Simultaneous Gaussian Model-Based Clustering for Samples of Multiple Origins

Alexandre Lourme, Christophe Biernacki

Computational Statistics manuscript No.
(will be inserted by the editor)

Simultaneous Gaussian Model-Based Clustering for Samples of Multiple Origins

A. Lourme · **C. Biernacki**

Received: date / Accepted: date

Abstract Gaussian mixture model-based clustering is now a standard tool to estimate some hypothetical underlying partition of a single dataset. In this paper, we aim to cluster several different datasets at the same time in a context where underlying populations, even though different, are not completely unrelated: All individuals are described by the same features and partitions of identical meaning are expected. Justifying from some natural arguments a stochastic linear link between the components of the mixtures associated to each dataset, we propose some parsimonious and meaningful models for a so-called simultaneous clustering method. Maximum likelihood mixture parameters, subject to the linear link constraint, can be easily estimated by a Generalized Expectation Maximization (GEM) algorithm that we describe. Some promising results are obtained in a biological context where simultaneous clustering outperforms independent clustering for partitioning three different subspecies of birds. Further results on ornithological data show that the proposed strategy is robust to the relaxation of the exact descriptor concordance which is one of its main assumptions.

Keywords Stochastic linear link · Gaussian mixture · Model-based clustering · EM algorithm · Model selection · Biological features

1 Introduction

Clustering aims to separate a sample into classes in order to reveal some hidden but meaningful structure in data. In a probabilistic context it is standard practice to

A. Lourme
Université de Pau & IUT Génie Biologique, Mont de Marsan, France
Tel.: +33 5 58 51 37 02
Fax: +33 5 58 51 37 37
E-mail: Alexandre.Lourme@univ-pau.fr

C. Biernacki
Université Lille 1 & CNRS & INRIA, Villeneuve d'Ascq, France
E-mail: Christophe.Biernacki@math.univ-lille1.fr

suppose that the data arise from a mixture of parametric distributions and to draw a partition by assigning each data point to the prevailing component (see McLachlan and Peel, 2000, for a review). In particular, in the multivariate continuous situation, Gaussian mixture model-based clustering has found successful applications in diverse fields: Genetics (Schork and Thiel, 1996), Medicine (McLachlan and Peel, 2000), Magnetic Resonance Imaging (Banfield and Raftery, 1993), Astronomy (Celeux and Govaert, 1995), ... Consequently, nowadays, involving such models for clustering a given dataset could be considered as familiar to every statistician as to more and more practitioners.

However, in many situations one needs to cluster several datasets which arise from *related* but *different* populations. More precisely, such populations are related since they correspond to similar individuals described also by the same features and, thus, from which partitions of identical meaning are expected to be discovered. But, such populations are different since there exists a shift between them, typically involving time, place, culture, ethnic group, *etc.* Here are some examples of such situations (some other ones can be found in Lourme, 2011):

- *Finance*: In Du Jardin and Séverin (2010), several samples of firms differing over the time period are described by a common set of econometric variables. Within each sample, clusters are expected to reveal healthy and bankruptcy companies but the well-known fast chronological evolution of the global firm features makes the populations very different.
- *Geology*: Eruptions of the Old Faithful geyser are now famous and are usually described by waiting time between eruptions and the duration of the eruption. However, Lourme (2011) shows that the distribution of the outbursts evolves over time (considering a ten years shift) but that the traditional structure in two groups (short/long) of Old Faithful eruptions remains.
- *Biology*: Thibault et al (1997) consider three samples of seabirds corresponding to distinct Shearwater subspecies, differing over their geographical range, but described by the same five morphological features (tarsus, bill length, *etc.*). Despite these both place and subspecies shifts, each sample is expected to be partitioned according to the birds gender.

Such situations involve commonly two kinds of standard clustering processes. The samples are clustered traditionally either as if all units arose from the same distribution, or on the contrary as if the samples came from distinct and unrelated populations. But a third situation should be considered: As the datasets share statistical units of same nature and as they are described by features of same meaning, there may exist some link between the samples.

In the Gaussian mixture model-based clustering context, we propose a probabilistic model which enables us to simultaneously classify all individuals instead of applying several independent Gaussian clustering methods, without ever considering that all units have the same origin. Assuming a linear stochastic link between the samples, what can be justified from some simple but realistic assumptions, will be the basis of this work. This link allows us to estimate, by maximum likelihood (ML), all Gaussian mixture parameters at the same time which is a novelty for independent clustering, and consequently allows us to cluster the diverse datasets simultaneously.

Any likelihood-based model choice criterion such as *BIC* (Schwarz, 1978) enables us then to compare the three clustering methods: The simultaneous clustering method which assumes a stochastic link between the populations, the independent clustering method which considers that populations are unrelated and the clustering method involved by considering that all data arise from a common origin.

In fact, generalizing a one-sample method to *several* samples is quite common in statistical literature. Flury (1983), for example, proposes the use a particular Principal Component Analysis based on common principal components for representing several samples in a mutual lower-dimensional space when their covariance matrices share a common form and orientation. Another example is given by Gower (1975) who generalizes to K samples ($K \geq 3$) the classical Procrustes analysis and estimates a geometrical link, established between two samples.

Note also that terminology "simultaneous clustering" that can be found in literature is very different from this one we present in this work since it usually refers to a *single* sample. For example hierarchical mixture models (Vermunt and Magidson, 2005, pp.176–183) aim to cluster simultaneously lower- and higher-level units (nested levels) of a unique three-way dataset with an extension of standard mixture model-based clustering. In addition, model-based co-clustering aims to cluster simultaneously the sets of rows and columns of a data matrix of a unique dataset (see for instance Govaert and Nadif, 2008).

In Section 2, starting from the standard solution of some independent Gaussian mixture model-based clustering methods, we present the principle of simultaneous clustering. Some parsimonious and meaningful models on the established stochastic link are then proposed in Section 3. Section 4 gives the formulae required by the ML inference of the parameter, and also proposes, for some models, a simplified alternative estimation combining a cheap standardization step and a standard ML for Gaussian mixture step. Some experiments on seabird samples show encouraging results for our new method. They will be presented in Section 5. Finally in Section 6 we plan extensions of this work.

2 From independent to simultaneous Gaussian clustering

We aim to separate H samples into K groups. Describing standard Gaussian model-based clustering (Subsection 2.1) in this apparently more complex context (H samples instead of one), will be later convenient for introducing simultaneous Gaussian model-based clustering (Subsection 2.2). Each sample $\mathbf{x}^h$ ($h \in \{1,\ldots,H\}$) is composed of n^h individuals $\mathbf{x}_i^h$ ($i = 1,\ldots,n^h$) of $\mathbb{R}^d$, and arises from a population h. In addition, all populations are described by the same d continuous variables. Let us remind also here that, in each sample the same number of clusters has to be discovered, and that the obtained partition has the same meaning for each sample.

2.1 Standard solution: Several independent Gaussian clusterings

Standard Gaussian model-based clustering assumes that individuals $\mathbf{x}_i^h$ of each sample $\mathbf{x}^h$ are independently drawn from the random vector $\boldsymbol{X}^h$ distributed as a mixture

P^h of K non-degenerate Gaussian components C_k^h ($k = 1,\dots,K$), with probability density function:

$$f(\boldsymbol{x};\boldsymbol{\psi}^h) = \sum_{k=1}^{K} \pi_k^h \boldsymbol{\Phi}_d(\boldsymbol{x};\boldsymbol{\mu}_k^h,\boldsymbol{\Sigma}_k^h),\ \boldsymbol{x} \in \mathbb{R}^d.$$

Coefficients π_k^h ($k = 1,\dots,K$) are the mixing proportions (for all k, $\pi_k^h > 0$ and $\sum_{k=1}^K \pi_k^h = 1$), $\boldsymbol{\mu}_k^h$ and $\boldsymbol{\Sigma}_k^h$ correspond respectively to the center and the covariance matrix of C_k^h component, and $\boldsymbol{\Phi}_d(\boldsymbol{x};\boldsymbol{\mu}_k^h,\boldsymbol{\Sigma}_k^h)$ denotes its probability density function. The whole parameter of P^h mixture is $\boldsymbol{\psi}^h = (\boldsymbol{\psi}_k^h)_{k=1,\dots,K}$ where $\boldsymbol{\psi}_k^h = (\pi_k^h,\boldsymbol{\mu}_k^h,\boldsymbol{\Sigma}_k^h)$ is a vector combining in a properly manner a scalar, a vector and a matrix.

