PUB. IRMA, LILLE 2010
Vol. 70, N° VII

Simultaneous Gaussian Model-Based Clustering for Samples of Multiple Origins

Alexandre LOURME [a, b], Christophe BIERNACKI [b]

Abstract

Mixture model-based clustering usually assumes that the data arise from a mixture population in order to estimate some hypothetical underlying partition of the dataset. In this work, we are interested in the case where several samples have to be clustered at the same time, that is when the data arise not only from one but possibly from several mixtures. In the multinormal context, we establish a linear stochastic link between the components of the mixtures wich enables the joint-estimate of their parameters–estimations are performed here by maximum likelihood–and the simultaneous classification of the diverse samples. We propose several useful models of constraint on this stochastic link, and we give their parameter estimators. The interest of these models is highlighted in a biological context where some birds belonging to several species have to be classified according to their sex. We show firstly that our simultaneous clustering method does improve the partition obtained by clustering independently each sample. We then show that this method is also efficient in assessing the cluster number when assuming it is unknown. Finally some additional experiments are performed to show the robustness of our simultaneous clustering method when one of its main assumptions is relaxed.

Résumé

Lorsqu'on classifie des données il est courant de supposer qu'elles proviennent d'une population mélange pour en estimer une éventuelle partition sous-jacente. Nous nous intéressons ici au cas où plusieurs échantillons doivent être classifiés en même temps, c'est-à-dire au cas où la donnée ne provient pas seulement d'une, mais éventuellement de plusieurs populations mélange. Dans un contexte multinormal nous établissons un lien linéaire stochastique entre les composantes des mélanges, qui permet d'estimer de façon conjointe leur paramètre–les estimations sont réalisées ici par maximum de vraisemblance–et de classifier simultanément les différents échantillons. Nous proposons plusieurs modèles de contrainte, utiles et réalistes, portant sur le lien stochastique établi, et nous donnons

a. Université de Pau et des Pays de l'Adour, IUT département Génie Biologique, 371 rue du Ruisseau, 40000 Mont de Marsan, France.

b. Laboratoire P. Painlevé, UMR 8524 CNRS Université Lille I, Bât M2, Cité Scientifique, F-59655 Villeneuve d'Ascq Cedex, France.

l'estimateur de leur paramètre. L'intérêt de ces modèles est mis en lumière dans un contexte biologique où des oiseaux d'espèces différentes doivent être classifiés selon leur sexe. Nous montrons dans un premier temps que notre méthode de classification simultanée améliore la partition obtenue en classifiant indépendamment les échantillons. Nous montrons ensuite que cette méthode est aussi efficace pour déterminer le nombre de groupes lorsqu'on l'ignore. Des expériences complémentaires sont finalement réalisées pour montrer la robustesse de notre méthode de classification simultanée à la relaxation de l'une de ses principales hypothèses.

MSC 2000 subject classifications. Primary-?????; secondary-?????.

Key words and phrases. Biological features; Distributional relationship; EM algorithm; Gaussian mixture; Model-based clustering; Model selection.

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 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 [13] for a review). In particular, in the multivariate continuous situation, Gaussian mixture model-based clustering has found successful applications in diverse fields: Genetics [15], medicine [13], magnetic resonance imaging [1], astronomy [4]. Consequently, nowadays, involving such models for clustering a given dataset could be considered as familiar to every statistician as to more and more practitioners.

In many situations, one needs to cluster several datasets, possibly arising from different populations, instead of a single one, into partitions having both the same number of clusters and identical meaning. For instance, in biology, Thibault et al. [17] described three samples of seabirds living in several geographic zones, leading to very different morphological variables (tarsus, bill length, etc.). The clustering purpose here could be to retrieve the sex of birds from these features. In such a situation, a standard clustering process could be independently applied to each dataset. 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. 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–estimations are performed here 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 [16] enables us then to compare both clustering methods: The simultaneous clustering method which assumes a stochastic link between the populations, and the independent clustering method which considers that populations are unrelated.

Generalizing a one-sample method to several samples is common in statistical literature. Flury [8], for example, proposes the use a particuliar 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. Gower [10] generalizes to K samples ($K \geq 3$) the classical Procrustes analysis which estimates a geometrical link, established between two samples. Hierarchical mixture models [18] for a last example, devoted to nested data classification, can be viewed as specific mixtures allowing to classify several samples at the same time. Our models differ from those on our knowledge of level-2 cluster memberships and also on our exclusive multinormal conditional population hypothesis.

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 less-expensive least square 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). Let us remind 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. 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 P^h. In addition, all populations are described by the same d continuous variables.

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 $\boldsymbol{X}^h$ following a K-modal mixture P^h of non degenerate Gaussian components C_k^h ($k = 1, \ldots, K$), with probability density function:

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

Coefficients π_k^h ($k = 1, \ldots, K$) 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 $_d(\boldsymbol{x}; \boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h)$ denotes its probability density function. The whole parameter of P^h mixture is $\boldsymbol{\psi}^h = (\boldsymbol{\psi}_k^h)_{k=1,\ldots,K}$ where $\boldsymbol{\psi}_k^h = (\pi_k^h, \boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h)$.
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 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 $(\pi_1^h, \ldots, \pi_K^h)$.
The complete data model assumes that couples $(\mathbf{x}_i^h, \mathbf{z}_i^h)_{i=1,\ldots,n^h}$ 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 k-th component Z_k^h equals 1

(and the others 0) with probability π_k^h, and $\left(\boldsymbol{X}^h | Z_k^h = 1\right) \sim \quad_d(\,.\,; \boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h)$. We note also $\boldsymbol{z}^h = \{\boldsymbol{z}_1^h, \ldots, \boldsymbol{z}_{n^h}^h\}$.

Estimating $\boldsymbol{\psi} = (\boldsymbol{\psi}^h)_{h=1,\ldots,H}$, 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 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 [13] 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 $\hat{\boldsymbol{\psi}}$:

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

Since the partition estimated by independent clustering is arbitrarily numbered, the practitioner has if necessary, to renumber some clusters in order to assign the same index to clusters having the same meaning for all populations. 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 H_1, H_2, H_3 described bellow.
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:

$$(\boldsymbol{X}^{h'} | Z_k^{h'} = 1) \sim \xi_k^{h,h'}\left(\boldsymbol{X}^h | Z_k^h = 1\right). \tag{2}$$

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 ($j \in \{1, \ldots, d\}$) component $\left(\xi_k^{h,h'}\right)^{(j)}$ of $\xi_k^{h,h'}$ map depends only on the j-th component $\boldsymbol{x}^{(j)}$ of $\boldsymbol{x}$, situation that is expressed by the following hypothesis:

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

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 covariate $(\boldsymbol{X}^h | Z_k^h = 1)^{(j)}$ into the corresponding conditional Gaussian covariate $(\boldsymbol{X}^{h'} | Z_k^{h'} = 1)^{(j)}$. Assuming moreover that $\left(\xi_k^{h,h'}\right)^{(j)}$ is continuously differentiable–this assumption about all superscripts j is noted H_2–, then the only possible transformation is an affine map. Indeed, De Meyer et al. [6] have shown that for two given non-degenerate univariate normal distributions, there exists 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.
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:

$$(\boldsymbol{X}^{h'} | Z_k^{h'} = 1) \sim \boldsymbol{D}_k^{h,h'} (\boldsymbol{X}^h | Z_k^h = 1) + \boldsymbol{b}_k^{h,h'}. \tag{3}$$

Relation (2) constitutes the keystone of the simultaneous Gaussian model-based clustering framework, and (3) is its affine form involved from the two previous hypotheses H_1 and H_2.

