Université Lille1
Sciences et Technologies

N° d'ordre: 40552

THÈSE

Présentée pour obtenir

LE GRADE DE DOCTEUR EN
MATHÉMATIQUES DE L'UNIVERSITÉ
LILLE I

Spécialité: Mathématiques Appliquées

par

Alexandre LOURME

Contribution à la Classification par Modèles de Mélange et Classification Simultanée d'Echantillons d'Origines Multiples

Soutenue le 17 Juin 2011 devant la Commission d'examen:

M.	Christophe BIERNACKI	(Directeur de thèse)
M.	Gérard GOVAERT	(Rapporteur)
M.	Joseph NGATCHOU-WANDJI	(Rapporteur)
M.	Jérôme SARACCO	(Rapporteur)

Thése préparée au
Département de Mathématiques Appliquées
Laboratoire Paul Painlevé (UMR CNRS 8524)
Université Lille 1
59 655 Villeneuve d'Ascq CEDEX

Résumé

L'usage de modèles de mélange est devenu courant en classification automatique. Contrairement aux méthodes de partitionnement traditionnelles, la classification à base de mélanges permet à la fois (i) de déterminer sans ambiguïté la structure d'un jeu de données, en définissant de façon rigoureuse la notion de sous-groupe homogène et (ii) de proposer une interprétation significative de la partition inférée.
Dans la première partie de ce travail nous passons en revue la classification par modèle de mélange. En particulier nous décrivons une famille de mélanges gaussiens usuels basés sur une décomposition spectrale des matrices de covariances et dont la parcimonie porte sur des paramètres géométriques. Ces modèles permettent une réduction graduelle et interprétable de la dimension du paramètre ; ils sont d'un usage très large ; mais ils souffrent de défauts rédhibitoires.
Nous leur opposons une nouvelle famille de mélanges gaussiens dits RTV, basés sur une décomposition variance-corrélation des matrices de covariances, et dont la parcimonie porte sur des paramètres statistiques. Les modèles RTV remédient à l'ensemble des inconvénients liés aux modèles géométriques. Ils sont stables par modification des unités de mesure. Une telle modification n'altère pas le choix de modèle par des critères de vraisemblance classiques comme AIC, BIC ou ICL. Les modèles RTV sont également stables par projection dans n'importe quel sous-espace canonique ; cela permet de les représenter de façon fidèle en dimension réduite.

Dans la seconde partie de ce travail nous présentons la classification simultanée. Cette procédure de partitionnement s'inscrit dans la tradition de nombreuses méthodes statistiques, dédiées initialement à un échantillon puis étendues à plusieurs échantillons. Nous montrons d'abord que la classification d'un échantillon revient très souvent au partitionnement de plusieurs échantillons ; puis nous proposons d'établir un lien entre la population d'origine des différents échantillons. Ce lien, dont la nature varie selon le contexte, a toujours pour vocation de formaliser de façon réaliste une information commune aux données à classifier.
Lorsque les échantillons sont décrits par des variables de même signification et que l'on cherche le même nombre de groupes dans chacun d'eux, nous établissons un lien stochastique entre populations conditionnelles. Cette forme de lien est envisagée d'abord dans un cadre usuel de mélanges gaussiens ; nous l'étendons ensuite aux mélanges de Student pour la classification de données bruitées ; nous évoquons enfin son intérêt dans les mélanges de Factor Analyzers pour classifier de données en grande dimension.
Lorsque les variables sont différentes mais sémantiquement proches d'un échantillon à l'autre, il se peut que leur pouvoir discriminant soit similaire et que l'imbrication des données conditionnelles soit comparable. Nous envisageons des mélanges spécifiques à ce contexte, liés par un chevauchement homogène de leurs composantes : nous supposons d'abord que l'erreur de Bayes des différents mélange est identique, puis que l'entropie globale de leurs composantes est invariante.
Sur de nombreux exemples de données réelles, nous montrons que l'hypothèse d'un lien entre populations, qu'il soit conditionnel ou global, permet très souvent d'améliorer l'adéquation du modèle, la qualité de la partition estimée, et que cette hypothèse donne lieu à une interprétation convaincante de la structure inférée.

Mots-clefs : classification non-supervisée, mélanges parcimonieux, corrélations, algorithme EM, lien entre populations, chevauchement des classes.