The component that may have generated an individual $\mathbf{x}_i^h$ constitutes a missing data. We represent it by a binary vector $\mathbf{z}_i^h \in \{0,1\}^K$ of which the k-th component $z_{i,k}^h$ equals 1 if and only if $\mathbf{x}_i^h$ arises from C_k^h. The vector $\mathbf{z}_i^h$ is assumed to arise from the K-variate multinomial distribution of order 1 and of parameter $(\pi_1^h,\dots,\pi_K^h)$.

The complete data model assumes that couples $(\mathbf{x}_i^h,\mathbf{z}_i^h)_{i=1,\dots,n^h}$ are realizations of independent random vectors identically distributed to $\left(\boldsymbol{X}^h,\boldsymbol{Z}^h\right)$ in $\mathbb{R}^d \times \{0,1\}^K$ where $\boldsymbol{Z}^h$ denotes a random vector of which the k-th component Z_k^h equals 1 (and the others 0) with probability π_k^h, and $\left(\boldsymbol{X}^h | Z_k^h = 1\right) \sim \boldsymbol{\Phi}_d(\,.\,;\boldsymbol{\mu}_k^h,\boldsymbol{\Sigma}_k^h)$. We note also $\boldsymbol{z}^h = \{\boldsymbol{z}_1^h,\dots,\boldsymbol{z}_{n^h}^h\}$.

Estimating $\boldsymbol{\psi} = (\boldsymbol{\psi}^h)_{h=1,\dots,H}$, by maximizing its log-likelihood

$$\ell(\boldsymbol{\psi};\mathbf{x}) = \sum_{h=1}^{H}\sum_{i=1}^{n^h} \log\left[f(\mathbf{x}_i^h;\boldsymbol{\psi}^h)\right] = \sum_{h=1}^{H} \ell^h(\boldsymbol{\psi}^h;\mathbf{x}^h),$$

computed on the observed data $\mathbf{x} = \bigcup_{h=1}^H \mathbf{x}^h$, leads to maximizing independently each log-likelihood $\ell^h(\boldsymbol{\psi}^h;\mathbf{x}^h)$ of the parameter $\boldsymbol{\psi}^h$ computed on $\mathbf{x}^h$ sample. Invoking an EM algorithm to perform the maximization is a classical method. One can see McLachlan and Peel (2000) for a review.

Then the observed data $\mathbf{x}_i^h$ is allocated by the Maximum A Posteriori principle (MAP) to the group corresponding to the highest estimated posterior probability of membership computed at the ML estimate $\hat{\boldsymbol{\psi}}$:

$$t_{i,k}^h(\hat{\boldsymbol{\psi}}) = E(Z_k^h | \boldsymbol{X}^h = \mathbf{x}_i^h;\hat{\boldsymbol{\psi}}). \tag{1}$$

In a parametric model-based clustering context the *BIC* criterion (see Schwarz, 1978; Lebarbier and Mary-Huard, 2006) is commonly used for selecting a model and/or the number of clusters (see Roeder and Wasserman, 1997; Fraley and Raftery, 1998). It is defined here by:

$$BIC = -\ell(\hat{\boldsymbol{\psi}};\mathbf{x}) + \frac{\nu}{2}\log(n), \tag{2}$$

where $\ell(\hat{\boldsymbol{\psi}};\mathbf{x})$ denotes the maximized log-likelihood of the parameter $\boldsymbol{\psi}$ computed on the observed data $\mathbf{x}$, ν the dimension of $\boldsymbol{\psi}$, and n the size of the data ($n = \sum_{h=1}^H n^h$).

At this step, we have to remember that identical meaning partitions are expected in all samples and the question is now to match classes of same meaning across different samples. Three different strategies can be applied:

- Either, the practitioner has to perform himself the matching from the main features of the estimated clusters (centers, *etc.*);
- Or, in a more systematic way, we can try to minimize a global "distance" (as a Kullback-Leibler divergence) between the matched components;
- Or, if an external partition exists, the classes can be labelled so as to obtain a global error rate as low as possible but this is likely the less frequent solution since an external partition is rarely available[1].

The simultaneous clustering method that we present now, aims both to improve the partition estimation and to automatically give the same numbering to the clusters with identical meaning.

2.2 Proposed solution: Using a linear stochastic link between populations

From the beginning the groups that have to be discovered consist in a same meaning partition of each sample and samples are described by the same features. In that context, since involved populations are so related, we establish a distributional relationship between the identically labelled components C_k^h ($h = 1,\ldots,H$). Formalizing thus some link between the conditional populations constitutes the key idea of the so-called simultaneous clustering method, and this idea will be specified thanks to three additional hypotheses $\mathscr{H}_1$, $\mathscr{H}_2$, $\mathscr{H}_3$ described below.

For all $(h,h') \in \{1,\ldots,H\}^2$ and all $k \in \{1,\ldots,K\}$, a map $\xi_k^{h,h'} : \mathbb{R}^d \to \mathbb{R}^d$ is assumed to exist, so that:

$$(\boldsymbol{X}^{h'} | Z_k^{h'} = 1) \stackrel{\mathscr{D}}{=} \xi_k^{h,h'} \left(\boldsymbol{X}^{h} | Z_k^{h} = 1 \right), \tag{3}$$

where ' $\stackrel{\mathscr{D}}{=}$ ' refers to the equality of two random vectors in distribution. (Note that $\xi_k^{h,h'}$ map is not symmetric with respect to (h,h').) This model implicates that individuals from some Gaussian component C_k^h are stochastically transformed (via $\xi_k^{h,h'}$) into individuals of $C_k^{h'}$.

In addition, as samples are described by the same features, it is natural, in many practical situations, to expect from a variable in some population to depend mainly on the same feature, in another population. So we assume that the j-th ($j \in \{1,\ldots,d\}$) map component $\left(\xi_k^{h,h'}\right)^{(j)}$ of $\xi_k^{h,h'}$ map depends only on the j-th map component $\boldsymbol{x}^{(j)}$ of $\boldsymbol{x}$, situation that is expressed by the following hypothesis:

$$\mathscr{H}_1 : \forall j \in \{1,\ldots,d\}, \forall (\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^d \times \mathbb{R}^d, \boldsymbol{x}^{(j)} = \boldsymbol{y}^{(j)} \Rightarrow \left(\xi_k^{h,h'}\right)^{(j)} (\boldsymbol{x}) = \left(\xi_k^{h,h'}\right)^{(j)} (\boldsymbol{y}).$$

[1] However, we will use this strategy in our biological example in Section 5 in order to put independent clustering in this "ideal" context.

In other words, $\left(\xi_k^{h,h'}\right)^{(j)}$ corresponds to a map from $\mathbb{R}$ into $\mathbb{R}$ that transforms, in distribution, the conditional Gaussian variable $(\boldsymbol{X}^h|Z_k^h=1)^{(j)}$ into the corresponding conditional Gaussian variable $(\boldsymbol{X}^{h'}|Z_k^{h'}=1)^{(j)}$. $\mathscr{H}_1$ can be seen as a "first order" approximation of the link between variables. We will discuss possibility to relax it (and other hypotheses) in Section 6.

Assuming moreover that $\left(\xi_k^{h,h'}\right)^{(j)}$ is continuously differentiable–this assumption about all superscripts j is noted $\mathscr{H}_2$–, then the only possible transformation is an affine map. Indeed, according to Appendix A in Biernacki et al (2002), for two given non-degenerate univariate normal distributions, there exist only two continuously differentiable maps from $\mathbb{R}$ into $\mathbb{R}$ that transforms, in distribution, the first one into the second one, and they are both affine. The same hypotheses have been also used by Biernacki et al (2002) in a Gaussian supervised context.