For now as components C_k^h are non degenerate, $\boldsymbol{D}_k^{h,h'}$ matrices are non singular. Let us assume henceforward that any couple of corresponding conditional covariables $(\boldsymbol{X}^h | Z_k^h = 1)^{(j)}$ and $(\boldsymbol{X}^{h'} | Z_k^{h'} = 1)^{(j)}$ are positively correlated. That assumption–noted H_3–involves that $\boldsymbol{D}_k^{h,h'}$ matrices are positive, and means that covariable correlation signs, within some conditional population, remain through the populations. Although it seems to be realistic in many practical contexts as in our biological example below (Section 5), this assumption 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:

$$\boldsymbol{\Sigma}_k^{h'} = \boldsymbol{D}_k^{h,h'} \boldsymbol{\Sigma}_k^h \boldsymbol{D}_k^{h,h'} \quad \text{and} \quad \boldsymbol{\mu}_k^{h'} = \boldsymbol{D}_k^{h,h'} \boldsymbol{\mu}_k^h + \boldsymbol{b}_k^{h,h'}. \tag{4}$$

(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 (4) characterizes henceforward the whole parameter space 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 useful and statistically meaningful interpretation of the linear stochastic link

Each covariance matrix can be decomposed into :

$$\boldsymbol{\Sigma}_k^h = \boldsymbol{T}_k^h \boldsymbol{R}_k^h \boldsymbol{T}_k^h, \tag{5}$$

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 (5) is unique, Relation (4) 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 (3) is equivalent therefore to postulating that conditional correlations are equal through the populations.

This interpretation of the affine link between the conditional populations (3) allows the model to keep all its sense when simultaneous clustering is envisaged in a relaxed context–as in Subsection 5.4–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 on both mixing proportions and Gaussian parameters, within each mixture, with meaningful constraints on the parametric link (4) between conditional populations.

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 ($\boldsymbol{\Sigma}_k^h = \boldsymbol{\Sigma}^h$) or heteroscedastic, their mixing proportions may be equal (π) or free (π_k) (see [13], 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 [4] 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

Thus we can also imagine some meaningful constraints on the parametric link (4). In the most general case, $\boldsymbol{D}_k^{h,h'}$ matrices are definite-positive and diagonal. Moreover they could be variable-independent ($\boldsymbol{D}_k^{h,h'} = \alpha_k^{h,h'}\boldsymbol{I}, \alpha_k^{h,h'} \in \mathbb{R}_*^+$), component-independent ($\boldsymbol{D}_k^{h,h'} = \boldsymbol{D}^{h,h'}$), both component and variable-independent ($\boldsymbol{D}_k^{h,h'} = \alpha^{h,h'}\,\boldsymbol{I}, \alpha^{h,h'} \in \mathbb{R}_*^+$). They could even be all equal to identity matrix ($\boldsymbol{D}_k^{h,h'} = \boldsymbol{I}$) when considering that components C_k^h ($h = 1, \ldots, H$) only differ in their center. The vectors $\boldsymbol{b}_k^{h,h'}$ themselves may be unconstrained ($\boldsymbol{b}_k^{h,h'}$ free), component-independent ($\boldsymbol{b}_k^{h,h'} = \boldsymbol{b}^{h,h'}$), or null ($\boldsymbol{b}_k^{h,h'} = \boldsymbol{0}$). Finally we can suppose the mixing proportion vectors $(\pi_1^h, \ldots, \pi_K^h)$ ($h = 1, \ldots, H$) to be free (π^h) or equal (π). These models will be called *interpopulation models* and they have to be combined with some intrapopulation model.
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 (4), 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.

Some of the previous interpopulation models have a meaningful statistical interpretation. Assuming $\boldsymbol{b}_k^{h,h'}$ vectors to be null with unconstrained $\boldsymbol{D}_k^{h,h'}$ matrices for example leads us to suppose that each conditional covariable has identical coefficients of variation through the populations. Indeed in that case (4) becomes:

$$\boldsymbol{\Sigma}_k^{h'} = \boldsymbol{D}_k^{h,h'}\,\boldsymbol{\Sigma}_k^h\,\boldsymbol{D}_k^{h,h'} \quad \text{and} \quad \boldsymbol{\mu}_k^{h'} = \boldsymbol{D}_k^{h,h'}\boldsymbol{\mu}_k^h. \tag{6}$$

As the first equality involves the following relation between the conditional standard deviation matrices:

$$\boldsymbol{T}_k^{h'} = \boldsymbol{D}_k^{h,h'}\,\boldsymbol{T}_k^h, \tag{7}$$

we deduce then from the second one:

$$\left(\boldsymbol{T}_k^{h'}\right)^{-1}\boldsymbol{\mu}_k^{h'} = \left(\boldsymbol{T}_k^h\right)^{-1}\boldsymbol{\mu}_k^h. \tag{8}$$

This signifies that $\left(\boldsymbol{T}_k^h\right)^{-1}\boldsymbol{\mu}_k^h$ vectors do not depend on h and therefore that any conditional covariable has equal coefficients of variation across the populations.

3.3 Combining intra and interpopulation models

The most general model of simultaneous clustering is noted

$$\left(\pi^h, \boldsymbol{D}_k^{h,h'}, \boldsymbol{b}_k^{h,h'}; \pi_k, \boldsymbol{\Sigma}_k^h\right).$$

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, $\boldsymbol{b}_k^{h,h'}$ 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 $\left(\pi, \boldsymbol{D}^{h,h'}, \boldsymbol{b}^{h,h'}; \pi, \boldsymbol{\Sigma}^h\right)$ 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 combinings. 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 $\left(\pi^h\ ,\ .\ ,\ .\ ;\ \pi\ ,\ .\right)$ 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 (4)–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.

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.*

			Intrapopulation models			
			π		π_k	
	Interpopulation models		$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
π (π^h)	$\boldsymbol{I}$, $\alpha^{h,h'}\,\boldsymbol{I}$, $\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	• (.)	• (.)	• (•)	• (•)
		$\boldsymbol{b}^{h,h'}$	• (.)	• (.)	• (•)	• (•)
		$\boldsymbol{b}_k^{h,h'}$	○ (.)	• (.)	• (•)	• (•)
	$\alpha_k^{h,h'}\,\boldsymbol{I}$, $\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	. (.)	• (.)	. (.)	• (•)
		$\boldsymbol{b}_k^{h,h'}$	. (.)	• (.)	. (.)	• (•)

3.4 Requirements about identifiability

For a given permutation σ in S_H (symmetric group on $\{1,\ldots,H\}$), and another one τ in S_K, $\boldsymbol{\psi}_\tau^\sigma$ will denote the parameter $\boldsymbol{\psi}$, in which population labels have been permuted as σ, and component labels as τ, that is:

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

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 model is said to be identifiable when it satisfies:

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

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

Although most of the proposed models are identifiable, some of them, which we have to take care about, authorize different component label permutations depending on the population, and, as a consequence, some crossing of the link between Gaussian components. Let us assume for instance that each mixture has homoscedastic components ($\boldsymbol{\Sigma}_k^h = \boldsymbol{\Sigma}^h$) with equal mixing proportions ($\pi_k^h = 1/K$), that $\boldsymbol{D}_k^{h,h'}$ matrices in (4) only depend on population labels ($\boldsymbol{D}_k^{h,h'} = \boldsymbol{D}^{h,h'}$), and that $\boldsymbol{b}_k^{h,h'}$ vectors are free. It is easy to show in that case, that any component may be linked to any other one. This model is not identifiable.

