{4a} : $\lambda_{k0} = 1$ and $\lambda_{kj} = \lambda_j \quad \forall 1 \leq j \leq d$,
  - MM_{4b} : $\Lambda_k = \Lambda$ with Λ a diagonal matrix,
- MM_5 assumes Λ_k is unconstrained, which leads to estimate the mixture regression model for Ω^* by using only $\mathcal{S}^*$ (σ_k^* free).

Moreover, the mixing proportions are allowed to be the same in each population or to be different. In the latter case, they consequently have to be estimated using the sample $\mathcal{S}^*$. Corresponding notations for the models are respectively $\pi MM_\bullet$ when the mixing proportions π_k^* of Ω^* have to be estimated and $MM_\bullet$ when the have not. Table 10 gives the number of parameters to estimate for each model. If the mixing proportions are different from Ω to Ω^*, $K-1$ parameters to estimate must be added to these values.

Model	MM_1	MM_{2a-c}	MM_{2d}	MM_{3a-c}	MM_{3d}	MM_{4a}	MM_{4b}	MM_5
Param. nb.	0	1	2	K	$2K$	$d+K$	$d+K+1$	$K(d+2)$

Table 10. Number of parameters to estimate for the transfer learning models in the regression mixture case.

Model inference The estimation of the link parameters is carried out by ML using a missing data approach *via* the EM algorithm [11]. Indeed, we do not know from which component of the mixture arises each observation $(\boldsymbol{x}_i, y_i)$ or also each observation $(\boldsymbol{x}_i^*, y_i^*)$. This technique is certainly the most popular approach for inference in mixtures of regressions (MCMC approaches can be also used, see [6]). More details on the EM algorithm for the link parameter estimation can be found in [6]. Then, by plug-in, it provides an estimate $\hat{\boldsymbol{\theta}}^*$ of $\boldsymbol{\theta}^*$.

Prediction rule Once the model parameters have been estimated, the prediction $\hat{y}^*$ of the response variable corresponding to an observation $\boldsymbol{x}^*$ of $\mathcal{X}$ is obtained by a two step procedure. First, the component membership of $\boldsymbol{x}^*$ is estimated by the MAP rule. Then, $\hat{y}^*$ is predicted using the kth regression model of the mixture:

$$\hat{y}^* = \hat{\boldsymbol{\beta}}_k^{*\prime} \boldsymbol{\Psi}(\tilde{\boldsymbol{x}}^*).$$

Model selection In order to select among the differents transfer learning models the most appropriate model of transformation between the populations Ω and Ω^* , we propose to use two well-known criteria, already presented: PRESS and BIC. Let us recall that, for both criteria, the most adapted model is the one with the smallest criterion value.

3.2.2. Economic-environmental application

In this experiment, the link between CO_2 emission and gross national product (GNP) of various countries is investigated. The sources of the data are *The official United Nations site for the Millennium Development Goals Indicators* and the *World Development Indicators of the World Bank*. Figure 6 plots the CO_2 emission per capita *versus* the logarithm of GNP per capita for 111 countries, in 1980 (left) and 1999 (right).

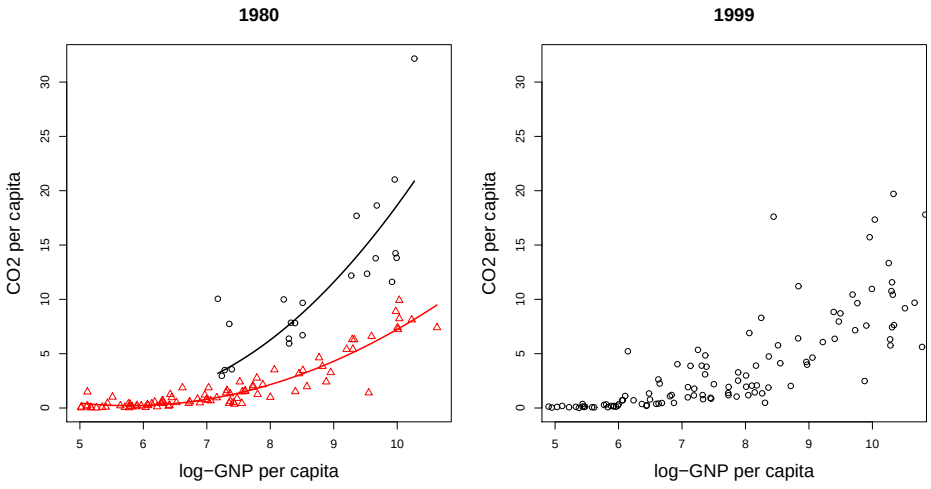


Figure 6. Emission of CO_2 per capita *versus* GNP per capita in 1980 (left) and 1999 (right).

A mixture of second order polynomial regressions seems to be particularly well adapted to fit these data and will be used in the following. Let remark that regression model with heteroscedasticity could also be appropriated for such data, but these kind of models are out of the topic of the present work. For the 1980's data, two groups of countries are easily distinguishable: A first minority group (about 25% of the whole sample) is made of countries for which a grow in the GNP is linked to a high grow of the CO_2 emission, whereas the second group (about 75%) seems to have more environmental political orientations. As pointed out by [20], the study of such data could be particularly useful for countries with low GNP in order to clarify in which development path they are embarking. This country discrimination in two groups is more difficult to obtain on the 1999's data: It seems that countries which had high CO_2 emission in 1980 have adopted a more environmental development than in the past, and a two-component mixture regression model could be more difficult to exhibit.
In order to help this distinction, parametric transfer learning models are used to estimate the mixture regression model on the 1999's data. The ten parametric transfer learning models, with free component proportions π_k^*, πMM_{2a} to πMM_{4b}, classical mixture of second order polynomial regressions with two components (MR) and usual second order polynomial regression (UR) are considered. Different sample size of the 1999's data are tested: 30%, 50%, 70% and 100% of the $\mathcal{S}^*$ size ($n^* = 111$). The experiments have been repeated 20 times in order to average the results. Table 11 summarizes these results where MSE corresponds to the mean square error. In this application, the total number of available data in the 1999 population is not sufficiently large to separate them into two training and test samples. For this reason, MSE is computed on the whole $\mathcal{S}^*$ sample, even though a part of it has been used for the training (from 30% for the first experiment to 100% for the last one). Consequently, MSE is a significant indicator of predictive ability of the model when 30% and 50% of the whole dataset are used as training set since 70% and 50% of the samples used to compute the MSE remain independent from the training stage. However, MSE is a less significant indicator of predictive ability for the two last experiments and the PRESS should be preferred in these situations as indicator of predictive ability.