As a consequence, for all $(h,h') \in \{1,\ldots,H\}^2$ and all $k \in \{1,\ldots,K\}$, there exists $\boldsymbol{D}_k^{h,h'} \in \mathbb{R}^{d\times d}$ diagonal and $\boldsymbol{b}_k^{h,h'} \in \mathbb{R}^d$ so that:

$$(\boldsymbol{X}^{h'}|Z_k^{h'}=1) \overset{\mathscr{D}}{=} \boldsymbol{D}_k^{h,h'}(\boldsymbol{X}^h|Z_k^h=1)+\boldsymbol{b}_k^{h,h'}. \tag{4}$$

Relation (3) constitutes the keystone of the simultaneous Gaussian model-based clustering framework, and (4) is its affine form involved from the two previous hypotheses $\mathscr{H}_1$ and $\mathscr{H}_2$.

Since components C_k^h are non-degenerate, $\boldsymbol{D}_k^{h,h'}$ matrices are non singular. In addition we assume henceforward that $\boldsymbol{D}_k^{h,h'}$ matrices are positive–assumption noted $\mathscr{H}_3$–for two reasons. First, it makes the model identifiable. Second, this assumption involves that for any couple of conditional variables, the sign of their correlation keeps unchanged through the populations what seems to be realistic in many practical contexts (for instance in our biological example below, Section 5). Notice that hypothesis may be weakened as we remark it at the end of Subsection 4.4.

Thus, any couple of identically labelled component parameters, $\boldsymbol{\psi}_k^h$ and $\boldsymbol{\psi}_k^{h'}$, has now to satisfy the following property: There exists 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:

$$\boldsymbol{\Sigma}_k^{h'} = \boldsymbol{D}_k^{h,h'}\,\boldsymbol{\Sigma}_k^h\,\boldsymbol{D}_k^{h,h'} \text{ and } \boldsymbol{\mu}_k^{h'} = \boldsymbol{D}_k^{h,h'}\boldsymbol{\mu}_k^h+\boldsymbol{b}_k^{h,h'}. \tag{5}$$

(Let us note then that $\boldsymbol{D}_k^{h,h'} = \left(\boldsymbol{D}_k^{h',h}\right)^{-1}$ and $\boldsymbol{b}_k^{h,h'} = -\boldsymbol{D}_k^{h,h'}\boldsymbol{b}_k^{h',h}$.)

Property (5) characterizes henceforward the whole parameter space $\boldsymbol{\Psi}$ of $\boldsymbol{\psi}$ and the so-called simultaneous clustering method is based on $\boldsymbol{\psi}$ parameter inference in that so constrained parameter space.

2.3 A statistical interpretation of the linear stochastic link

Each covariance matrix can be decomposed into:

$$\boldsymbol{\Sigma}_k^h = \boldsymbol{T}_k^h\boldsymbol{R}_k^h\boldsymbol{T}_k^h, \tag{6}$$

where $\boldsymbol{T}_k^h$ is the diagonal matrix of conditional standard deviations in C_k^h component– for all $(i,j) \in \{1,\ldots,d\}^2$: $\boldsymbol{T}_k^h(i,j) = \sqrt{\boldsymbol{\Sigma}_k^h(i,j)}$ if $i=j$ and 0 otherwise–and $\boldsymbol{R}_k^h = \left(\boldsymbol{T}_k^h\right)^{-1} \boldsymbol{\Sigma}_k^h \left(\boldsymbol{T}_k^h\right)^{-1}$ is the conditional correlation matrix of the class. As each decomposition (6) is unique, Relation (5) involves for every $(h,h') \in \{1,\ldots,H\}^2$ and every $k \in \{1,\ldots,K\}$ both $\boldsymbol{T}_k^{h'} = \boldsymbol{D}_k^{h,h'} \boldsymbol{T}_k^h$ and $\boldsymbol{R}_k^{h'} = \boldsymbol{R}_k^h$. The previous model (4) is equivalent therefore to postulating that conditional correlations are equal through the populations.

This interpretation of the affine link between the conditional populations (4) allows the model to keep all its sense when simultaneous clustering is envisaged in a relaxed context–as in Subsection 5.3–where the samples to be classified are described by different descriptor sets.

3 Parsimonious models

This section displays some parsimonious models established by combining classical assumptions within each mixture on both mixing proportions and Gaussian parameters (intrapopulation models), with meaningful constraints on the parametric link (5) between conditional populations (interpopulation models).

3.1 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 ($\forall k : \boldsymbol{\Sigma}_k^h = \boldsymbol{\Sigma}^h$) or heteroscedastic, their mixing proportions may be equal ($\forall k : \pi_k^h = \pi^h$) or free (see McLachlan and Peel, 2000, Chapter 3). These models will be called *intrapopulation models*.

Although they are not considered here, some other intrapopulation models can be assumed. Celeux and Govaert (1995) for example propose some parsimonious models of Gaussian mixtures based on an eigenvalue decomposition of the covariance matrices which can be envisaged as an immediate extension of our intrapopulation models.

3.2 Interpopulation models

In the most general case, $\boldsymbol{D}_k^{h,h'}$ matrices are definite-positive and diagonal, and $\boldsymbol{b}_k^{h,h'}$ are unconstrained. We can also consider component independent situations on $\boldsymbol{D}_k^{h,h'}$ ($\forall k : \boldsymbol{D}_k^{h,h'} = \boldsymbol{D}^{h,h'}$) and/or on $\boldsymbol{b}_k^{h,h'}$ ($\forall k : \boldsymbol{b}_k^{h,h'} = \boldsymbol{b}^{h,h'}$). Other constraints on $\boldsymbol{D}_k^{h,h'}$ and $\boldsymbol{b}_k^{h,h'}$ can be easily proposed but are not considered in this paper (see Lourme and Biernacki, 2010, for other examples). We can also suppose the mixing proportion vectors $(\pi_1^h,\ldots,\pi_K^h)$ $(h=1,\ldots,H)$ to be equal or free depending on whether the ratios $\alpha_k^{h,h'} = \pi_k^{h'}/\pi_k^h$ $(k=1,\ldots,K, h=1,\ldots,H, h'=1,\ldots,H)$ are equal to one ($\forall k : \alpha_k^{h,h'} =$

1) or free. These models will be called *interpopulation models* and they have to be combined with some intrapopulation model.

Remark There we can see that some of the previous constraints cannot be set simultaneously on the transformation matrices and on the translation vectors. When $\boldsymbol{b}_k^{h,h'}$ vectors do not depend on k for example, then neither do $\boldsymbol{D}_k^{h,h'}$ matrices. Indeed, from (5), we obtain $\boldsymbol{\mu}_k^h = \left(\boldsymbol{D}_k^{h,h'}\right)^{-1} \boldsymbol{\mu}_k^{h'} - \left(\boldsymbol{D}_k^{h,h'}\right)^{-1} \boldsymbol{b}_k^{h,h'}$, and consequently $\boldsymbol{b}_k^{h',h} = -\left(\boldsymbol{D}_k^{h,h'}\right)^{-1} \boldsymbol{b}_k^{h,h'}$ depends on k once $\boldsymbol{D}_k^{h,h'}$ or $\boldsymbol{b}_k^{h,h'}$ does.

3.3 Combining intra and interpopulation models

The most general model of simultaneous clustering is noted $\left(\alpha_k^{h,h'}, \boldsymbol{D}_k^{h,h'}, \boldsymbol{b}_k^{h,h'}; \pi_k^h, \boldsymbol{\Sigma}_k^h\right)$. It assumes that mixing proportion vectors may be different between populations (so π_k^h coefficients are free on h), $\boldsymbol{D}_k^{h,h'}$ matrices are just diagonal definite-positive (unconstrained case), $\boldsymbol{b}_k^{h,h'}$ vectors are unconstrained, and that each mixture has heteroscedastic components with free mixing proportions (thus π_k^h coefficients are also free on k). The model $\left(1, \boldsymbol{D}^{h,h'}, \boldsymbol{b}^{h,h'}; \pi^h, \boldsymbol{\Sigma}^h\right)$, for another example, assumes all mixing proportions to be equal to $1/K$, $\boldsymbol{D}_k^{h,h'}$ matrices, $\boldsymbol{b}_k^{h,h'}$ vectors to be component independent and each mixture to have homoscedastic components.