Identifiable models among the allowed matchings of intra and interpopulation models are displayed in Table 1.

Assuming the data arise from a model which is not identifiable must not be rejected. It just leads to combinatorial possibilities in constituting groups of identical labels from the components C_k^h. In that case, simultaneous clustering provides a partition of the data, but the practitioner keeps some freedom in renumbering the components in each population.

3.5 Model selection

In a parametric model-based clustering context the BIC criterion (see [16] and see also [11] for a review) is commonly used, when the cluster number is known, in order to select a model within some model set, but also for assessing the number of clusters when this one is ignored [14] [9].
The BIC of a model is defined here by:

$$BIC = -\ell(\hat{\boldsymbol{\psi}};\mathbf{x}) + \frac{\nu}{2}\log(n), \tag{9}$$

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 ($n = \sum_{h=1}^H n^h$). Table 2 indicates the values of ν corresponding to the diverse intra and interpopulation model combinations. The model selected among competing ones corresponds to the smallest computed BIC value.

Let us remark that BIC appears also, here, as a natural way for selecting between independent clustering (Subsection 2.1) and simultaneous clustering (Subsection 2.2).

Table 2: *Dimension ν of the parameter $\boldsymbol{\psi}$ 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.*

		$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
$\boldsymbol{I}$	$\mathbf{0}$	$\beta+\gamma$	$\beta+K\gamma$
	$\boldsymbol{b}^{h,h'}$	$\beta+\gamma+d(H-1)$	$\beta+K\gamma+d(H-1)$
	$\boldsymbol{b}_k^{h,h'}$	$\beta+\gamma+dK(H-1)$	$\beta+K\gamma+dK(H-1)$
$\alpha^{h,h'}\ \boldsymbol{I}$	$\mathbf{0}$	$\beta+\gamma+(H-1)$	$\beta+K\gamma+(H-1)$
	$\boldsymbol{b}^{h,h'}$	$\beta+\gamma+(d+1)(H-1)$	$\beta+K\gamma+(d+1)(H-1)$
	$\boldsymbol{b}_k^{h,h'}$	$\beta+\gamma+(dK+1)(H-1)$	$\beta+K\gamma+(dK+1)(H-1)$
$\alpha_k^{h,h'}\ \boldsymbol{I}$	$\mathbf{0}$	.	$\beta+K\gamma+K(H-1)$
	$\boldsymbol{b}_k^{h,h'}$	.	$\beta+K\gamma+K(d+1)(H-1)$
$\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	$\beta+\gamma+d(H-1)$	$\beta+K\gamma+d(H-1)$
	$\boldsymbol{b}^{h,h'}$	$\beta+\gamma+2d(H-1)$	$\beta+K\gamma+2d(H-1)$
	$\boldsymbol{b}_k^{h,h'}$	$\beta+\gamma+d(K+1)(H-1)$	$\beta+K\gamma+d(K+1)(H-1)$
$\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	.	$\beta+K\gamma+dK(H-1)$
	$\boldsymbol{b}_k^{h,h'}$	.	$\beta+K\gamma+2dK(H-1)$

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. 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 (4) allows a new parameterization of the model at hand, which is useful and meaningful for estimating $\boldsymbol{\psi}$.
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 $\boldsymbol{\psi}$, the following vectors: $\boldsymbol{\theta}^1 = \boldsymbol{\psi}^1$, and for all $h \in \{2, \ldots, H\}$, $\boldsymbol{\theta}^h = \left[\left(\pi_k^h, \boldsymbol{D}_k^h, \boldsymbol{b}_k^h\right); k = 1, \ldots, K\right]$, where $\boldsymbol{D}_k^h = \boldsymbol{D}_k^{1,h}$ and $\boldsymbol{b}_k^h = \boldsymbol{b}_k^{1,h}$. Let us note the space described by the vector $\boldsymbol{\theta} = (\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^H)$ when $\boldsymbol{\psi}$ scans the parameter space . There exists a canonical bijective map between and . Thus $\boldsymbol{\theta}$ constitutes a new parameterization of the model at hand, and estimating $\boldsymbol{\psi}$ or $\boldsymbol{\theta}$ by maximizing their likelihood, respectively on or , is equivalent.

$\boldsymbol{\theta}^1$ appears to be a 'reference population parameter' whereas $(\boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^H)$ 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 P^1 population. Indeed the bijective correspondance between the parameter spaces and 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:

$$l_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{h=1}^{H} \sum_{i=1}^{n^h} \sum_{k=1}^{K} z_{i,k}^h \log\left(\pi_k^h \quad_d \left(\mathbf{x}_i^h; \boldsymbol{D}_k^h \boldsymbol{\mu}_k^1 + \boldsymbol{b}_k^h, \boldsymbol{D}_k^h \boldsymbol{\Sigma}_k^1 \boldsymbol{D}_k^h\right)\right), \qquad (10)$$

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 $GL_d(\mathbb{R})$ and $\boldsymbol{b}_k^1$ is the null vector of $\mathbb{R}^d$. But Dempster et al. [7] showed that an EM algorithm is not required 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 $E[l_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) | \boldsymbol{X} = \mathbf{x}]$ 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.
- GM-step: The conditional expectation of the completed log-likelihood, obtained by substituting $z_{i,k}^h$ for $t_{i,k}^h$ in (10), can be alternatively maximized with respect to the two following component sets of $\boldsymbol{\theta}$ parameter: $\{\pi_k^h, \boldsymbol{\mu}_k^1, \boldsymbol{\Sigma}_k^1\}$ and $\{\boldsymbol{D}_k^h, \boldsymbol{b}_k^h\}$ $(h = 1, \ldots, H)$. It provides the estimator $\boldsymbol{\theta}^+$ that is used as $\boldsymbol{\theta}$ at the next iteration of the current GM-step.

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

Let us detail now the GM-step since it depends on the intra and interpopulation model at hand.

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:

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

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

– *Vectors* $\boldsymbol{b}_k^h$

Noting $\mathbf{x}_k^h = (1/\hat{n}_k^h) \sum_{i=1}^{n^h} t_{i,k}^h \mathbf{x}_i^h$ 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 $\boldsymbol{b}_k^{h^+} = \mathbf{x}_k^h - \boldsymbol{D}_k^h \boldsymbol{\mu}_k^{1^+}$, and by:

$$\boldsymbol{b}_k^{h^+} = \left[\sum_{k=1}^{K} \hat{n}_k^h \left(\boldsymbol{D}_k^h \boldsymbol{\Sigma}_k^{1^+} \boldsymbol{D}_k^h\right)^{-1}\right]^{-1} \left[\sum_{k=1}^{K} \hat{n}_k^h \left(\boldsymbol{D}_k^h \boldsymbol{\Sigma}_k^{1^+} \boldsymbol{D}_k^h\right)^{-1} \left(\mathbf{x}_k^h - \boldsymbol{D}_k^h \boldsymbol{\mu}_k^{1^+}\right)\right], \tag{11}$$

when supposing they are equal.