Table 11 first allows to remark that the 1999's data are actually made of two components as in the 1980's data since both PRESS and MSE are better for MR (2 components) than UR (1 component) for all sizes n^* of $\mathcal{S}^*$. This first result validates the assumption that both the reference population Ω and the new population Ω^* have the same number $K = 2$ components, and consequently the use of transfer learning techniques makes sense for this data. Secondly, AMR models turn out to provide very satisfying predictions for all values of n^* and particularly outperforms the other approaches when n^* is relatively small (less than 77 here). Indeed, both BIC, PRESS and MSE testify that the transfer learning models provide better predictions than the other studied methods when n^* is equal to 30%, 50% and 70% of the whole sample. Furthermore, it should be noticed that transfer learning models provide stable results according to variations on n^*. In particular, the models πMM_2 are those which appear the most efficient on this dataset and this means that the link between both populations Ω and Ω^* is mixture component independent.

This application illustrates the interest of combining informations on both past (1980) and present (1999) situations in order to analyse the link between CO_2 emissions and gross national product for several countries in 1999, especially when the number of data for the

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

model	BIC	PRESS	MSE
πMM_{2a}	13.09	**3.38**	3.40
πMM_{2b}	12.73	3.89	**3.32**
πMM_{2c}	12.79	5.48	3.68
πMM_{2d}	11.54	4.99	3.73
πMM_{3a}	12.14	4.20	3.76
πMM_{3b}	11.72	4.87	4.00
πMM_{3c}	**11.50**	5.09	3.86
πMM_{3d}	22.83	5.52	3.64
πMM_{4a}	18.72	5.15	4.01
πMM_{4b}	22.01	6.21	5.04
UR	27.08	7.46	7.66
MR	32.89	5.54	5.11

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

model	BIC	PRESS	MSE
πMM_{2a}	**10.18**	4.11	3.44
πMM_{2b}	13.54	3.73	3.37
πMM_{2c}	13.89	4.25	3.45
πMM_{2d}	22.35	4.38	4.80
πMM_{3a}	12.00	3.84	4.49
πMM_{3b}	12.00	4.47	3.86
πMM_{3c}	17.53	3.97	**3.28**
πMM_{3d}	25.39	4.77	3.67
πMM_{4a}	20.65	**3.68**	3.44
πMM_{4b}	24.92	5.57	4.19
UR	20.87	7.95	7.21
MR	39.69	4.82	4.77

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

model	BIC	PRESS	MSE
πMM_{2a}	14.76	**3.65**	3.35
πMM_{2b}	14.73	3.91	3.39
πMM_{2c}	**14.53**	4.49	3.53
πMM_{2d}	18.90	4.30	3.72
πMM_{3a}	18.84	4.33	3.85
πMM_{3b}	18.80	4.40	3.85
πMM_{3c}	18.81	4.41	3.26
πMM_{3d}	27.05	3.91	**3.17**
πMM_{4a}	22.29	5.25	4.00
πMM_{4b}	26.55	4.92	4.03
UR	22.08	8.00	7.10
MR	43.91	5.06	3.33

($n^* = 111$)

model	BIC	PRESS	MSE
πMM_{2a}	15.51	4.78	3.32
πMM_{2b}	15.44	3.81	3.37
πMM_{2c}	**15.39**	4.84	3.47
πMM_{2d}	20.05	4.45	3.59
πMM_{3a}	20.18	4.29	3.79
πMM_{3b}	20.03	4.38	3.77
πMM_{3c}	20.05	3.94	3.10
πMM_{3d}	29.37	4.08	3.34
πMM_{4a}	23.98	4.21	4.13
πMM_{4b}	28.58	5.21	4.52
UR	23.62	7.53	6.99
MR	47.19	**3.66**	**2.89**

Table 11. MSE on the whole 1999's sample, PRESS and BIC criterion for the 10 parametric transfer learning models (πMM_{2a} to πMM_{4b}), usual regression model (UR) and classical regressions mixture model (MR), for 4 sizes of the 1999's sample: 33, 55, 77 and 111 (whole sample). Lower BIC, PRESS and MSE values for each sample size are in bold character.

present situation is not sufficiently large. Moreover, the competition between the parametric transfer learning models is also informative. Effectively, it seems that three models are particularly well adapted to model the link between the 1980's data and those of 1999's data: πMM_{2a}, πMM_{2b} and πMM_{2c}. The particularity of these models is that they consider the same transformation for both classes of countries, which means, conversely to what one might *prima facie* have thought, that all the countries have made an effort to reduce their CO_2 emissions and not only those which had the higher ones.