As a model of simultaneous clustering consists of a combination of some intra and interpopulation models, one will have to pay attention to non-allowed combinations. It is impossible for example, to assume both that mixing proportion vectors are free through the diverse populations, and that each of them has equal components. Then a model $\left(\alpha_k^{h,h'}, \ldots; \pi^h, .\right)$ is not allowed. In the same way, we cannot suppose–it is straightforward from the relationship between $\boldsymbol{\Sigma}_k^h$ and $\boldsymbol{\Sigma}_k^{h'}$ in (5)–both $\boldsymbol{D}_k^{h,h'}$ transformation matrices to be free, and, at the same time, each mixture to have homoscedastic components. A model $\left(., \boldsymbol{D}_k^{h,h'}, .; ., \boldsymbol{\Sigma}^h\right)$ is then prohibited.

Table 1 displays all allowed combinations of intra and interpopulation models, leading to 15 models.

3.4 Requirements about identifiability

For a given permutation σ in $\mathscr{S}_H$ (symmetric group on $\{1,\ldots,H\}$), and another one τ in $\mathscr{S}_K$, ψ_τ^σ will denote the parameter ψ, in which population labels have been permuted as σ, and component labels as τ, that is:

$$\forall k \in \{1,\ldots,K\}, \forall h \in \{1,\ldots,H\} : (\psi_\tau^\sigma)_k^h = \psi_{\tau(k)}^{\sigma(h)}.$$

Identifiability of a model is defined up to a permutation of population labels, and up to the same component label permutation within each population, that is, formally, a

Table 1 *Allowed intra/interpopulation model combinations and identifiable models. We note '.' some non-allowed combination of intra and interpopulation models, '∘' some allowed but non-identifiable model, and '•' some allowed and identifiable model.*

Interpopulation models			*Intrapopulation models* π^h $\boldsymbol{\Sigma}^h$	π^h $\boldsymbol{\Sigma}_k^h$	π_k^h $\boldsymbol{\Sigma}^h$	π_k^h $\boldsymbol{\Sigma}_k^h$
$1\ (\alpha_k^{h,h'})$	$\boldsymbol{D}^{h,h'}$	$\boldsymbol{b}^{h,h'}$	• (.)	• (.)	•(•)	•(•)
		$\boldsymbol{b}_k^{h,h'}$	∘ (.)	• (.)	•(•)	•(•)
	$\boldsymbol{D}_k^{h,h'}$	$\boldsymbol{b}_k^{h,h'}$	.(.)	•(.)	.(.)	•(•)

model is said to be identifiable when it satisfies:

$$\left(\exists(\boldsymbol{\psi},\tilde{\boldsymbol{\psi}}) \in \boldsymbol{\Psi}^2, \forall \boldsymbol{x} \in \mathbb{R}^d, g(\boldsymbol{x};\boldsymbol{\psi}) = g(\boldsymbol{x};\tilde{\boldsymbol{\psi}})\right) \Rightarrow \left(\exists \sigma \in \mathscr{S}_H, \exists \tau \in \mathscr{S}_K : \tilde{\boldsymbol{\psi}} = \boldsymbol{\psi}_\tau^\sigma\right),$$

where $g(\boldsymbol{x};\boldsymbol{\psi})$ denotes the probability density function of an observed data $\boldsymbol{x}$.

All admissible models are identifiable except one of them, $\left(1, \boldsymbol{D}^{h,h'}, \boldsymbol{b}_k^{h,h'}; \boldsymbol{\pi}^h, \boldsymbol{\Sigma}^h\right)$, since it authorizes different component label permutations depending on the population, and, as a consequence, some crossing of the link between Gaussian components. Indeed, it is easy to show in that case, that any component may be linked to any other one.

However, assuming the data arise from this unidentifiable model must not be rejected since it just leads to combinatorial possibilities in constituting groups of identical labels from the components C_k^h. In that case, matching strategies presented at the end of Subsection 2.1 for independent clustering can be used.

3.5 Model selection

We propose to select some of the previous parsimonious models for simultaneous clustering with the *BIC* criterion defined in (2). Table 2 indicates the number of parameters ν corresponding to the diverse intra and interpopulation model combinations. The maximum log-likelihood value ℓ has to be calculated for each model at hand (see the next section about estimation). The model selected among competing ones corresponds to the smallest computed *BIC* value.

Remark It is important to notice that *BIC* appears also, here, as a natural way for selecting between independent clustering (Subsection 2.1), simultaneous clustering (Subsection 2.2) or clustering under the common origin assumption.

4 Parameter estimation

After a useful reparameterization (Subsection 4.1), a GEM procedure for estimating the model parameters by maximum likelihood is described in Subsections 4.2 to 4.4.

Table 2 *Dimension ν of the parameter ψ in simultaneous clustering in case of equal mixing proportions. $\beta = Kd$ represents the degree of freedom in the parameter component set $\{\boldsymbol{\mu}_k^1\}$ and $\gamma = \frac{d^2+d}{2}$ is the size of $\boldsymbol{\Sigma}_1^1$ parameter component. If mixing proportions π_k^h are free on both h and k (resp. free on k only), then one must add $H(K-1)$ (resp. $K-1$) to the indicated dimensions below.*

		$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
$\boldsymbol{D}^{h,h'}$	$\boldsymbol{b}^{h,h'}$	$\beta+\gamma+2d(H-1)$	$\beta+K\gamma+2d(H-1)$
	$\boldsymbol{b}_k^{h,h'}$	$\beta+\gamma+d(K+1)(H-1)$	$\beta+K\gamma+d(K+1)(H-1)$
$\boldsymbol{D}_k^{h,h'}$	$\boldsymbol{b}_k^{h,h'}$	.	$\beta+K\gamma+2dK(H-1)$

An alternative and simplified estimation process is proposed then, in Subsection 4.5, for some specific models.

4.1 A useful reparameterization

The parametric link between the Gaussian parameters (5) allows a new parameterization of the model at hand, which is useful and meaningful for estimating ψ. It is easy to verify that for any identifiable model, each $\boldsymbol{D}_k^{h,h'}$ matrix is unique and each $\boldsymbol{b}_k^{h,h'}$ vector also. It has sense then to define from any value of the parameter ψ, the following vectors: $\boldsymbol{\theta}^1 = \psi^1$, and for all $h \in \{2,\dots,H\}$: $\boldsymbol{\theta}^h = \left[\left(\alpha_k^h, \boldsymbol{D}_k^h, \boldsymbol{b}_k^h\right); k = 1,\dots,K\right]$, where $\alpha_k^h = \alpha_k^{1,h}$, $\boldsymbol{D}_k^h = \boldsymbol{D}_k^{1,h}$ and $\boldsymbol{b}_k^h = \boldsymbol{b}_k^{1,h}$. Let us note Θ the space described by the vector $\boldsymbol{\theta} = (\boldsymbol{\theta}^1,\dots,\boldsymbol{\theta}^H)$ when ψ scans the parameter space Ψ. There exists a canonical bijective map between Ψ and Θ. Thus $\boldsymbol{\theta}$ constitutes a new parameterization of the model at hand, and estimating ψ or $\boldsymbol{\theta}$ by maximizing their likelihood, respectively on Ψ or Θ, is equivalent.

$\boldsymbol{\theta}^1$ appears to be a "reference population parameter" whereas $(\boldsymbol{\theta}^2,\dots,\boldsymbol{\theta}^H)$ corresponds to a "link parameter" between the reference population and the other ones. But in spite of appearance the estimated model does not depend on the initial choice of population 1. Indeed the bijective correspondence between the parameter spaces Θ and Ψ ensures that the model inference is invariant by relabelling the populations.