– *Matrices* $\boldsymbol{D}_k^h$

When $\boldsymbol{D}_k^h$ $(k = 1, \ldots, K$ and $h = 2, \ldots, H)$ are some homothety matrices, that is when $\boldsymbol{D}_k^h = \alpha_k^h \boldsymbol{I}$ $(\alpha_k^h \in \mathbb{R}_*^+)$, or $\boldsymbol{D}_k^h = \alpha^h \boldsymbol{I}$ $(\alpha^h \in \mathbb{R}_*^+)$, according to their depending (or not depending) on the components, they are estimated respectively thanks to the two following formulas:

$$\alpha_k^{h^+} = \frac{-u_k^h + \sqrt{(u_k^h)^2 + 4d\hat{n}_k^h v_k^h}}{2d\hat{n}_k^h} \quad \text{or} \quad \alpha_k^{h^+} = \frac{-u^h + \sqrt{(u^h)^2 + 4d\hat{n}^h v^h}}{2dn^h},$$

where

– $u_k^h = \sum_{i=1}^{n^h} t_{i,k}^h \left(\mathbf{x}_i^h - \boldsymbol{b}_k^{h^+}\right)' \left(\boldsymbol{\Sigma}_k^{1^+}\right)^{-1} \boldsymbol{\mu}_k^{1^+}$ and $u^h = \sum_{k=1}^{K} u_k^h$,

– $v_k^h = \sum_{i=1}^{n^h} t_{i,k}^h \left(\mathbf{x}_i^h - \boldsymbol{b}_k^{h^+}\right)' \left(\boldsymbol{\Sigma}_k^{1^+}\right)^{-1} \left(\mathbf{x}_i^h - \boldsymbol{b}_k^{h^+}\right)$ and $v^h = \sum_{k=1}^{K} v_k^h$.

In the other more general cases, $\boldsymbol{D}_k^h$ matrices can not be estimated explicitly. Nevertheless, as the conditional expectation of $\boldsymbol{\theta}$ completed log-likelihood 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.

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, first, identifiability of the model is preserved, and secondly the conditional expectation of $\boldsymbol{\theta}$ completed log-likelihood $E[l_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z})|\boldsymbol{X} = \mathbf{x}]$, keeps on being concave with respect to $(\boldsymbol{D}_k^h)^{-1}$ on the parameter space . 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, $\boldsymbol{\psi}$ 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 some of the models of simultaneous clustering allow an alternative sequential estimation which does not maximize $\boldsymbol{\psi}$ likelihood in general, but which is simpler than the previous GEM algorithm and which leads also to consistent estimates.

When the interpopulation model is $(\pi, \boldsymbol{D}^{h,h'}, \boldsymbol{b}^{h,h'})$ (or one of its parsimonious models obtained by assuming $\boldsymbol{D}^{h,h'} = \alpha^{h,h'}\boldsymbol{I}$, $\boldsymbol{D}^{h,h'} = \boldsymbol{I}$ or $\boldsymbol{b}^{h,h'} = \mathbf{0}$) the conditional link (3) stretches over unconditional populations:

$$\boldsymbol{X}^{h'} \sim \boldsymbol{D}^{h,h'}\boldsymbol{X}^h + \boldsymbol{b}^{h,h'}. \tag{12}$$

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 is 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 a least square methodology leading to explicit estimates given in Table 3.

Table 3: *Link parameter least-square estimates in the sequential estimation method.* $\mathbf{x}^h = (1/n^h)\sum_{i=1}^{n^h}\mathbf{x}_i^h$ *and* $\hat{\boldsymbol{S}}^h = (1/n^h)\sum_{i=1}^{n^h}(\mathbf{x}_i^h - \mathbf{x}^h)(\mathbf{x}_i^h - \mathbf{x}^h)'$ *denote respectively the empirical center and the empirical covariance matrix of the whole population* P^h.

Interpopulation model	$\hat{\boldsymbol{D}}^h$	$\hat{\boldsymbol{b}}^h$
$(\boldsymbol{I}, \boldsymbol{b}^{h,h'})$	$\boldsymbol{I}$	$\hat{\boldsymbol{b}}^h = \mathbf{x}^h - \mathbf{x}^1$
$(\alpha^{h,h'}\boldsymbol{I}, \mathbf{0})$	$\dfrac{(\mathbf{x}^h)'(\mathbf{x}^1)}{(\mathbf{x}^1)'(\mathbf{x}^1)}\boldsymbol{I}$	$\mathbf{0}$
$(\alpha^{h,h'}\boldsymbol{I}, \boldsymbol{b}^{h,h'})$	$\hat{\alpha}^{1,h} = \left[\mathrm{tr}\left(\hat{\boldsymbol{S}}^1\hat{\boldsymbol{S}}^h\right)/\mathrm{tr}\left((\hat{\boldsymbol{S}}^1)^2\right)\right]^{1/2}$	$\hat{\boldsymbol{b}}^h = \mathbf{x}^h - \hat{\alpha}^{1,h}\mathbf{x}^1$
$(\boldsymbol{D}^{h,h'}, \mathbf{0})$	$\{\hat{\boldsymbol{D}}^h\}_{jj} = \{\mathbf{x}^h\}_j/\{\mathbf{x}^1\}_j$	$\mathbf{0}$
$(\boldsymbol{D}^{h,h'}, \boldsymbol{b}^{h,h'})$	$\left(\mathrm{diag}\hat{\boldsymbol{S}}^h\right)^{1/2}\left(\mathrm{diag}\hat{\boldsymbol{S}}^1\right)^{-1/2}$	$\hat{\boldsymbol{b}}^h = \mathbf{x}^h - \hat{\boldsymbol{D}}^h\mathbf{x}^1$

Since in case of the most complex model considered in this subsection, $(\pi, \boldsymbol{D}^{h,h'}, \boldsymbol{b}^{h,h'})$, the least square estimator of $\boldsymbol{D}^h$ parameter requires a numerical procedure, we give an alternative but explicit and consistent estimator of $\boldsymbol{D}^h$ based on the relation $[\boldsymbol{S}^h = \boldsymbol{D}^h\boldsymbol{S}^1\boldsymbol{D}^h] \Rightarrow [(\mathrm{diag}\ \boldsymbol{S}^h) = \boldsymbol{D}^h(\mathrm{diag}\ \boldsymbol{S}^1)\boldsymbol{D}^h]$, where $\boldsymbol{S}^h$ denotes the covariance matrix of the whole population P^h.

The second step of the strategy is the following: As all the transformed data points $(\boldsymbol{D}^h)^{-1}(\mathbf{x}_i^h - \boldsymbol{b}^h)$ ($h = 1, \ldots, H$, $k = 1, \ldots, K$) are assumed to arise inde-

pendently from P^1 population, a simple and traditional EM algorithm devoted to Gaussian mixture estimation, can be involved. Softwares as MIXMOD [3] are now available for practitioners to perform that estimation.

Remark: That alternative estimation procedure still consists of a ML estimate of $\boldsymbol{\psi}$ parameter but now under the constraint of the previously estimated and plugged in link parameter. Although estimators given in Table 3 depend on which sample holds the label 1, the constraint set on $\boldsymbol{\psi}$ likelihood does not depend on this population label choice in case of interpopulation models $(\pi, \boldsymbol{D}^{h,h'}, \boldsymbol{b}^{h,h'})$, $(\pi, \boldsymbol{D}^{h,h'}, \mathbf{0})$ or $(\pi, \boldsymbol{I}, \boldsymbol{b}^{h,h'})$. Indeed for these models, the link parameter owns some symmetry and transitivity properties which are also satisfied by the corresponding estimators of Table 3. In case of both other interpopulation models the symmetry and transitivity properties of the link parameter are no more satisfied by the estimators of Table 3 and then the sequential estimation does depend on the population label choice. Nevertheless next section will suggest that, in these cases, sequential estimates are still close to ML estimates obtained by the previous GEM algorithm (Subsections 4.3 and 4.4).