Contribution to Model-Based Clustering and Simultaneous Clustering of Samples Arising from Multiple Origins

Abstract

Nowadays mixture models are a useful tool to determine an underlying structure within data. Unlike traditional partitioning methods, mixture model-based clustering enables both to (i) draw unambiguous classes by defining rigorously what are homogeneous subgroups within some dataset and (ii) provide a significant interpretation of the estimated partition.
In the first part of this work we review the mixture model-based clustering method. We describe in particular, a family of common Gaussian mixtures based on a spectral decomposition of the covariance matrices. The parsimony of these models focuses on geometric parameters. They allow a gradual and meaningful reduction of the parameter size, they are broadly used, but they suffer from crippling flaws.
Then we display new Gaussian mixtures called RTV, based on a variance-correlation decomposition of the covariance matrices. The parsimony of the RTV models focuses on statistical parameters and they overcome all drawbacks associated to the geometric models. The change of the measurement units does not violate the parsimonious assumptions which characterize the RTV models. The choice of some particular measurment units has no effect on the model selection according to some classical likelihood criterion like AIC, BIC or ICL. The RTV models are also stable by projection on any canonical subspace; this enables to represent each of them faithfuly in low dimension.

In the second part of this work we display the so-called simultaneous clustering method. This partitioning process continues the tradition of many statistical methods, originally dedicated to one sample and extended then to several ones. We first show that the classification of a single sample can often be seen as a multiple sample clustering problem; then we propose to establish a link between the original population of the diverse samples. This link varies depending on the context but it always tries to formalize in a realistic way some common information of the samples to classify.
When samples are described by variables with identical meaning and when the same number of groups is researched within each of them, we establish a stochastic link between the conditional populations. This kind of link is first considered in a usual context of Gaussian mixtures, and then extended to Student's ones for the classification of noisy data. We finally mention its interest in mixtures of Factor Analyzers for classifying high dimensional data.
When the variables are different but semantically close through the diverse samples nevertheless their discriminant power may be similar and the nesting of the conditional data can be comparable. We plan specific mixtures dedicated to this context. The link beween the populations consists then in an homogeneous overlap of the components: We assume first an identical Bayes error rate through the diverse mixtures, and then the invariance of the global component entropy.
We show on many real dataset examples, that the hypothesis of a link between populations, whether conditional or global, improves the fit of the model, improves the quality of the estimated partition, and often leads to a convincing interpretation of the inferred structure.

Keywords : Unsupervised classification; Parsimonious mixtures; Correlations; EM algorithm; Link between populations; Classes overlap.

Remerciements

Je remercie Christophe d'avoir accepté d'encadrer ma thèse, d'avoir eu confiance en moi, de sa patience lorsque je piétinais, de ses encouragements discrets dans les moments de doute.
Je le remercie également pour sa franchise, sa rigueur et pour les jugements sans complaisance qu'il a toujours portés sur mon travail.

Je souhaite exprimer ma gratitude à Charles Suquet qui m'a accueilli comme ATER au laboratoire Paul Painlevé de janvier à juin 2008. J'ai pris beaucoup de plaisir aux TD de probabilités à l'ENIC et aux TD de géométrie différentielle avec les étudiants de L3, Pierre, Clara, Anaïs et les autres.

Je n'aurais pas pu soutenir ma thèse sans le soutien indéfectible de ma compagne Dilek. Elle a supporté sans marquer d'exaspération (ou peu), que je m'endorme régulièrement en faisant des mathématiques. Les périodes durant lesquelles je parviens à me concentrer sur une liste de courses, une invitation ou l'organisation d'un déplacement sont rares et de courte durée. Dilek a pris en charge l'organisation matérielle de notre vie quotidienne sans jamais se plaindre et sans jamais perdre patience.
Je la remercie pour tout ce qu'elle est, pour sa douceur et sa générosité particulièrement.

Je salue nos amis Jean, Isabelle, Emmanuel, Gégé(raldine), Camille, René, Agnès, Olivier, Nat(acha), Cat(herine), Fabrice, etc. ainsi que mes parents. Je salue également les anciens thésards du laboratoire Paul Painlevé, Vincent, Alexis, Benoît, Shuyan etc. avec lesquels j'ai passé des moments très agréables.

L'accord d'écart-type au pluriel fait débat et divise gravement la communauté statistique. J. Goupy (`http://www-rocq.inria.fr/axis/modulad/archives/numero-35/Notule-Goupy-35/Goupy-35.pdf`) passe en revue les différents accords possibles et leur fréquence dans la littérature. J'ai opté dans ce qui suit pour la forme la plus répandue qui consiste à accorder les deux noms.

Je dédie cette thèse à ma fille Aurore

Un homme doit toujours étudier même s'il oublie ce qu'il lit, même s'il ne comprend pas [...]. L'école ne doit pas être fermée même pour rebâtir le Temple de Jérusalem[1].