4.2 Invoking a GEM algorithm

The log-likelihood of the new parameter $\boldsymbol{\theta}$, computed on the observed data, has no explicit maximum, neither does its completed log-likelihood:

$$l_c(\boldsymbol{\theta};\mathbf{x},\mathbf{z}) = \sum_{h=1}^{H}\sum_{i=1}^{n^h}\sum_{k=1}^{K} z_{i,k}^h \log\left(\alpha_k^h \pi_k^1 \Phi_d\left(\mathbf{x}_i^h; \boldsymbol{D}_k^h\boldsymbol{\mu}_k^1 + \boldsymbol{b}_k^h, \boldsymbol{D}_k^h\boldsymbol{\Sigma}_k^1\boldsymbol{D}_k^h\right)\right), \qquad (7)$$

with $\mathbf{z} = \bigcup_{h=1}^{H}\mathbf{z}^h$ and where we adopt the convention that for all k, $\boldsymbol{D}_k^1$ is the identity matrix of $\mathbb{R}^{d\times d}$ and $\boldsymbol{b}_k^1$ is the null vector of $\mathbb{R}^d$. But Dempster et al (1977) showed

that an EM algorithm is not required in the M-step to converge to a local maximum of the parameter likelihood in an incomplete data structure. The conditional expectation of its completed log-likelihood has just to increase at each M-step instead of being maximized. This algorithm, called GEM (Generalized EM), can be easily implemented here; It consists, at its GM-step, on an alternating optimization of $E\left[l_c(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Z})|\boldsymbol{X}=\mathbf{x}\right]$ where $\boldsymbol{X}$ and $\boldsymbol{Z}$ denote respectively the random version of $\mathbf{x}$ and $\mathbf{z}$. Starting from some initial value of the parameter $\boldsymbol{\theta}$, it alternates the two following steps.

- E-step: From the current value of $\boldsymbol{\theta}$, the expected component memberships (1) are computed and we note them shortly $t_{i,k}^h$.
- GM-step: The conditional expectation of the completed log-likelihood, obtained by substituting $z_{i,k}^h$ for $t_{i,k}^h$ in (7), can be alternatively maximized with respect to the two following component sets of $\boldsymbol{\theta}$ parameter: $\{\pi_k^1,\boldsymbol{\mu}_k^1,\boldsymbol{\Sigma}_k^1\}$ and $\{\alpha_k^h,\boldsymbol{D}_k^h,\boldsymbol{b}_k^h\}$ ($h=2,\ldots,H$). It provides the estimator $\boldsymbol{\theta}^+$ that is used as $\boldsymbol{\theta}$ at the next iteration of the current GM-step. The detail of the GM-step is given in the following two subsections since it depends on the intra and interpopulation model at hand.

The algorithm stops either when reaching stationarity of the likelihood or after a given iteration number.

4.3 Estimation of the reference population parameter $\boldsymbol{\theta}^1$

Mixing proportions π_k^1 Noting $\hat{n}_k^h=\sum_{i=1}^{n^h}t_{i,k}^h$ and $\hat{n}_k=\sum_{h=1}^{H}\hat{n}_k^h$, we obtain $\pi_k^{1^+}=\hat{n}_k^1/n^1$ when assuming that mixing proportions are free, $\pi_k^{1^+}=\hat{n}_k/n$ when they only depend on the component, and $\pi_k^{1^+}=1/K$ when they neither depend on the component nor on the population.

Centers $\boldsymbol{\mu}_k^1$ Component centers in the reference population are estimated by:

$$\boldsymbol{\mu}_k^{1^+}=\frac{1}{\hat{n}_k}\sum_{h=1}^{H}\sum_{i=1}^{n^h}t_{i,k}^h\left(\boldsymbol{D}_k^h\right)^{-1}\left(\mathbf{x}_i^h-\boldsymbol{b}_k^h\right).$$

Covariance matrices $\boldsymbol{\Sigma}_k^1$ If mixtures are assumed to have heteroscedastic components, the covariance matrices in the reference population are given by:

$$\boldsymbol{\Sigma}_k^{1^+}=\frac{1}{\hat{n}_k}\sum_{h=1}^{H}\sum_{i=1}^{n^h}t_{i,k}^h\left[\left(\boldsymbol{D}_k^h\right)^{-1}\left(\mathbf{x}_i^h-\boldsymbol{b}_k^h\right)-\boldsymbol{\mu}_k^{1^+}\right]\left[\left(\boldsymbol{D}_k^h\right)^{-1}\left(\mathbf{x}_i^h-\boldsymbol{b}_k^h\right)-\boldsymbol{\mu}_k^{1^+}\right]'.$$

Otherwise, when supposing each mixture has homoscedastic components, the covariance matrices in P^1 are estimated by:

$$\boldsymbol{\Sigma}^{1^+}=\frac{1}{n}\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{i=1}^{n^h}t_{i,k}^h\left[\left(\boldsymbol{D}^h\right)^{-1}\left(\mathbf{x}_i^h-\boldsymbol{b}_k^h\right)-\boldsymbol{\mu}_k^{1^+}\right]\left[\left(\boldsymbol{D}^h\right)^{-1}\left(\mathbf{x}_i^h-\boldsymbol{b}_k^h\right)-\boldsymbol{\mu}_k^{1^+}\right]'.$$

4.4 Estimation of the link parameters $\boldsymbol{\theta}^h$ ($h \geq 2$)

Scalars α_k^h We have $\alpha_k^{h^+} = \hat{n}_k^h / (n^h \pi_k^{1^+})$ when assuming that the mixing proportions are free and $\alpha_k^{h^+} = 1$ when the mixing proportions only depend on the component or when they neither depend on the component nor on the population.

Vectors $\boldsymbol{b}_k^h$ Noting $\overline{\mathbf{x}}_k^h = (1/\hat{n}_k^h) \sum_{i=1}^{n^h} t_{i,k}^h \mathbf{x}_i^h$ the empirical mean of C_k^h component, when vectors $\boldsymbol{b}_k^h$ ($k = 1, \ldots, K$) are assumed to be free for any $h \in \{2, \ldots, H\}$, they are estimated by the differences $\boldsymbol{b}_k^{h^+} = \overline{\mathbf{x}}_k^h - \boldsymbol{D}_k^h \boldsymbol{\mu}_k^{1^+}$, and by:

$$\boldsymbol{b}^{h^+} = \left[\sum_{k=1}^{K} \hat{n}_k^h \left(\boldsymbol{D}_k^h \boldsymbol{\Sigma}_k^{1^+} \boldsymbol{D}_k^h \right)^{-1} \right]^{-1} \left[\sum_{k=1}^{K} \hat{n}_k^h \left(\boldsymbol{D}_k^h \boldsymbol{\Sigma}_k^{1^+} \boldsymbol{D}_k^h \right)^{-1} \left(\overline{\mathbf{x}}_k^h - \boldsymbol{D}_k^h \boldsymbol{\mu}_k^{1^+} \right) \right], \tag{8}$$

when supposing they are equal.

Matrices $\boldsymbol{D}_k^h$ $\boldsymbol{D}_k^h$ matrices can not be estimated explicitly but, as the conditional expectation of the completed log-likelihood $E\left[l_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) | \boldsymbol{X} = \mathbf{x} \right]$ is concave with respect to $(\boldsymbol{D}_k^h)^{-1}$ (whatever are $h \in \{2, \ldots, H\}$ and $k \in \{1, \ldots, k\}$), we obtain $\boldsymbol{D}_k^{h^+}$ by any convex optimization algorithm. This concavity property can be proved in the same way as Theorem 2 in Biernacki et al (2002), Appendix A.

Remark Until now we have supposed that $\boldsymbol{D}_k^h$ matrices were positive. If that assumption is weakened by simply fixing each $\boldsymbol{D}_k^h$ matrix coefficient sign (positive or negative), then, firstly, identifiability of the model is preserved, and, secondly, the conditional expectation of the completed log-likelihood is still concave with respect to $(\boldsymbol{D}_k^h)^{-1}$ on the parameter space Θ. Then we will always be able to get $\boldsymbol{D}_k^{h^+}$ at the GM-step of the GEM algorithm, numerically at less.