4. Parametric transfer learning in clustering

4.1. The statistical model

Clustering aims to partition a sample $\mathcal{S} = \mathbf{x}$ of n observed data into K groups. The standard model-based clustering procedure assumes that any observed data $\boldsymbol{x}_i \in \mathcal{S}$ $(i = 1,\ldots,n)$ is i.i.d. drawn from a random vector $\boldsymbol{X}$ of the K-component mixture of parametric distributions $f(\bullet;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\bullet;\boldsymbol{\alpha}_k)$. We recall that π_k denotes the mixing proportions of the component k, $\boldsymbol{\alpha}_k$ its parameter and $\boldsymbol{\theta} = \{(\pi_k,\boldsymbol{\alpha}_k) : k = 1,\ldots,K\}$ the whole mixture parameter. It can be assumed equivalently that (i) $\boldsymbol{x}_i$ has been generated by the component $y_i \in \{1,\ldots,K\}$ with probability π_{y_i} and that (ii) the component of origin y_i is a lost information. Thus the vector $\mathbf{y} = (y_1,\ldots,y_n)$ constitutes some hidden data (see [27], p. 7).

The clustering procedure consists on three stages. First, the parameter $\boldsymbol{\theta}$ has to be estimated by the ML principle, usually by the mean of an EM algorithm. Second observed data $\mathbf{x}$ are allocated by the MAP rule to the group corresponding to the highest estimated conditional probability of membership computed at the ML estimate $\hat{\boldsymbol{\theta}}$: See (1) and (2) in this chapter (see also [27] p. 31). Finally, the BIC criterion is commonly used for selecting some parsimonious model and/or the number of clusters (see [14, 30]).

4.2. Gaussian transfer learning and its estimation

Thereafter we aim to partition, in a Gaussian context, not a single sample, but H n^h-samples $\mathcal{S}^h = \mathbf{x}^h$ $(h = 1,\ldots,H)$, with $\mathbf{x}^h = (\boldsymbol{x}_1^h,\ldots,\boldsymbol{x}_{n^h}^h)$, described by a set of d continuous variables (so $\mathcal{X} = \mathbb{R}^d$) into K groups each. Thus, in this section, we are no longer limited to two populations, a source one (Ω) and a target one (Ω^*), but we are in a more general situation where H populations $\Omega^1,\ldots,\Omega^H$ are present. Note that all populations will now play a symmetric role (instead of previous discriminant analysis and regression situations), so the concept of "source" and "target" population becomes totally arbitrary and unimportant. In addition, we make the assumptions that (i) the samples share statistical units of same nature, (ii) they are described by identical features and (iii) all researched partitions $\mathbf{y}^h = (y_1^h,\ldots,y_{n^h}^h)$ share the same meaning.

Such situations are numerous and it is easily to exhibited examples through well-known data sets. For instance the Old Faithful geyser located in the Yellowstone National Park (Wyoming, USA) is regularly subject to clustering investigations. Figure 7 displays two samples of the geyser eruptions differing over ten years and described by the same variables: Duration and inter-eruption waiting time interval (in minutes). Some structure of the eruptions is frequently researched within one of these samples, but never by using the whole information that they both provide. Sections 4.3. and 4.4. present also two other situations, the first one in Biology and the second one in Finance, where several samples of identical statistical units described by the same features, have to be clustered into same meaning partitions.

4.2.1. Independent clustering of several populations

Standard Gaussian model-based clustering assumes that each individual $\boldsymbol{x}_i^h$ of the sample $\mathbf{x}^h$ $(h \in \{1,\ldots,H\})$ is i.i.d. drawn from a population Ω^h modelled by a mixture of K nor-

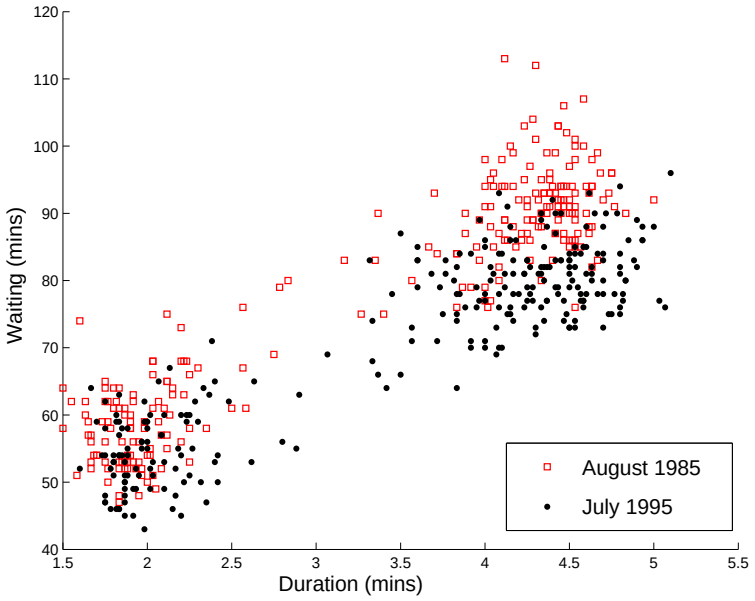


Figure 7. Two samples of Old Faithful geyser eruptions differing over 10 years.

mal d-dimensional distributions. In addition, all samples $\mathbf{x}^h$ are mutually independent. The component k his weighted by $\pi_k^h > 0$ ($\sum_{j=1}^K \pi_j^h = 1$), centered in $\boldsymbol{\mu}_k^h \in \mathbb{R}^d$ with covariance matrix $\boldsymbol{\Sigma}_k^h \in \mathbb{R}^{d\times d}$ (symmetric positive-definite) and $\mathbf{y}^h = (y_1^h, \ldots, y_K^h)$ is the missing response variable indicating the component from which each $\boldsymbol{x}_i^h$ arises. So the mixture $\boldsymbol{\Omega}^h$ is entirely parameterized by $\boldsymbol{\theta}^h = \{(\pi_k^h, \boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h) : k = 1, \ldots, K\}$.