1. Talmud

Table des matières

Enable page numbering

Introduction générale

Etablir des catégories, regrouper des objets, classifier des entités est une (pré)occupation constante et universelle. Les philosophes établissent des catégories de citoyens (Platon), de discours (Aristote), de jugements (Kant), des catégories de l'être (Spinoza). La biologie classifie le vivant (taxinomie et systématique), la linguistique classifie les langues (typologie), l'économie, les marchés (segmentation), la médecine, les maladies (nosologie), etc.
Quel que soit le contexte, l'objectif de la classification est (i) de déterminer si une collection d'objets est homogène et, lorsqu'elle ne l'est pas, (ii) d'établir une partition de cette collection en sous-ensembles homogènes (que l'on appelle des classes). Mais cette définition de bon sens, conduit immédiatement à des difficultés.
Lorsque le critère d'homogénéité n'est pas explicite, lorsqu'il n'est pas défini de façon rigoureuse ou de façon unanime, les classes estimées ne sont pas automatiques : la classification peut varier d'un expert à l'autre. La classification traditionnelle des gastéropodes selon J. Thiele (1935) par exemple, repose sur des observations phénotypiques. On remarque qu'elle est remise en question depuis quelques années par une autre classification, proposée par W. Ponder et D.R. Lindberg (1997), intégrant, elle, des caractéristiques génotypiques. Une des raisons du succès récent des modèles de mélange en classification, tient à ce qu'ils permettent de définir de façon rigoureuse la notion de sous-groupe homogène de données.
Mais une définition rigoureuse de sous-groupes homogènes ne suffit pas. Puisque les classes d'une partition sont relatives à un critère d'homogénéité, elles changent selon le critère considéré. Comparer deux partitions d'un même ensemble d'objets, c'est comparer le critère d'homogénéité sur lequel repose chacune d'elles. Or nous verrons que certaines méthodes même récentes, ne permettent pas une telle comparaison. C'est le cas par exemple de la méthode dite des 'k-means' (dédiée au partitionnement de données continues). Cette procédure subordonne la partition estimée au choix d'une mesure des distances (que l'on appelle la métrique) dans l'espace des données. Rien dans cette méthode, ne permet de comparer les partitions relatives à deux métriques différentes. Ainsi, il ne suffit pas de définir rigoureusement le critère d'homogénéité qui sous-tend la définition des classes. Il est souhaitable en plus, de pouvoir comparer les critères d'homogénéité eux-mêmes. Or là encore les modèles de mélange apportent une réponse satisfaisante. En plaçant la classification dans le contexte plus général de la modélisation probabiliste, les mélanges permettent à la classification d'hériter des outils puissants de l'inférence paramétrique et, en particulier, des critères de choix de modèle.
Mais l'intérêt des mélanges ne réside pas seulement dans la rigueur qu'ils confèrent aux procédures de classification. Nous verrons en effet dans ce travail, que les mélanges comme les critères de choix de modèle relatent une information, qu'ils sont porteurs d'un sens véritable, et que ce sens est pour beaucoup dans les performances et le succès de la classification à base de mélanges.

La classification par un mélange peut prendre plusieurs formes selon l'information disponible. Dans tous les cas elle consiste d'abord à inférer un modèle, puis à en déduire

une règle de classement. Il arrive que certaines données soient déjà classées, et la classification se réduit alors à répartir les autres données (les données sans label) dans les classes existantes. On parle dans ce cas de classification semi-supervisée ou d'analyse discriminante selon que les données sans label participent ou non, à l'apprentissage du modèle. Mais il se peut aussi qu'aucune donnée ne soit classée, que l'on ignore tout des groupes recherchés, jusqu'à leur nombre et leur interprétation. On parle alors de classification non supervisée, de clustering ou encore de classification automatique. Le travail qui suit se place résolument dans un contexte non supervisé. Cependant les modèles dont nous parlerons sont tous utilisables en analyse discriminante, ou dans un contexte semi-supervisé.