5 A biological example

5.1 The data

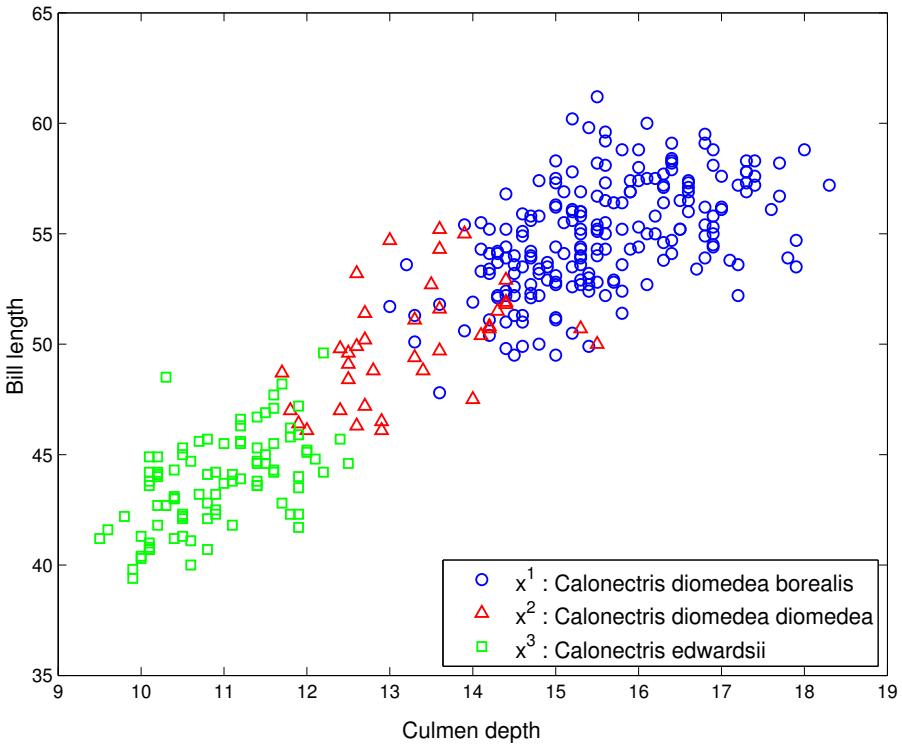
In [17] three seabird subspecies ($H = 3$) of Shearwaters, differing over their geograpical range, are described. *Borealis* (sample $\mathbf{x}^1$, size $n^1 = 206$ individuals, 45% female) are living in the Atlantic Islands (Azores, Canaries, etc.), *Diomedea* (sample $\mathbf{x}^2$, size $n^2 = 38$ individuals, 58% female), in Mediterranean Islands (Balearics, Corsica, etc.), and *Edwardsii* (sample $\mathbf{x}^3$, size $n^3 = 92$ individuals, 52% female), 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 retrieve the sex of the birds ($K = 2$).

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. We aim to distinguish males and females for each of them and, so, three standard Gaussian model-based clusterings should be considered. However, let us remark that the researched partition (males, females) has the same meaning in each sample, and the three samples are described by the same five morphological features. Then the data set is suitable for some simultaneous clustering process.

5.2 Partitioning when the cluster number is known

We applied on the three seabird samples each of the 66 allowed models of simultaneaous clustering displayed in Table 1. Since the birds must be clustered

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



according to their sex, the number of groups is set to 2. The clustering procedure consists in estimating the parameter of each model by a GEM algorithm (5 trials for each procedure, 500 iterations and 5 directional maximizations at each GM step (see Subsection 4.2)) and selecting the model which gives the smallest BIC value. Results are constituted by the empirical error rate (obtained thanks to the known true partition) and by the BIC value of each model.

BIC criterion allows also to compare the simultaneous clustering procedure to the independent one. Indeed, one can also estimate the parameter $\boldsymbol{\psi}$ assuming that the stochastic link (3) does not hold in the three seabird populations and compute then the BIC value of the model so inferred. In Table 4, the BIC values obtained by the independent clustering method, have been computed according to (9). Comparing them with BIC obtained from simultaneous clustering, leads to choose the simultaneous clustering method.

BIC criterion and error rate are quite different statistics. BIC translates in some particular sense the adequacy of a model to the data, whereas the error

rate translates the overlapping of components in a mixture model. Some model well adapted to the data may be quite inefficient to determine well-separated clusters and conversely. Table 4 shows that BIC and error rate seem to behave, here, in the same manner. The model selected by BIC, $(\pi, \boldsymbol{D}^{h,h'}, \mathbf{0}; \pi, \boldsymbol{\Sigma}^h)$, corresponds also to the smallest error rate (10.42%). According to this model, $\boldsymbol{b}_k^{h,h'}$ vectors are all null. Biernacki et al. performed in [2] some test on the empirical covariance matrices $\hat{\boldsymbol{\Sigma}}_k^h$ estimated from the sexed samples, in order to corroborate this hypothesis. That model involves also that the mixture components are homoscedastic. Some cross-validation criterion can show that males and females should constitute some homoscedastic components, at less among *Borealis* and *Diomeda* (see [2]).

Remark: Table 5 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 four of them which correspond to a parameter located on a degeneracy path of the likelihood–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 provide for less 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 some peculiar models of simultaneous clustering.

5.3 The general situation: Partitioning when the cluster number is unknown

Experiments exhibited in the previous paragraph were extended to less or more than two clusters. We considered successively that bird species were partitioned into one (no structure), two, three or four underlying groups and results are respectively displayed in Tab. 6, 4, 7 and 8. Obviously no empirical error rate is displayed when $K \neq 2$.

When the cluster number was set equal to 2, the best model inferred by simultaneous clustering was better than the best model obtained in independent clustering. By comparing the best BIC values obtained in both methods, Table 9 confirms when $K = 1, 3,$ or 4, that advantage of the simultaneous clustering method on the independent one. Indeed, whatever is K among $\{1, 2, 3, 4\}$, the best model is always obtained by simultaneous clustering, which shows how relevant may be the specific parsimony of simultaneous clustering models.

According to Table 9, selecting the cluster number thanks to the best BIC values obtained by independent clustering leads to an error (indeed it corresponds to $K = 1$), whereas the best BIC obtained in simultaneous clustering selects the cluster number which is researched ($K = 2$).

5.4 Some robustness study of the simultaneous clustering method: Relaxing the exact variable concordance

Simultaneous clustering relies, among other things, on the assumption that samples to be classified are described by variables of identical meaning. However in many concrete situations descriptors do not have exactly the same sense in some sample or other. The parsimonious models of simultaneous clustering are still relevant in those cases if it remains realistic to suppose that conditional correlations are invariant through the populations for some variable permutation within each population. Then the practitioner will have in that relaxed context to propose, if possible, a realistic correspondance between all involved population variables.

The following example shows that the models of simultaneous clustering may still be of interest when relaxing the covariable concordance assumption.

We dispose of another bird sample $\mathbf{x}^4$ (size $n^4 = 22$ individuals, 54% female) [5] composed of White-throated Dippers (*Cinclus cinclus cinclus*) living in Lorraine (France), which size is close to Calonectris diomedea diomeda sample's one. Birds of $\mathbf{x}^4$ are described by their tarsus and the length of their folded wing, that is two variables close in meaning to the couple tarsus-wing length which describes among others $\mathbf{x}^2$ sample.