The standard independent procedure considers that the populations $\boldsymbol{\Omega}^h$ ($h = 1, \ldots, H$) are unrelated and so, that the parameters $\boldsymbol{\theta}^h$ are algebraically free. Then estimating the whole model parameter $\boldsymbol{\theta} = (\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^H)$ by maximum likelihood, is equivalent to estimating each parameter $\boldsymbol{\theta}^h$ independently from the others. Indeed, $\ell(\boldsymbol{\theta}) = \sum_{h=1}^H \ell^h(\boldsymbol{\theta}^h)$ where $\ell(\boldsymbol{\theta})$ is the log-likelihood of $\boldsymbol{\theta}$ computed on the whole observed data $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^H)$, and $\ell^h(\boldsymbol{\theta}^h)$ is the log-likelihood of $\boldsymbol{\theta}^h$ computed on the observed data $\mathbf{x}^h$ only. Then, all partitions $\mathbf{y}^h$ are estimated by performing the MAP rule on observed data $\mathbf{x}^h$ by using the obtained ML estimate of $\boldsymbol{\theta}^h$.

That first standard method proceeds as if the diverse samples to be classified were unrelated and prohibits any transfer learning between the populations. But let us remind that (i) the samples share statistical units of same nature (ii) they are described by the same features and (iii) the groups that have to be discovered consist in a same meaning partition of each sample. The simultaneous clustering [23] method that we present now, formalizes the previous informations by establishing a link between the populations, in order to improve the model fit and the estimated partitions.

4.2.2. Simultaneous clustering of several populations

Let assume that observations of $(\mathbf{x}^h, \mathbf{y}^h)$ are i.i.d. realizations of a random couple $(\boldsymbol{X}^h, Y^h)$. For all $(h,h') \in \{1,\ldots,H\}^2$ and all $k \in \{1,\ldots,K\}$, we suppose that there exist $\boldsymbol{D}_k^{h,h'} \in \mathbb{R}^{d\times d}$ diagonal, regular, positive and $\boldsymbol{b}_k^{h,h'} \in \mathbb{R}^d$ such that:

$$\boldsymbol{X}^{h'}_{|Y^{h'}=k} \stackrel{\mathscr{D}}{=} \boldsymbol{D}_k^{h,h'} \boldsymbol{X}^{h}_{|Y^{h}=k} + \boldsymbol{b}_k^{h,h'}. \tag{13}$$

Some arguments for justifying this affine form have been already discussed in Section 2.1.2. and can be also find in [23]. Note that the condition that $\boldsymbol{D}_k^{h,h'}$ be positive is imposed for identifiability reasons. This positivity involves that the correlation sign of any couple of conditional variables keeps unchanged through the populations. That constraint seems to be realistic in many experimental situations.

Equivalently to (13), the following parametric link can be established between populations. Whatever are k, h, h', there exist some diagonal positive-definite matrix $\boldsymbol{D}_k^{h,h'} \in \mathbb{R}^{d\times d}$ and some vector $\boldsymbol{b}_k^{h,h'} \in \mathbb{R}^d$, such that:

$$\boldsymbol{\Sigma}_k^{h'} = \boldsymbol{D}_k^{h,h'}\, \boldsymbol{\Sigma}_k^{h}\, \boldsymbol{D}_k^{h,h'} \text{ and } \boldsymbol{\mu}_k^{h'} = \boldsymbol{D}_k^{h,h'} \boldsymbol{\mu}_k^{h} + \boldsymbol{b}_k^{h,h'}. \tag{14}$$

Property (14) characterizes henceforward the whole parameter space of $\boldsymbol{\theta}$ and the so-called simultaneous clustering method is based on $\boldsymbol{\theta}$ parameter inference in that so constrained parameter space.

Let us set: $p_k^{h,h'} = \pi_k^{h'}/\pi_k^h$ for all k, h and h'. Then matrices $\boldsymbol{D}_k^{h,h'}$, vectors $\boldsymbol{b}_k^{h,h'}$ and scalars $p_k^{h,h'}$ constitute a whole parametric bond between populations, which is helpful (i) for defining some meaningful parsimonious models of stochastic transformations and (ii) for estimating $\boldsymbol{\theta}$.

Parsimonious models and estimation Several parsimonious models can be considered by combining classical assumptions within each mixture on both mixing proportions and Gaussian parameters (*intrapopulation* models), with meaningful constraints on the link parameters $\boldsymbol{D}_k^{h,h'}$, $\boldsymbol{b}_k^{h,h'}$ and $p_k^{h,h'}$ (*interpopulation* models).

Intrapopulation models. Inspired by standard Gaussian model-based clustering, one can envisage several classical parsimonious models of constraints on the Gaussian mixtures P^h: Their components may be homoscedastic ($\boldsymbol{\Sigma}_k^h = \boldsymbol{\Sigma}^h$) or heteroscedastic, their mixing proportions may be equal ($\pi_k^h = \pi^h$) or free.

Interpopulation models. In the most general case, $\boldsymbol{D}_k^{h,h'}$ matrices are positive-definite and diagonal, $\boldsymbol{b}_k^{h,h'}$ vectors are unconstrained and $p_k^{h,h'}$ scalars are positive. We can also consider component independent situations on $\boldsymbol{D}_k^{h,h'}$ ($\boldsymbol{D}_k^{h,h'} = \boldsymbol{D}^{h,h'}$), on $\boldsymbol{b}_k^{h,h'}$ ($\boldsymbol{b}_k^{h,h'} = \boldsymbol{b}^{h,h'}$) and/or on $p_k^{h,h'}$ ($p_k^{h,h'} = p^{h,h'}$).

Let us mention briefly that some combinations of the previous constraints are not allowed. For example, the mixing proportions cannot be assumed to be homogeneous within each mixture (π^h) and free through the populations ($p_k^{h,h'}$). All allowed combinations of intra and interpopulation models are available in [23]. They constitute a family of Gaussian mixture-based simultaneous clustering models.