Le chapitre 1 passe en revue la classification basée sur des mélanges.
Après avoir défini les mélanges, nous mettons en évidence à la section 1.1.1 leur interprétation générative, et nous en exhibons quelques exemples (Section 1.1.2). Nous décrivons à la section 1.1.3 des mélanges gaussiens reposant sur une interprétation géométrique des classes, puis à la section 1.1.4 des mélanges d'inspiration similaire dédiés aux données de grande dimension. Nous exposons ensuite (Section 1.1.5) quelques problématiques récurrentes en matière de mélanges (identifiabilité, émergence de fausses composantes, dégénérescence). Nous détaillons à la section 1.1.6, la procédure de classification elle-même, ainsi que les notions afférentes de classe, de groupe, de règle de classement, de frontière, etc. Nous exposons (Section 1.2.1) le principe général de l'algorithme EM et en particulier, nous montrons (Section 1.2.2) qu'il est adapté à l'estimation du paramètre d'un mélange par maximum de vraisemblance. Grâce à une décomposition particulière de la vraisemblance complétée (Section 1.2.3), EM peut être vu comme un algorithme d'optimisation alternée d'un critère portant à la fois sur le paramètre du mélange et sur une partition des données. Cette interprétation d'EM (i) permettra d'interpréter certaines méthodes de classification historiques de façon probabiliste et (ii) constituera la base d'un nouveau critère pour l'inférence du paramètre en classification simultanée. Nous décrivons à la section 1.3 quelques critères de choix de modèle utilisés dans ce travail, et nous montrons que l'usage de l'un ou l'autre de ces critères est toujours lié à un objectif particulier.

Les mélanges gaussiens d'inspiration géométrique décrits à la section 1.1.3 sont largement utilisés, mais ils présentent des inconvénients majeurs. La modification des unités de mesure enfreint généralement les hypothèses parcimonieuses sur lesquelles ils reposent. Le changement des unités de mesure (et en particulier la réduction des variables) altère le choix de l'un de ces modèles par certains critères de vraisemblance classiques comme AIC, BIC ou ICL. Les modèles géométriques ne sont pas stables par projection dans les plans canoniques ; cela empêche de représenter en dimension 2 un modèle de dimension supérieure. Aussi proposons-nous au chapitre 2, de nouveaux modèles de mélanges dont la parcimonie porte sur des paramètres d'interprétation statistique et non plus géométrique : l'écart-type et le coefficient de variation des variables ainsi que la corrélation des covariables. Ces modèles que nous appelons mélanges RTV, sont stables par projection dans les plans canoniques (Section 2.3.1) ; ils peuvent donc être représentés de façon fidèle en dimension réduite. Le changement des unités de mesure n'enfreint aucune des hypothèses parcimonieuses qui les caractérisent. Le choix d'un modèle dans la famille RTV par l'un des critères de vraisemblance précédents, est

invariant à la modification des unités de mesure (Section 2.2.1) et invariant (donc) à la réduction des données. Enfin, contrairement aux autres familles de mélanges gaussiens parcimonieux, beaucoup de modèle RTV permettent une interprétation dynanique des classes inférées (Section 2.4). Nous mettrons en évidence sur de nombreux exemples l'avantage manifeste des modèles RTV sur les mélanges d'inspiration géométrique.

Nous proposons dans la seconde partie de ce travail, une méthode nouvelle dite de classification simultanée. Dans de nombreux cas la classification d'un seul échantillon se ramène à la classification de plusieurs échantillons, et cette transformation du contexte de classification permet très souvent d'améliorer à la fois l'adéquation du modèle et la qualité de la partition estimée.
Le chapitre 3 s'emploie tout entier à montrer que les situations de classification simultanée sont très nombreuses, à décrire le principe de la méthode, et à justifier ses performances.
La classification simultanée repose toujours sur la formalisation d'un lien entre populations. Lorsque les échantillons à classifier sont décrits par des variables de même signification, nous verrons (Chapitre 4) qu'il peut être judicieux d'établir un lien stochastique affine entre populations conditionnelles. Nous envisagerons d'abord ce type de lien dans un contexte général de mélanges gaussiens (Section 4.1). Nous l'étendrons aux mélanges de Student (Section 4.3) destinés à classifier des données bruitées. Enfin nous proposerons de façon étayée à la section 4.4, d'appliquer cette forme de la classification simultanée, aux données de grande dimension modélisées par des mélanges de Factor Analyzers.
Nous proposons au chapitre 5 une forme de la classification simultanée dédiée à un contexte plus large. Il arrive que les descripteurs des échantillons à classifier n'aient pas tout à fait le même sens, mais possèdent le même pouvoir discriminant. On supposera dans ce cas que l'imbrication des données conditionnelles est semblable d'un échantillon à l'autre, et que les composantes des différents mélanges se chevauchent de façon similaire. Nous traduirons cette hypothèse en supposant homogène l'erreur de Bayes des différents mélanges (Chapitre 5.2), ou l'entropie globale de leurs composantes (Chapitre 5.3). En particulier nous présentons un algorithme appelé $\tilde{\text{E}}$M (Section 5.2), basé sur l'interprétation d'EM selon la formule d'Hathaway (Section 1.2.3) et dédié à l'inférence de mélanges (gaussiens ou non) dont l'entropie des composantes est homogène.