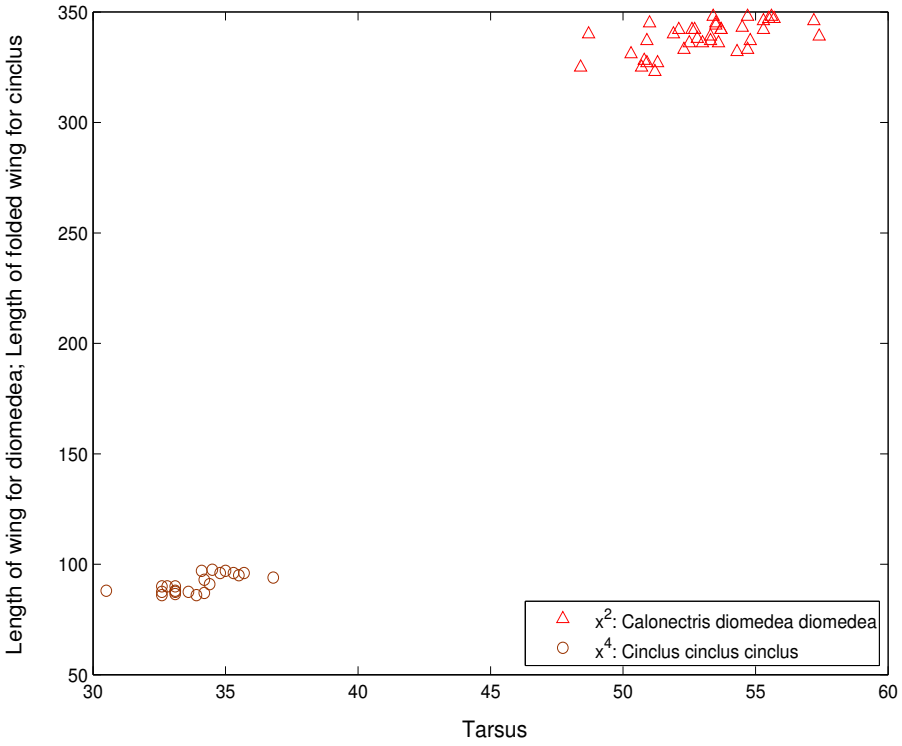
We aim to classify simultaneously the 60 birds of $\mathbf{x}^2$ and $\mathbf{x}^4$ (see Figure 2) according to their sex and then the cluster number is set to 2. Table 10 displays BIC values of the 66 allowed combinations of intra and interpopulation models of simultaneous clustering, BIC values of the 4 parsimonious models of independent clustering, and the corresponding error rates obtained thanks to the known true partitions.

In that relaxed context, the best BIC value (309.8) is still obtained from the simultaneous clustering method, as the second and the third best one (respectively 309.9 and 310.1), and they all correspond to a model in which $\boldsymbol{D}_k^{h,h'}$ matrices are equal among males and females and $\boldsymbol{b}_k^{h,h'}$ vectors also. Moreover these models provide some error rates (respectively 23.33%, 30% and 18.33%) which are often better than the error rate corresponding to the best model of independent clustering (25.00%).

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

Figure 2: *Two bird samples described by variables close in meaning.*



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 biologist: Explicit least square estimates followed by a standard EM algorithm for Gaussian mixtures.

We showed the efficiency of the models on biological data which true partition was known. Experiments revealed that for some given number of clusters, the model inferred from simultaneous clustering was better than the model estimated by several independent clustering methods. On the other hand, feigning to ignore the true cluster number, the models available in simultaneous clustering did select it naturally. We noticed at last that the so-called simultaneous clustering method had some kind of robustness to one of its main assumptions relaxation that is to say the exact concordance of population descriptors.

If the subspecies of each Shearwater that we classified in Subsection 5.2 were unknown and had to be determined so as its sex, our model of simultaneous clustering could easily be extended to hierarchical mixtures for nested data structures [18]–level-1 groups consisting on the bird sex and level-2 ones

on subspecies–by considering some additional latent variable in the model, indicating each bird subspecies.

Gaussian mixtures are widespread in model-based clustering but the literature mentions many other distributions useful in that context. Mixtures of factor analyzers are used in order to assess groups in high-dimensional data sets [13], mixtures of Student distributions are applied when the data include outliers [13]. Some combined use of both factor analyzers and t-distributions seems to give interesting results in microarray gene-expression data clustering [12]. Studying the possibility and the efficiency of performing some simultaneous clustering method based on t-mixtures or factor analyzer mixtures, in those situations, would be of interest.

The simultaneous clustering method relies in this work on an affine stochastic link between the components of diverse mixtures. Some other kinds of link can be envisaged which should improve–if they translate some realistic constraint on the populations–the standard method consisting on several independent sample clusterings. For example some close overlappings of the groups within the diverse samples to be classified should make as difficult every sample clustering. Formalizing that information by supposing all mixtures to have equal global component entropies (or identical error rates) and setting this as a constraint on the model should improve the sample classification insofar as this constraint is close to truth.

Acknowledgements

The authors 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 also thank Sandra McJannett and Anne-Marie Pollaud-Dulian for their advice.

References

[1] Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

[2] Biernacki, C., Beninel, F. and Bretagnolle, V. (2002). Generalized discriminant rule when training population and test population differ on their descriptive parameters, *Biometrics*, **49**, 803-821.

[3] Biernacki, C., Celeux, G., Govaert, G. and Langrognet, F. (2006). Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. *Computational Statistics and Data Analysis*, **51**, 2, 587-600.

[4] Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models, *Pattern Recognition*, **28(5)**, 781-793.

[5] D'Amico, F., Lalanne, Y., O'Halloran, J. and Smiddy, P. (2009). Personal communication.

[6] De Meyer, B., Roynette, B., Vallois, P. and Yor, M. (2000). On independent times and positions for Brownian motion. Technical Report 1, Les prépublications de l'Institut Elie Cartan, Institut Elie Cartan, Vandoeuvre lès Nancy, France.

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

[8] Flury, B.N. (1983). Common principal components in k groups, *Journal of the American Statistical Association*, **79**, 892–898.

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

[10] Gower, J.C. (1975). Generalized Procrustes Analysis, *Psychometrika*, **40**, 33–51.

[11] Lebarbier, E. et Mary-Huard, T. (2006). Le critère BIC, fondements théoriques et interprétation, *Journal de la Société Française de Statistique*, **1**, 39-57.

[12] McLachlan, G.J., Bean, R.W. and Ben-Tovim Jones L. (2006). Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution, *Computational Statistics & Data Analysis*, **51**, 5327–5338.

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

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

[15] Schork, N.J. and Thiel, B. (1996). Mixture distributions in human genetics. *Statistical Methods in Medical Research*, **39**, 155–178.

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

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

[18] Vermunt, J.K. and Magidson, J. (2005). Hierarchical mixture models for nested data structures, *Classification: The Ubiquitous Challenge*, Weihs, C. and Gaul, W., eds., Springer, Heidelberg, 176–183.