Assuming that H samples $\mathcal{S}^1,\ldots,\mathcal{S}^H$ are drawn from the H populations $\Omega^1,\ldots,\Omega^H$, estimation of the model parameter by ML is carried out by the EM algorithm. We refer to [23] for more details.

4.3. Biological application of Gaussian transfer learning

In [32] three seabird subspecies ($H = 3$) of Cory's Shearwaters, differing over their geographical range, are described. *Borealis* (size $n^1 = 206$ individuals, 45% female) are living in the Atlantic Islands (Azores, Canaries, etc.), *diomedea* (size $n^2 = 38$ individuals, 58% female), in Mediterranean Islands (Balearics, Corsica, etc.), and *edwardsii* (size $n^3 = 92$ individuals, 52% female), in Cape Verde Islands. Only the two first subspecies have been considered in Application 2.1.3. Individuals are described in all species by the same five morphological variables ($d = 5$): Culmen (bill length), tarsus, wing and tail lengths, and culmen depth. We aim to cluster each subspecies.

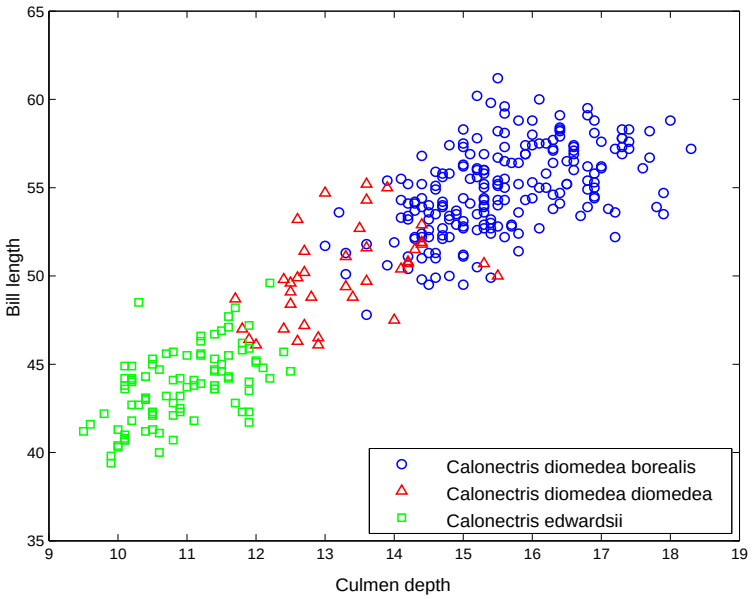


Figure 8. Three samples of Cory's Shearwaters described by identical features.

Figure 8 displays the birds in the plane of the culmen depth and the bill length. Samples seem clearly to arise from three different populations, so three standard independent Gaussian model-based clusterings should be considered. But as all of them arise from the same species *calonectris diomedea*, the researched partitions could be expected to have the same number of clusters with the same partition meaning in each sample. In addition, the three samples are described by the same five morphological features, thus the data set could be suitable for some simultaneous clustering process. As a consequence, it is quite reasonable that both simultaneous and independent clustering compete. The following paragraphs compare the results obtained from simultaneous and independent clustering.

Selecting the number of clusters Table 12 displays the best BIC criterion value among all models for the two clustering strategies. The overall best BIC value (4071.8) is ob-

Cluster Number	1	2	3	4
Simultaneous Clustering	4073.3	**4071.8**	4076.7	4082.4
Independent Clustering	4102.6	4139.8	4137.7	4159.6

Table 12. Best BIC values obtained in clustering the Cory's Shearwaters simultaneously and independently, with different number of clusters.

tained from simultaneous clustering for $K = 2$ groups. This value is widely better than the best BIC obtained from independent clustering ($BIC = 4102.6$). So BIC clearly prefers the simultaneous clustering method and rejects, here, the standard independent clustering method.

Determing the gender of birds Retaining the two cluster solution ($\hat{K} = 2$) we propose now to compare the partition estimated in each method with the gender partition of birds (males/females). The error rate associated to the best model of simultaneous clustering (BIC = 4071.8) is 10.71% whereas the best model of independent clustering (BIC = 4139.8) reaches 12.50%. Let us add that 10.71% is also the overall best error rate observed in both methods. So the best model according to BIC (provided by simultaneous clustering) is also the overall best global classifier. Moreover, Figure 9 and the confusion tables in Table 13 (a) and Table 13 (b), highlight the following point: By sexing differently few birds from the independent clustering, the simultaneous method improves not only the global error rate, but also the correlation between the estimated clusters and the bird genders.

(a) Independent clustering (BIC=4139.8).

		cluster 1	cluster 2
borealis	male	20	93
	female	88	5
diomedea	male	1	15
	female	18	4
edwardsii	male	7	37
	female	43	5

(b) Simultaneous clustering (BIC=4071.8).

		cluster 1	cluster 2
borealis	male	18	95
	female	89	4
diomedea	male	2	14
	female	18	4
edwardsii	male	5	39
	female	45	3

Table 13. Confusion tables obtained by comparing the inferred clusters within each subspecies ($K = 2$ groups) to the sex of the birds.