4.5 An alternative sequential estimate

According to Subsections 4.3 and 4.4, ψ estimate based on ML relies on an alternate likelihood optimization with respect to the reference parameter $\boldsymbol{\theta}^1$ and to the link parameter $\boldsymbol{\theta}^h$ ($h \geq 2$). However the interpopulation model $(1, \boldsymbol{D}^{h,h'}, \boldsymbol{b}^{h,h'})$ allows an alternative two steps estimation which does not maximize ψ likelihood in general, but which is simpler than the previous GEM algorithm and which leads also to consistent estimates. Indeed, the conditional link (4) stretches over unconditional populations:

$$\boldsymbol{X}^{h'} \overset{\mathscr{D}}{=} \xi_k^{h,h'} \boldsymbol{D}^{h,h'} \boldsymbol{X}^h + \boldsymbol{b}^{h,h'}. \tag{9}$$

Still using both notations $\boldsymbol{D}^h = \boldsymbol{D}^{1,h}$ and $\boldsymbol{b}^h = \boldsymbol{b}^{1,h}$, the first step of the proposed strategy corresponds to a *standardization* process aiming to estimate each population link parameter $(\boldsymbol{D}^h, \boldsymbol{b}^h)$ with each sample pair $(\mathbf{x}^1, \mathbf{x}^h)$ ($h = 2, \ldots, H$). This can be performed very simply by the following simple and explicit estimates:

$$\hat{\boldsymbol{D}}^h = \left(\operatorname{diag} \hat{\boldsymbol{S}}^h \right)^{1/2} \left(\operatorname{diag} \hat{\boldsymbol{S}}^1 \right)^{-1/2} \quad \text{and} \quad \hat{\boldsymbol{b}}^h = \overline{\mathbf{x}}^h - \hat{\boldsymbol{D}}^h \overline{\mathbf{x}}^1, \tag{10}$$

where $\overrightarrow{\mathbf{x}}^h = (1/n^h)\sum_{i=1}^{n^h}\mathbf{x}_i^h$ and $\hat{\boldsymbol{S}}^h = (1/n^h)\sum_{i=1}^{n^h}(\mathbf{x}_i^h - \overrightarrow{\mathbf{x}}^h)(\mathbf{x}_i^h - \overrightarrow{\mathbf{x}}^h)'$ denote respectively the empirical center and the empirical covariance matrix of the whole h-th population. The estimate of $\boldsymbol{D}^h$ is based on the relation $[\boldsymbol{S}^h = \boldsymbol{D}^h\boldsymbol{S}^1\boldsymbol{D}^h] \Rightarrow [(\text{diag } \boldsymbol{S}^h) = \boldsymbol{D}^h(\text{diag } \boldsymbol{S}^1)\boldsymbol{D}^h]$ which is a direct consequence of (9) and where $\boldsymbol{S}^h$ denotes the covariance matrix of the whole population h.

The second step of the strategy relies on the following point: According to (9) all standardized data points $\tilde{\mathbf{x}}_i^h = (\boldsymbol{D}^h)^{-1}(\mathbf{x}_i^h - \boldsymbol{b}^h)$ $(i = 1,\ldots,n^h)$ arise independently from P^1, whatever is h. Then the second step consists on involving a simple and traditional EM algorithm devoted to Gaussian mixture estimation, on the whole standardized data $\tilde{\mathbf{x}}_i^h$ $(h = 1,\ldots,H,\ i = 1,\ldots,n^h)$ obtained by *plug-in* of estimates $\hat{\boldsymbol{D}}^h$ and $\hat{\boldsymbol{b}}^h$. Softwares as MIXMOD (Biernacki et al, 2006) are now available for practitioners to perform that estimation.

Consistency of the estimate $\hat{\psi}$ related to this alternative sequential estimation is preserved since the whole procedure combines in a *plug-in* way consistent estimates:

- $\overrightarrow{\mathbf{x}}^h$ and $\hat{\boldsymbol{S}}^h$ are well-known consistent empirical estimates;
- $\hat{\boldsymbol{D}}^h$ and $\hat{\boldsymbol{b}}^h$ in (10) are estimators obtained by equalizing both moments of first order, and moments of second order from (9);
- P^1 parameter is based on ML.

Remark $\hat{\psi}$ related to this sequential estimation does not depend on which sample holds the label 1 since (*i*) the constraint set on ψ likelihood does not depend on this population label choice and (*ii*) the link parameter owns some symmetry and transitivity properties which are also satisfied by these new estimators.

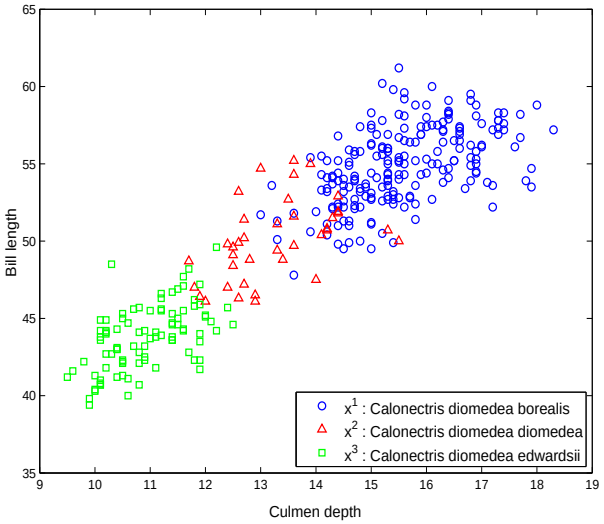
5 A biological example

5.1 The data

In Thibault et al (1997) three seabird subspecies ($H = 3$) of Shearwaters, differing over their geographical range, are described. *Borealis* (sample $\mathbf{x}^1$, size $n^1 = 206$ individuals) are living in the Atlantic Islands (Azores, Canaries, etc.), *diomedea* (sample $\mathbf{x}^2$, size $n^2 = 38$ individuals), in Mediterranean Islands (Balearics, Corsica, etc.), and *edwardsii* (sample $\mathbf{x}^3$, size $n^3 = 92$ individuals), in Cape Verde Islands. 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 1 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. However, let us remark that the researched partitions could be expected to have the same number of clusters with the same partition meaning in each sample since all of them arise from the same species *calonectris diomedea*. 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

Fig. 1 *Three samples of Cory's Shearwaters described by variables of identical meaning.*



simultaneous and independent clustering compete. Results are given in the next subsection.

5.2 Results of simultaneous *vs.* independent clustering

We applied on the three seabird samples each of the 15 allowed models of simultaneous clustering displayed in Table 1 for different number of clusters ($K = 1,\dots,4$) and with the GEM algorithm (5 trials for each procedure, 500 iterations and 5 directional maximizations at each GM step; see Subsection 4.2). Independent clustering is also applied for each sample with a common value of K in the same range than simultaneous clustering ($K = 1,\dots,4$) and for four standard models: Homoscedastic or heteroscedastic with free or not mixing proportions (same model for each sample). At last four models are inferred under the common origin assumption: Each one assumes that the conditional covariance matrices (resp. the mixing proportions) of the common original Gaussian mixture, are free or equal.

Table 3 displays the best *BIC* criterion value among all models for the three clustering strategies. The overall best *BIC* value (4071.8) is obtained from simultaneous clustering for $K = 2$ groups. This value is widely better than the best *BIC* obtained from independent clustering ($BIC = 4102.6$), or from the models of common origin ($BIC = 4341.7$). So *BIC* clearly prefers the simultaneous clustering method and rejects, here, the two standard other ones.

As the two best *BIC* values from simultaneous clustering (4071.8 and 4073.3) are close, they produce an ambiguity about the structure of each bird subspecies: $K = 2$ *vs* $K = 1$ group. But this indecision of *BIC* should not be considered as a drawback since

it reveals a real overlapping between different genders (we can see it on Figure 1 for instance where it is difficult to guess gender). Note that the simultaneous procedure is the only method which informs the practitioner about this fact since independent and "common" clusterings lead to select respectively $K = 1$ and $K = 4$ groups without ambiguity (but with a worse *BIC* value than in the simultaneous clustering situation).

Table 3 *Best BIC values obtained in clustering the Cory's Shearwaters simultaneously (full ML estimates), independently or under the assumption of common origin, with different number of clusters.*

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
Common Origin	4472.0	4371.8	4349.3	4341.7

Retaining the two cluster solution for all models of all clustering strategies, we propose now to compare the estimated partition with the gender partition of birds (males/females). The associated errors rates given in Table 4 clearly indicate that the best simultaneous clustering model $(1, \boldsymbol{D}^{h,h'}, \boldsymbol{b}^{h,h'}; \boldsymbol{\pi}^h, \boldsymbol{\Sigma}^h)$ leads to a smaller global error rate (10.71%) than the best independent clustering model for $K = 2$, $(\boldsymbol{\pi}^h, \boldsymbol{\Sigma}^h)$, which reaches 12.50%. Moreover Figure 2 and the confusion table given in Table 5, highlight the following point. The partitions estimated from simultaneous and from independent clustering are close to each other; however the partition related to the simultaneous methodology improves not only the global error rate, but provides also a higher agreement between the estimated clusters and the true gender partition of the birds. In addition, the selected interpopulation model clearly indicates that evolution of subspecies over the geographical range is independent on the gender. It is a readable and meaningful information for biologists. We retrieve also what biologists usually know which is homoscedastic components with equal mixing proportions (information given by the intrapopulation model).