Table 4: *BIC value and (error rate) in simultaneous (full ML estimates) and independent clustering (2 groups) of Shearwaters.*

			π		π_k	
			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}^h_k$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}^h_k$
π	$\boldsymbol{I}$	$\mathbf{0}$	4392.9 (43.45)	4392.5 (44.94)	4371.8 (45.24)	4383.6 (43.45)
		$\boldsymbol{b}^{h,h'}$	4064.5 (11.61)	4089.8 (11.61)	4067.4 (11.61)	4091.2 (15.77)
		$\boldsymbol{b}^{h,h'}_k$	4084.4 (12.20)	4110.1 (13.10)	4080.0 (41.96)	4107.4 (26.49)
	$\alpha^{h,h'}\ \boldsymbol{I}$	$\mathbf{0}$	4254.0 (33.04)	4279.7 (29.17)	4246.2 (42.56)	4276.0 (41.37)
		$\boldsymbol{b}^{h,h'}$	4056.8 (11.61)	4081.7 (11.61)	4059.7 (11.01)	4083.7 (14.88)
		$\boldsymbol{b}^{h,h'}_k$	4079.6 (11.61)	4105.2 (11.90)	4079.9 (40.77)	4095.8 (45.83)
	$\alpha^{h,h'}_k\ \boldsymbol{I}$	$\mathbf{0}$	.	4282.9 (32.14)	.	4279.4 (38.69)
		$\boldsymbol{b}^{h,h'}_k$	.	4110.4 (12.50)	.	4110.4 (16.07)
	$\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	**4047.0** (**10.42**)	4071.9 (11.61)	4049.7 (11.31)	4073.9 (11.88)
		$\boldsymbol{b}^{h,h'}$	4071.8 (10.71)	4096.9 (12.20)	4074.7 (10.71)	4099.3 (14.58)
		$\boldsymbol{b}^{h,h'}_k$	4094.9 (33.33)	4122.2 (11.31)	4101.9 (41.96)	4122.7 (15.77)
	$\boldsymbol{D}^{h,h'}_k$	$\mathbf{0}$	.	4097.5 (11.90)	.	4099.2 (14.88)
		$\boldsymbol{b}^{h,h'}_k$	.	4154.5 (38.39)	.	4147.9 (25.29)
π^h	$\boldsymbol{I}$	$\mathbf{0}$	.	.	4194.9 (43.45)	4186.1 (45.54)
		$\boldsymbol{b}^{h,h'}$	.	.	4058.0 (40.48)	4088.5 (25.89)
		$\boldsymbol{b}^{h,h'}_k$	.	.	4084.4 (41.96)	4110.5 (44.05)
	$\alpha^{h,h'}\ \boldsymbol{I}$	$\mathbf{0}$	.	.	4095.2 (47.32)	4123.7 (47.32)
		$\boldsymbol{b}^{h,h'}$	.	.	4059.4 (40.48)	4090.1 (26.19)
		$\boldsymbol{b}^{h,h'}_k$	.	.	4081.5 (41.96)	4102.9 (45.83)
	$\alpha^{h,h'}_k\ \boldsymbol{I}$	$\mathbf{0}$	.	.	.	4129.5 (47.32)
		$\boldsymbol{b}^{h,h'}_k$	.	.	.	4107.8 (45.83)
	$\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	.	.	4055.5 (11.01)	4079.5 (15.18)
		$\boldsymbol{b}^{h,h'}$	.	.	4079.9 (39.88)	4107.8 (40.18)
		$\boldsymbol{b}^{h,h'}_k$	.	.	4107.6 (42.86)	4128.5 (15.18)
	$\boldsymbol{D}^{h,h'}_k$	$\mathbf{0}$	.	.	.	4101.8 (45.24)
		$\boldsymbol{b}^{h,h'}_k$	.	.	.	4153.6 (16.37)
Independent			**4139.8** (**12.50**)	4218.2 (38.39)	4143.0 (29.17)	4219.7 (40.18)

Table 5: *Sequential estimation: BIC value and (error rate) in simultaneous and independent clustering (2 groups) of Shearwaters.*

			π		π_k	
			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}^h_k$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}^h_k$
π	$\boldsymbol{I}$	$\mathbf{0}$	4392.9 (43.45)	4392.5 (44.94)	4371.8 (45.24)	4383.6 (43.45)
		$\boldsymbol{b}^{h,h'}$	4064.6 (11.61)	4090.9 (11.90)	4205.6 (37.20)	4337.0 (45.83)
	$\alpha^{h,h'}\ \boldsymbol{I}$	$\mathbf{0}$	4259.5 (32.74)	4283.5 (29.46)	4247.6 (43.45)	4278.1 (42.26)
		$\boldsymbol{b}^{h,h'}$	4057.0 (11.31)	4082.4 (11.61)	4059.6 (36.01)	4068.7 (46.13)
	$\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	**4047.0** (10.71)	4072.0 (11.90)	4049.0 (35.11)	4074.2 (14.28)
		$\boldsymbol{b}^{h,h'}$	4072.4 (**10.42**)	4097.5 (11.90)	4074.3 (34.52)	4099.7 (14.28)

Table 6: *BIC value in simultaneous (full ML estimates) and independent clustering (1 group) of Shearwaters.*

$\boldsymbol{I}$	$\mathbf{0}$	4472.0
	$\boldsymbol{b}^{h,h'}$, $\boldsymbol{b}^{h,h'}_k$	4061.8
$\alpha^{h,h'}\ \boldsymbol{I}$, $\alpha^{h,h'}_k\ \boldsymbol{I}$	$\mathbf{0}$	4246.4
	$\boldsymbol{b}^{h,h'}$, $\boldsymbol{b}^{h,h'}_k$	4057.3
$\boldsymbol{D}^{h,h'}$, $\boldsymbol{D}^{h,h'}_k$	$\mathbf{0}$	**4047.8**
	$\boldsymbol{b}^{h,h'}$, $\boldsymbol{b}^{h,h'}_k$	4073.3
Independent		4102.6

Table 7: *BIC value in simultaneous (full ML estimates) and independent clustering (3 groups) of Shearwaters.*

			π		π_k	
			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
π	$\boldsymbol{I}$	$\mathbf{0}$	4372.7	4405.3	4349.3	4409.9
		$\boldsymbol{b}^{h,h'}$	4074.5	4125.5	4067.9	4129.2
		$\boldsymbol{b}_k^{h,h'}$	4112.9	4167.2	4110.0	4160.1
	$\alpha^{h,h'}\,\boldsymbol{I}$	$\mathbf{0}$	4253.2	4317.0	4249.9	4307.6
		$\boldsymbol{b}^{h,h'}$	4065.7	4120.1	4060.8	4119.8
		$\boldsymbol{b}_k^{h,h'}$	4110.0	4161.8	4108.1	4157.0
	$\alpha_k^{h,h'}\,\boldsymbol{I}$	$\mathbf{0}$	.	4322.1	.	4311.9
		$\boldsymbol{b}_k^{h,h'}$	.	4174.2	.	4151.5
	$\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	4053.8	4105.5	**4051.0**	4103.2
		$\boldsymbol{b}^{h,h'}$	4078.7	4132.2	4076.7	4137.9
		$\boldsymbol{b}_k^{h,h'}$	4129.8	4181.6	4126.2	4173.5
	$\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	.	4153.3	.	4155.9
		$\boldsymbol{b}_k^{h,h'}$	.	4232.4	.	4216.6
π^h	$\boldsymbol{I}$	$\mathbf{0}$	.	.	4079.2	4137.7
		$\boldsymbol{b}^{h,h'}$	.	.	4070.2	4129.0
		$\boldsymbol{b}_k^{h,h'}$	.	.	4118.1	4159.8
	$\alpha^{h,h'}\,\boldsymbol{I}$	$\mathbf{0}$	.	.	4073.8	4143.3
		$\boldsymbol{b}^{h,h'}$	.	.	4068.6	4128.3
		$\boldsymbol{b}_k^{h,h'}$	.	.	4115.9	4159.5
	$\alpha_k^{h,h'}\,\boldsymbol{I}$	$\mathbf{0}$	.	.	.	4155.2
		$\boldsymbol{b}_k^{h,h'}$	.	.	.	4173.8
	$\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	.	.	4062.4	4119.9
		$\boldsymbol{b}^{h,h'}$	.	.	4089.2	4141.2
		$\boldsymbol{b}_k^{h,h'}$	.	.	4133.8	4174.4
	$\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	.	.	.	4153.9
		$\boldsymbol{b}_k^{h,h'}$	.	.	.	4236.3
Independent			**4137.6**	4289.3	4148.0	4291.3