Interpreting the selected model The overall best model (BIC = 4071.8) specifies the following points. Firstly, the matrices $\mathbf{\Sigma}_k^h$ are homogeneous on k (every mixture is homoscedastic). So the covariance of any couple of biometrical variables should be the same among males and females (in each subspecies). This assertion makes sense and is realistic

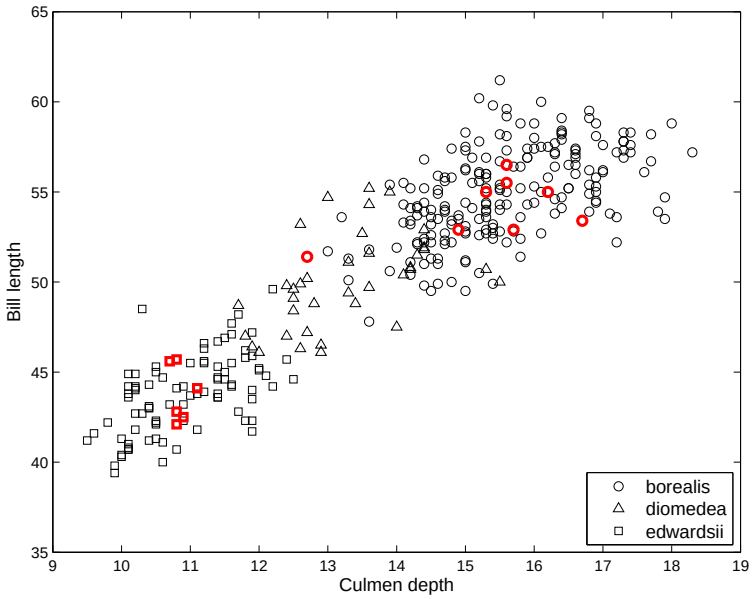


Figure 9. Similarities (in black) and differences (in red) between independent (BIC = 4139.8) and simultaneous clustering (BIC = 4071.8), in sexing each Shearwater sample.

for ornithologists. Secondly, $\pi_1^h = \pi_2^h$ whatever is $h \in \{1,2,3\}$. There are as many males than females in *borealis*, *diomedea* and *edwardsii* subspecies. This is realistic, given the gender partition of each sample. So, the two previous points are in accordance with the best model of independent clustering (BIC = 4139.8). In addition, the greater originality of the model provided by simultaneous clustering (BIC = 4071.8) lies in the following point: All transformation parameters $p_k^{h,h'}$, $\boldsymbol{b}_k^{h,h'}$ and $\boldsymbol{D}_k^{h,h'}$ are homogeneous on k. So, simultaneous clustering does not only provide the best model (according to BIC and according to the error rate), but this model allows also the following interpretation: There exists a mutual stochastic transformation of the males across the subspecies, there exists another transformation of the females, and these two transformations are identical.

4.4. Robust transfer learning and its estimation

As the Gaussian parameters are sensitive to extreme values, normal mixtures may be unsuitable for modeling the data when they are suspected to include noise or outliers. Alternatively, one can assume that the distribution of $\boldsymbol{X}^h$ conditionally to Y_k^h is no more Gaussian but corresponds to a d-dimensional Student's t distribution with degree of freedom $\nu_k^h \in \mathbb{R}_*^+$, location parameter $\boldsymbol{\mu}_k^h \in \mathbb{R}^d$ and (symmetric positive-definite) inner product matrix $\boldsymbol{\Sigma}_k^h \in \mathbb{R}^{d\times d}$ (see [27], chapter 7). We note in this case $\boldsymbol{\theta}^h = \{(\pi_k^h, \boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h, \nu_k^h) : k = 1,\ldots,K\}$ and again $\boldsymbol{\theta} = (\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^H)$.

[24] consider several parsimonious models of unlinked t-mixtures: Mixing proportions π_k^h are either free or homogeneous on k, as degrees of freedom ν_k^h and/or inner product matrices $\boldsymbol{\Sigma}_k^h$. Without any other constraint on the Student's parameters, the model at hand

involves an independent clustering procedure.

However, the mutual affine transformation of the conditional populations formalized by (13), can also be assumed in this new context of Student's t mixtures. This transformation is justified as in Section 4.2. when: (i) The samples share statistical units of same nature (ii) they are described by identical features and (iii) the expected partitions are about to be identically interpreted. Such an affine mutation of the conditional t-populations implicates that the degrees of freedom ν_k^h are homogeneous on h: $\nu_k^1 =, \ldots, = \nu_k^H = \nu_k$. Then interpopulation models are obtained by considering that $\boldsymbol{D}_k^{h,h'}$ matrices, $\boldsymbol{b}_k^{h,h'}$ vectors and/or $p_k^{h,h'}$ scalars are either free or homogeneous on k.

Combining the previous constraints on (i) the intrapopulation parameters π_k^h, ν_k^h, $\boldsymbol{\Sigma}_k^h$ and (ii) the transformation parameters $p_k^{h,h'}$, $\boldsymbol{b}_k^{h,h'}$, $\boldsymbol{D}_k^{h,h'}$, leads to a family of t-mixture-based simultaneous clustering models. As in the Gaussian case some of the proposed constraints cannot be matched and one has to refer to [24] in order to find all allowed combinations of intra and interpopulation models. The estimation of $\boldsymbol{\theta}$ by ML requires a Generalized EM algorithm (see [11]) and details are given again in [24].

Any likelihood-based criterion can help to (i) select a model of independent clustering, (ii) select a model of simultaneous clustering and (iii) determine the best clustering method among the two previous ones. The BIC criterion was used for that purpose in Section 4.2. But BIC reports essentially the adequacy of the model and sometimes at the expense of the interpretability of the associated partition. In order to avoid some spurious components (see [27]) due to the noise within the data, BIC can be replaced by the ICL criterion (see [4]) defined by

$$\mathrm{ICL} = \mathrm{BIC} - 2\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{\{i:\hat{y}_i^h=k\}} \ln t_k(\boldsymbol{x}_i^h;\hat{\boldsymbol{\theta}}^h),$$

where $\hat{y}_i^h$ denotes the MAP estimate of y_i^h and $t_k(\boldsymbol{x}_i^h;\hat{\boldsymbol{\theta}}^h)$ is the conditional probability of membership of $\boldsymbol{x}_i^h$ to the cluster k (it is a straightforward adaptation of (2)). ICL penalizes the models with strong overlap components, so it provides partitions where the clusters are well-separated and, consequently, easier to interpret. The model with the smallest ICL value has to be retained.