According to Subsection 2.3, the overall best model ($BIC = 4071.8$) involves that for males (as for females), any couple of biometrical variables is homogeneously correlated through the subspecies. On the one hand, this result is corroborated by a test of hypothesis (see Scherrer, 2007, p. 659): At a significance level of 5% any couple of conditional variables is identically correlated among *borealis*, *diomedea* and *edwardsii*. On the other hand this assumption, even though non standard, seems to make sense for ornithologists.

The search for a structure in the Shearwaters samples highlights some additional advantage of the simultaneous clustering models. Without any link between P^2 and other mixtures (independent clustering), the size of ψ^2 parameter is 83 for the most complex model $(\boldsymbol{\pi}_k^h, \boldsymbol{\Sigma}_k^h)$ when $K = 4$ whereas the corresponding sample size is only $n^2 = 38$ in $\mathbf{x}^2$ sample. As a consequence, the estimates of mixture P^2 based on ML could be somewhat hazardous... Contrariwise, simultaneous clustering involves all data sets $\{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3\}$ for estimating parameters of P^2 (and simultaneously of P^1 and

Table 4 *BIC value and (error rate in %) obtained in clustering the Cory's Shearwaters simultaneously (full ML estimates), independently or under the assumption of common origin, in case of $K = 2$ groups.*

			π^h		π_k^h	
			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
1	$\boldsymbol{D}^{h,h'}$	$\boldsymbol{b}^{h,h'}$	**4071.8** (**10.71**)	4096.9 (12.20)	4074.7 (10.71)	4099.3 (14.58)
		$\boldsymbol{b}_k^{h,h'}$	4094.9 (33.33)	4122.2 (11.31)	4101.9 (41.96)	4122.7 (15.77)
	$\boldsymbol{D}_k^{h,h'}$	$\boldsymbol{b}_k^{h,h'}$	.	4154.5 (38.39)	.	4147.9 (25.29)
$\alpha_k^{h,h'}$	$\boldsymbol{D}^{h,h'}$	$\boldsymbol{b}^{h,h'}$	.	.	4079.9 (39.88)	4107.8 (40.18)
		$\boldsymbol{b}_k^{h,h'}$	.	.	4107.6 (42.86)	4128.5 (15.18)
	$\boldsymbol{D}_k^{h,h'}$	$\boldsymbol{b}_k^{h,h'}$	.	.	.	4153.6 (16.37)
Independent			**4139.8** (**12.50**)	4218.2 (38.39)	4143.0 (29.17)	4219.7 (40.18)
Common Origin			4392.9 (**43.45**)	**4392.5** (44.94)	4371.8 (45.24)	4383.6 (43.45)

Table 5 *Comparison of the birds gender within each subspecies to the inferred clusters related to the best model ($K = 2$ groups) from independent and from simultaneous clustering.*

		independent clustering		simultaneous clustering	
		cluster 1	cluster 2	cluster 1	cluster 2
borealis	male	20	93	18	95
	female	88	5	89	4
diomedea	male	1	15	2	14
	female	18	4	18	4
edwardsii	male	7	37	5	39
	female	43	5	45	3

P^3) thanks to the link between all populations. For instance the parameter of the most complex model $(\alpha_k^{h,h'}, \boldsymbol{D}_k^{h,h'}, \boldsymbol{b}_k^{h,h'}; \pi_k^h, \boldsymbol{\Sigma}_k^h)$ has $\nu = 169$ degrees of freedom for $K = 4$ but the data involved in the estimation countains now all the $n = 336$ birds.

Remark Table 6 displays *BIC* values and all associated errors rates obtained by sequential estimation (Subsection 4.5). *BIC* values are greater than the corresponding *BIC* of Table 4–except one of them which corresponds to a parameter located on a degeneracy path of the likelihood and which is associated to a bad error rate–but both corresponding *BIC* values are often close to each other and the corresponding error rates also.

That example shows that the alternative sequential method can cheaply provide some acceptable partition close to the one which the full ML parameter estimate would lead to. Remember however that this alternative strategy is available only for four models of simultaneous clustering.

Fig. 2 *Similarities (hollow symbols) and differences (solid symbols) between independent clustering and simultaneous clustering in sexing each Shearwater sample ($K = 2$ groups).*

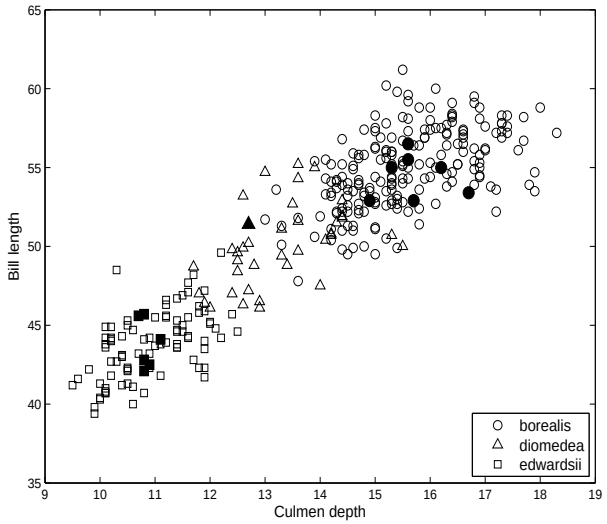


Table 6 *Sequential estimation: BIC value and (error rate in %) in simultaneous clustering (2 groups) of Shearwaters.*

			π^h		π_k^h	
			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
1	$\boldsymbol{D}^{h,h'}$	$\boldsymbol{b}^{h,h'}$	**4072.4 (10.42)**	4097.5 (11.90)	4074.3 (34.52)	4099.7 (14.28)

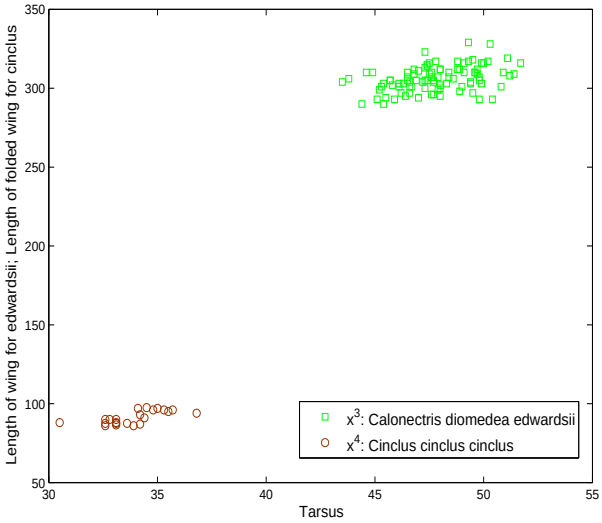
5.3 Some robustness study of the simultaneous clustering method

Simultaneous clustering relies on the assumption that samples to be classified are described by both variables and partition of identical meaning. In order to conduct a robustness study for simultaneous clustering, we propose now to slightly relax the first assumption.

We dispose of another bird sample $\mathbf{x}^4$ (size $n^4 = 22$ individuals) (D'Amico et al, 2009) composed of White-throated Dippers (*Cinclus cinclus cinclus*) living in Lorraine (France) which are described by only two morphological variables which are their tarsus and the length of their *folded* wing. We aim now to classify simultaneously this new sample $\mathbf{x}^4$ and the previous sample $\mathbf{x}^3$. In order to perform this partitioning, we keep the two variables in $\mathbf{x}^3$ which are the closer in meaning to the couple tarsus and length of folded wing, thus the couple tarsus-wing length. Figure 3 displays both samples $\mathbf{x}^3$ and $\mathbf{x}^4$ with this couple of axes. We have also to notice that species of $\mathbf{x}^3$ and $\mathbf{x}^4$ are different since they correspond respectively to *diomedea* and

cinclus. Since some assumptions of simultaneous clustering are now violated, it is clear that this strategy becomes somewhat more challenging.