Table 8: *BIC value in simultaneous (full ML estimates) and independent clustering (4 groups) of Shearwaters.*

			π		π_k	
			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}^h_k$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}^h_k$
π	$\boldsymbol{I}$	$\mathbf{0}$	4357.8	4429.2	4341.7	4444.6
		$\boldsymbol{b}^{h,h'}$	4075.9	4157.5	4079.9	4162.0
		$\boldsymbol{b}^{h,h'}_k$	4136.7	4225.1	4138.3	4219.0
	$\alpha^{h,h'}\ \boldsymbol{I}$	$\mathbf{0}$	4259.5	4351.3	4263.0	4354.0
		$\boldsymbol{b}^{h,h'}$	4067.4	4154.3	4071.3	4160.4
		$\boldsymbol{b}^{h,h'}_k$	4135.8	4222.3	4137.9	4219.0
	$\alpha^{h,h'}_k\ \boldsymbol{I}$	$\mathbf{0}$	.	4360.4	.	4362.6
		$\boldsymbol{b}^{h,h'}_k$	.	4238.2	.	4231.0
	$\boldsymbol{D}^{h,h'}$	0	**4055.7**	4147.6	4058.7	4151.7
		$\boldsymbol{b}^{h,h'}$	4082.4	4169.3	4085.4	4172.5
		$\boldsymbol{b}^{h,h'}_k$	4153.7	4243.5	4155.0	4229.0
	$\boldsymbol{D}^{h,h'}_k$	$\mathbf{0}$	.	4213.9	.	4207.9
		$\boldsymbol{b}^{h,h'}_k$	.	4320.8	.	4304.9
π^h	$\boldsymbol{I}$	$\mathbf{0}$	.	.	4078.7	4169.9
		$\boldsymbol{b}^{h,h'}$	.	.	4084.2	4165.9
		$\boldsymbol{b}^{h,h'}_k$	.	.	4151.5	4220.9
	$\alpha^{h,h'}\ \boldsymbol{I}$	$\mathbf{0}$	.	.	4078.7	4175.7
		$\boldsymbol{b}^{h,h'}$	.	.	4087.3	4163.7
		$\boldsymbol{b}^{h,h'}_k$	.	.	4151.9	4224.5
	$\alpha^{h,h'}_k\ \boldsymbol{I}$	$\mathbf{0}$	.	.	.	4193.2
		$\boldsymbol{b}^{h,h'}_k$	.	.	.	4235.8
	$\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	.	.	4073.1	4155.2
		$\boldsymbol{b}^{h,h'}$	.	.	4107.4	4175.0
		$\boldsymbol{b}^{h,h'}_k$	.	.	4168.7	4243.8
	$\boldsymbol{D}^{h,h'}_k$	$\mathbf{0}$	.	.	.	4228.5
		$\boldsymbol{b}^{h,h'}_k$	.	.	.	4318.1
Independent			**4159.6**	4363.4	4171.8	4359.3

Table 9: *Best BIC values obtained in simultaneous (full ML estimates) and independent clustering of Cory's Shearwaters with different number of clusters.*

Cluster Number	1	2	3	4
Simultaneous Clustering	4047.8	**4047.0**	4051.0	4055.7
Independent Clustering	4102.6	4139.8	4137.7	4159.6

Table 10: *BIC value and (error rate) obtained in simultaneous (full ML estimates) and independent clustering (2 groups) of two bird samples in some case of non concordant descriptors.*

			π		π_k	
			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
π	$\boldsymbol{I}$	$\mathbf{0}$	357.3 (46.67)	356.2 (46.67)	357.2 (46.67)	356.1 (46.67)
		$\boldsymbol{b}^{h,h'}$	318.9 (28.33)	321.7 (41.67)	318.9 (48.33)	328.8 (41.67)
		$\boldsymbol{b}_k^{h,h'}$	316.5 (30.00)	320.2 (45.00)	317.8 (45.00)	318.2 (21.67)
	$\alpha^{h,h'}\boldsymbol{I}$	$\mathbf{0}$	352.4 (46.67)	358.2 (46.67)	352.3 (46.67)	363.1 (18.33)
		$\boldsymbol{b}^{h,h'}$	**309.8** (23.33)	315.2 (25.00)	313.1 (33.33)	310.1 (**18.33**)
		$\boldsymbol{b}_k^{h,h'}$	311.5 (25.00)	315.6 (41.67)	311.0 (38.38)	312.0 (36.67)
	$\alpha_k^{h,h'}\boldsymbol{I}$	$\mathbf{0}$	.	468.8 (25.00)	.	465.1 (20.00)
		$\boldsymbol{b}_k^{h,h'}$	.	318.1 (43.33)	.	320.0 (41.67)
	$\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	319.0 (28.33)	322.7 (30.00)	318.8 (28.33)	316.9 (30.00)
		$\boldsymbol{b}^{h,h'}$	311.5 (23.33)	316.6 (23.33)	312.6 (28.33)	314.3 (18.33)
		$\boldsymbol{b}_k^{h,h'}$	313.6 (23.33)	318.4 (41.67)	312.8 (38.33)	314.4 (36.67)
	$\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	.	313.4 (20.00)	.	310.2 (40.00)
		$\boldsymbol{b}_k^{h,h'}$	.	320.8 (18.33)	.	314.5 (18.33)
π^h	$\boldsymbol{I}$	$\mathbf{0}$	.	.	319.8 (46.67)	318.7 (46.67)
		$\boldsymbol{b}^{h,h'}$	.	.	323.9 (43.33)	316.1 (21.67)
		$\boldsymbol{b}_k^{h,h'}$	.	.	319.8 (43.33)	318.6 (21.67)
	$\alpha^{h,h'}\ \boldsymbol{I}$	$\mathbf{0}$	.	.	314.9 (46.67)	320.7 (46.67)
		$\boldsymbol{b}^{h,h'}$	.	.	316.7 (43.33)	317.5 (21.67)
		$\boldsymbol{b}_k^{h,h'}$	.	.	312.4 (40.00)	313.2 (36.67)
	$\alpha_k^{h,h'}\ \boldsymbol{I}$	$\mathbf{0}$	.	.	.	447.2 (30.00)
		$\boldsymbol{b}_k^{h,h'}$	.	.	.	317.5 (28.33)
	$\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	.	.	311.9 (28.33)	309.9 (30.00)
		$\boldsymbol{b}^{h,h'}$	.	.	317.2 (43.33)	324.1 (41.67)
		$\boldsymbol{b}_k^{h,h'}$	.	.	314.5 (26.67)	315.1 (36.67)
	$\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	.	.	.	310.4 (40.00)
		$\boldsymbol{b}_k^{h,h'}$	.	.	.	314.9 (21.67)
	Independent		**310.9** (25.00)	315.8 (23.33)	313.9 (28.33)	318.2 (**20.00**)