4.5. Economic application of robust transfer learning

In [12] Du Jardin and Séverin display several samples of firms differing over the year and described by the same econometric variables. On two samples from these financial data, suspected by the authors to contain outliers, we propose to compare the simultaneous and the independent clustering strategies both based on t-mixtures. The first sample from 2002 consists of 428 firms (212 bankrupt ones) and the second sample from 2003, of 461 companies (220 bankrupt ones). Both samples are described by four financial ratios: EBITDA/Total Assets, Value Added/Total Sales, Quick Ratio, Accounts Payable/Total Sales. Figure 10 represents the two datasets in the canonical plane: [EBITDA/Total Assets, Quick ratio]. Table 14 displays the best ICL criterion value among all models for simultaneous clustering strategy. We notice that ICL retains a three clusters ($K = 3$) solution.

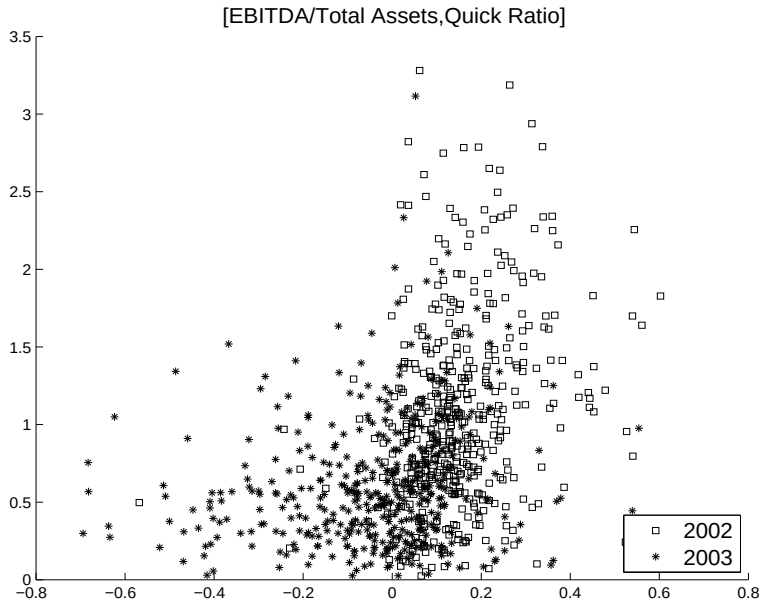


Figure 10. Two samples of companies differing over the year.

K	1	2	3	4	5
Simultaneous	−1169.7	−1191.3	**−1202.0**	−1183.4	−1131.3
Independent	−1154.6	**−1163.6**	−1072.1	−1127.7	−1098.3

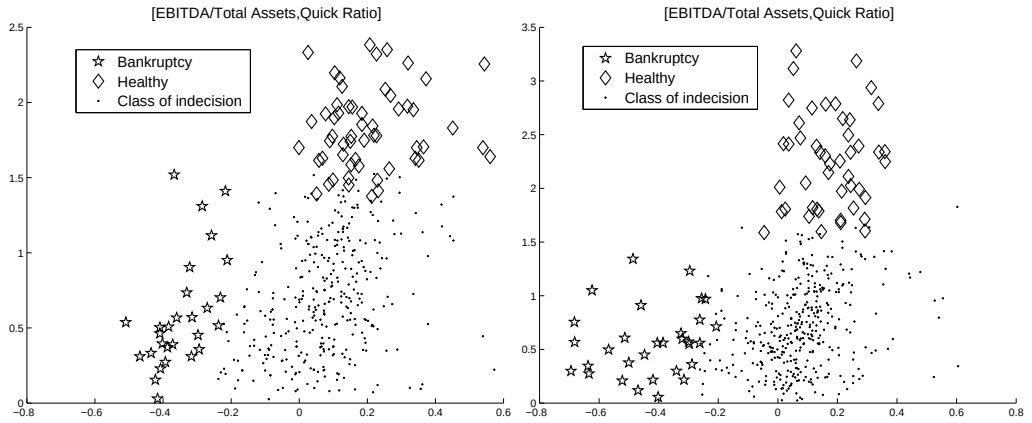
Table 14. Best ICL values, over all models, obtained in simultaneous and independent clustering with different number of clusters.

Table 15 gives the associated confusion table of this obtained partition in comparison to the bankruptcy and healthy specifications. We see that estimated Clusters 1 and 2 are highly correlated respectively to failed and no-failed companies, whereas Cluster 3 is clearly a group where failed and no-failed companies are indistinguishable.

	Cluster 1	Cluster 2	Cluster 3
Healthy	3	94	360
Bankruptcy	56	10	366

Table 15. Confusion table associated to the partition provided by the best simultaneous clustering model retained by ICL.

This typology indicates that it is easy to identify very well healthy and non-healthy companies (see Figure 11) for a small number of cases (Clusters 1 and 2 have respectively mixing proportions equal to 0.07 and 0.13) whereas it is expected to be a very hard task for most of them (Cluster 3 has a mixing proportion of 0.80).



(a) *year* 2002 (b) *year* 2003

Figure 11. Estimated partition of companies (Healthy, Bankruptcy, Indecision) for the two consecutive years (2002, 2003), obtained by a simultaneous t-mixture model-based clustering methodology.