Fig. 3 *Two bird samples described by variables close in meaning.*



Following a procedure similar to the one in Subsection 5.2. Table 7 displays the best *BIC* values among all models for both simultaneous and independent clustering. (We do not consider here that both samples might arise from the same population: Figure 3 clearly shows that any model based on this assumption would have some low likelihood and would be obviously rejected by *BIC*.) This time, both partitioning strategies which compete, retain two clusters but, again, the simultaneous one is preferred by *BIC*. Table 7 displays also *BIC* values and error rate (for the gender) and again the model retained by simultaneous clustering (model $(1, \boldsymbol{D}^{h,h'}, \boldsymbol{b}_k^{h,h'}; \pi^h, \boldsymbol{\Sigma}^h)$) leads to a lower error rate (21.93%) than this one of the model retained by independent clustering (model $(\pi^h, \boldsymbol{\Sigma}^h)$ and error rate 23.68%). Thus, relaxing requirements of simultaneous clustering leads here to select a more complex model than in the previous experiments but preserves the superiority of independent clustering both on *BIC* and error rate values.

Table 7 *Best BIC values obtained in simultaneous (full ML estimates) and independent clustering in the robustness study with different number of clusters.*

Cluster Number	1	2	3	4
Simultaneous Clustering	614.81	**613.51**	620.12	624.94
Independent Clustering	615.71	615.15	622.45	659.91

Table 8 *BIC value and (error rate in %) obtained in simultaneous (full ML estimates) and independent clustering (2 groups) of two bird samples in the robustness study.*

			π^h		π_k^h	
			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
1	$\boldsymbol{D}^{h,h'}$	$\boldsymbol{b}^{h,h'}$	617.12 (**20**.**18**)	622.67 (21.05)	618.42 (24.56)	624.79 (49.12)
		$\boldsymbol{b}_k^{h,h'}$	**613.51** (21.93)	619.25 (39.41)	615.66 (23.68)	630.14 (42.98)
	$\boldsymbol{D}_k^{h,h'}$	$\boldsymbol{b}_k^{h,h'}$	.	627.83 (38.60)	.	625.14 (47.37)
$\alpha_k^{h,h'}$		$\boldsymbol{b}^{h,h'}$	.	.	623.52 (40.35)	627.06 (46.49)
	$\boldsymbol{D}^{h,h'}$	$\boldsymbol{b}_k^{h,h'}$	.	.	617.98 (41.23)	624.69 (38.60)
	$\boldsymbol{D}_k^{h,h'}$	$\boldsymbol{b}_k^{h,h'}$	.	.	.	620.94 (46.49)
Independent			**615**.**15** (23.68)	622.92 (**21**.**93**)	619.96 (23.68)	625.96 (29.82)

6 Concluding remarks

This work is a scope enlargement of clustering based on Gaussian mixtures. It displays models allowing to classify automatically and simultaneously several samples even when they arise from different populations. It is based on the assumption of a linear stochastic link between the components of the mixtures which translates identical conditional correlations of the descriptors through the populations. Full ML estimates are proposed through a GEM procedure. Alternatively, for some models, it is possible to perform an estimation with traditional tools available for any statistician or practitioner: Explicit estimates given by a simple normalization followed by a standard EM algorithm for Gaussian mixtures. We showed the efficiency of the models on biological data both for selecting the number of groups and for retrieving a meaningful partition (here the gender of birds).

We noticed also that the so-called simultaneous clustering method had some kind of robustness when relaxing one of its main assumption which is the exact concordance of population descriptors. In the same spirit, it would be challenging in future works to relax another important assumption which is the canonical directional effects, denoted by $\mathscr{H}_1$ in the paper. Notice that $\mathscr{H}_2$ (regularity of the link) and $\mathscr{H}_3$ (identifiability constraint) are more minor hypotheses since they only correspond to quite technical and standard properties. We expect that relaxing $\mathscr{H}_1$ would lead to some interesting but quite opened mathematical problems for properly defining new possible stochastic links between conditional populations. Other relevant challenges can be also easily identified:

- A first natural question is to adapt the current Gaussian simultaneous model to the situation where some classes are empty within one or several samples (males are missing in one subpopulation for instance). If such a situation is suspected, it is possible to consider some mixtures with less components but this solution may lead to some combinatorial problems to be properly addressed.
- A second natural question is to extend the present method devoted to Gaussian mixture models to other kinds of distributions. For instance, we can consider (see

McLachlan and Peel, 2000, for each situation) mixtures of factor analyzers (high-dimensional data sets), mixtures of Student t-distributions (robust clustering) or other mixtures of non-continuous distributions (latent class model for categorical data for instance). Obviously, some other kinds of links (than the linear one) between populations can (have to) be considered, for instance some original ideas like group overlapping preservation, typically by using the partition entropy information.

Finally, we think that simultaneous clustering models could be combined with other existing ones. In particular, they may provide parsimonious and meaningful links between lower-level unit distributions depending on different higher-level units in the multilevel classification model of nested data (Vermunt and Magidson, 2005, pp.176–183).

Acknowledgements The authors wish to thank F. D'Amico, Y. Lalanne, J. O'Halloran and P. Smiddy for authorizing them to work on their White-throated Dipper data and V. Bretagnolle for his Cory's Shearwater dataset. They thank also Sandra McJannett and Anne-Marie Pollaud-Dulian for their advice and they express their gratitude to the Associate Editor and to two anonymous Reviewers for their very useful comments.

References

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49:803–821

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

Biernacki C, Celeux G, Govaert G, Langrognet F (2006) Model-based cluster and discriminant analysis with the mixmod software. Computational Statistics and Data Analysis 51:587–600

Celeux G, Govaert G (1995) Gaussian parsimonious clustering models. Pattern Recognition 28(5):781–793

D'Amico F, Lalanne Y, O'Halloran J, Smiddy P (2009) Personal communication

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39:1–38

Du Jardin P, Séverin E (2010) Dynamic analysis of the business failure process: a study of bankruptcy trajectories. In: Portuguese Finance Network, Ponte Delgada, Portugal

Flury BN (1983) Common principal components in k groups. Journal of the American Statistical Association 79:892–898

Fraley C, Raftery AE (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. Computer Journal 41:578–588

Govaert G, Nadif M (2008) Block clustering with mixture models: Comparison of different approaches. Computational Statistics and Data Analysis 52:3233–3245

Gower JC (1975) Generalized procrustes analysis. Psychometrika 40:33–51

Lebarbier E, Mary-Huard T (2006) Le critère bic, fondements théoriques et interprétation. Journal de la Société Francaise de Statistique 1:39–57

Lourme A (2011) Contribution à la Classification par Modèles de Mélange et Classification Simultanée d'Echantillons d'Origines Multiples. PhD thesis, Laboratoire Paul Painlevé, Université des Sciences et Techniques Lille 1

Lourme A, Biernacki C (2010) Simultaneous Gaussian models-based clustering for samples of multiples origins. Pub. Irma, University Lille 1, Lille

McLachlan GJ, Peel D (2000) Finite Mixture Models. Wiley, New York

Roeder K, Wasserman L (1997) Practical Bayesian density estimation using mixtures of normals. Journal of the American Statistical Association 92:894–902

Scherrer B (2007) Biostatistique, vol 1. Gaëtan Morin

Schork N, Thiel B (1996) Mixture distributions in human genetics. Statistical Methods in Medical Research 39:155–178

Schwarz G (1978) Estimating the number of components in a finite mixture model. Annals of Statistics 6:461–464

Thibault J, Bretagnolle V, Rabouam C (1997) Cory's shearwater calonectris diomedea. Birds of Western Paleartic Update 1:75–98

Vermunt J, Magidson J (2005) Hierarchical mixture models for nested data structures, Classification: The Ubiquitous Challenge. Wiley, Springer, Heidelberg, Weihs, C. and Gaul, W. eds.