In addition, by using t-parameters of each cluster, it is obviously possible to draw a synthetic description of each of them (classical analysis in model-based clustering so not reported here) but we focus on the specificity of simultaneous clustering which provides an information about the group evolution over the years. The retained model (ICL $= -1202.2$) states that every mutation parameter p_k^h, $\boldsymbol{b}_k^{h,h'}$, $\boldsymbol{D}_k^{h,h'}$ is homogeneous on k, that the degrees of freedom ν_k and the inner product matrices $\boldsymbol{\Sigma}_k^h$ do not depend on k either, and that the reference mixing proportions π_k^1 are free. Then, according to this model: (i) The mixing proportion of each cluster is invariant between 2002 and 2003 and (ii) other cluster features uniformly evolved over the years. More precisely, the associated estimated transition parameters are given by:

$$\begin{aligned} \hat{\boldsymbol{D}}^{1,2} &= \mathrm{diag}(1.12, 0.95, 1.20, 0.93) \\ \hat{\boldsymbol{b}}^{1,2} &= 10^{-3}.(-18.2, 2, -102, -1)', \end{aligned}$$

thus clusters from 2002 and 2003 appear to vary only through the two variables EBITDA/Total Assets and Quick Ratio. This result is meaningful: These two variables report respectively the liquidity and the performance of the firms, which are known to be the main features able to predict bankruptcy. Indeed the change of the financial structure is a consequence of the evolution of these two features.

For comparison, Table 14 displays also the best ICL criterion value among all models for independent clustering. We notice now that $K = 2$ clusters are retained and the associated confusion table (Table 16) indicates that estimated clusters bring poor information about the company health in comparison to the three components solution given by simultaneous clustering. In addition, independent clustering does not allow easy interpretation of the groups evolution over the years. Finally, it is worth noting that ICL prefers the simultaneous solution.

	Cluster 1	Cluster 2
Healthy	228	229
Bankruptcy	289	143

Table 16. Confusion table associated to the partition provided by the best independent clustering model retained by ICL.

5. Conclusion

In this chapter, parametric transfer learning and parametric unsupervised transfer learning have been addressed for classification, regression or clustering problems when several samples are involved in the analysis. In each situation, the proposed transfer function belongs to a parametric family what allows to obtain an easy understanding of the functional link between populations. It corresponds also to a very flexible approach since (i) many constraints on the parametric link can be proposed from the most to the less restrictive ones and (ii) model selection allows to retain automatically the most appropriate constraint.

It is worth noting also that proposed strategies are all easy to implement by practitioners since they are finally quite close to standard methods. Shortly speaking, they correspond to particular simple, but meaningful, constraints on classical models.

Face to the huge amount of available data today and especially in a near future, transfer learning may also work as a powerful and generic data reduction tool. Indeed, it allows to identify links between populations and, as a consequence, it is a way to obtain equivalence classes for them.

Finally, some challenges have still to be addressed by this emerging field. For instance, it would be useful to extend the proposed methods to high dimensional data sets or to other classical techniques. In addition, it would interesting to weaken some assumptions as the exact variable concordance between variables, property which is currently required.

References

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

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

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

[4] C. Biernacki and G. Celeux. Assessing a mixture for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:719–725, 2000.

[5] C. Bouveyron and J. Jacques. Adaptive linear models for regression: improving prediction when population has changed. *Pattern Recognition Letters*, 31(14):2237–2247, 2010.

[6] C. Bouveyron and J. Jacques. Adaptive mixtures of regressions: Improving predictive inference when population has changed. *Pub. IRMA Lille*, 70(VIII), 2010.

[7] V. Bretagnolle. personnal communication. 2006.

[8] G. Celeux and G. Govaert. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8:157–176, 1991.

[9] G. Celeux and G. Govaert. Parsimonious gaussian models in cluster analysis. *Pattern Recognition*, 28:781–793, 1995.

[10] B. De Meyer, B. Roynette, P. Vallois, and M. Yor. On independent times and positions for Brownian motions. *Revista Matemática Iberoamericana*, 18(3):541–586, 2002.

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

[12] P. Du Jardin and E. Séverin. Dynamic analysis of the business failure process: a study of bankruptcy trajectories. In *Portuguese Finance Network*, Ponte Delgada, Portugual, 2010.

[13] B. S. Everitt. *An introduction to latent variable models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1984.

[14] C. Fraley and A. E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, 41:578–588, 1998.

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

[16] D. Hand. *Discriminant and Classification*. Wiley, New York, 1996.

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2009.

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

[19] D.W. Hosmer and S. Lemeshow. *Applied logistic regression*. Wiley Series in Probability and Statistics. Wiley, 2000.

[20] M. Hurn, A. Justel, and C.P. Robert. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79, 2003.

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

[22] A. Lourme and C. Biernacki. Gaussian model-based classification when training and test population differ: Estimating jointly related parameters. In *First joint meeting of the Société Francophone de Classification and of the Classification and Data Analysis Group of SIS*, 2008.

[23] A. Lourme and C. Biernacki. Simultaneous Gaussian Model-Based Clustering for Samples of Multiple Origins. *preprint 70, VII, IRMA, Lille.*, 2010.

[24] A. Lourme and C. Biernacki. Simultaneous *t*-model-based clustering for data differing over time period: Application for understanding companies financial health. *Case Studies in Business, Industry and Government Statistics (CSBIGS)*, 4(2), 2011.

[25] G.J. McLachlan. *Discriminant analysis and statistical pattern recognition*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1992. , A Wiley-Interscience Publication.

[26] G.J. Mclachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2004.

[27] G.J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

[28] S.J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[29] B. Ripley. *Pattern Recognition and Neural Network*. Cambridge University Press, Cambridge, 1995.

[30] K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92:894–902, 1997.

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

[32] J.C. Thibault, V. Bretagnolle, and C. Rabouam. Cory's shearwater calonectris diomedea. *Birds of Western Paleartic Update*, 1:75–98, 1997.