L'hypothèse d'un lien entre populations, qu'il s'agisse d'un lien stochastique conditionnel ou de l'homogénéité du chevauchement des classes, améliore très souvent la qualité du modèle. Mais les nombreuses applications de la classification simultanée que nous citons, montrent de plus que cette hypothèse conduit à une structure des données dont l'interprétation est très souvent convaincante.

Première partie

Classification par modèles de mélange

Sommaire

Chapitre 1

Introduction

1.1 Modèles de mélange

1.1.1 Présentation et interprétation générative

Un mélange fini sur un espace $\mathcal{X}$ est une combinaison linéaire de lois de probabilités $f_k(\bullet)$ $(k = 1, \ldots, K)$ (appelées les *composantes* du mélange) définies sur $\mathcal{X}$:

$$f(\bullet) = \sum_{k=1}^{K} \pi_k f_k(\bullet). \tag{1.1}$$

Les coefficients π_k, appelés *proportions mélange* ou *poids* des composantes, sont positifs $(\forall k : \pi_k > 0)$ et leur somme vaut 1 $(\sum_{k=1}^{K} \pi_k = 1)$.

Cette définition générale recouvre une famille très large de lois de probabilités. On suppose souvent, selon le domaine d'intérêt, que les composantes f_k $(k = 1, \ldots, K)$ du mélange sont paramétrées par $\boldsymbol{\alpha}_k$ $(k = 1, \ldots, K)$. La loi mélange (1.1) sécrit alors :

$$f(\bullet\,; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\bullet\,; \boldsymbol{\alpha}_k), \tag{1.2}$$

et $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$ désigne le paramètre du mélange. Les composantes $f_k(\bullet\,; \boldsymbol{\alpha}_k)$ appartiennent souvent dans ce cas, à une même famille de lois dont le choix dépend des données à représenter. Il est classique de modéliser la distribution du nombre des couleurs dans une image ou la distribution du nombre des mots dans un texte (données discrètes), par un mélange multinomial ([105]) ; Mendoza-Rosas et Cruz-Reyna (2009) ([89]) modélisent le temps de repos d'un volcan entre deux erruptions (données continues) par un mélange de lois exponentielles. Les covariables des composantes sont parfois de nature différente : Jorgensen et Hunt (1996) ([66]) modélisent des données médicales par un mélange de lois mixtes pour des covariables discrètes et quantitatives. Les composantes d'un mélange n'appartiennent pas toujours à la même famille de lois : Dasgupta et Raftery (1998) ([33]) introduisent une composante uniforme dans un mélange gaussien pour modéliser des données bruitées. On peut réduire le déterminisme des populations conditionnelles en ne faisant aucune hypothèse (ou des hypothèses partielles) sur

les composantes $f_k(\bullet)$. Ce type de mélange, dit non paramétrique ou semi-paramétrique, rencontre un intérêt croissant notamment en économétrie ou en médecine ([75]).

Quelle que soit la spécificité des composantes f_k (ou leur défaut de spécificité pour les mélanges semi-paramétriques et non paramétriques), le mélange (1.1) peut être vu comme la loi marginale de la variable $\mathbf{X}$ pour un couple $(\mathbf{Z}, \mathbf{X})$ tel que (i) $\mathbf{Z}$ est un vecteur binaire à valeur dans $\{0;1\}^K$ distribué selon la loi multinomiale d'ordre 1 et de paramètre $(\pi_1, \ldots, \pi_k) : \mathcal{M}_K(1, \pi_1, \ldots, \pi_K)$ (la k^e composante Z_k de $\mathbf{Z}$ vaut 1 et les autres 0, avec probabilité π_k) et (ii) pour tout k, le vecteur conditionnel $(\mathbf{X}|Z_k = 1)$ à valeur dans $\mathcal{X}$ est distribué selon $f_k(\bullet)$.

Ainsi, générer une donnée $\mathbf{x} \in \mathcal{X}$ selon (1.1) revient à (i) choisir l'une des composantes du mélange (la composante k est choisie avec probabilité π_k) puis (ii) générer $\mathbf{x}$ selon $f_k(\bullet)$.

Inversement, à toute donnée $\mathbf{x}$ de $\mathcal{X}$ générée selon (1.1), correspond une réalisation $\mathbf{z}$ de la variable $\mathbf{Z}$ (appelée *variable latente*) indiquant (lorsqu'on la connaît) la composante dont $\mathbf{x}$ est issue.

Leur interprétation générative combinée à la variété de lois conditionnelles $f_k(\bullet)$ possibles et à la souplesse offerte par le choix du nombre K de composantes, est déterminante dans le succès croissant des modèles de mélange pour classifier des données et plus généralement pour les modéliser.

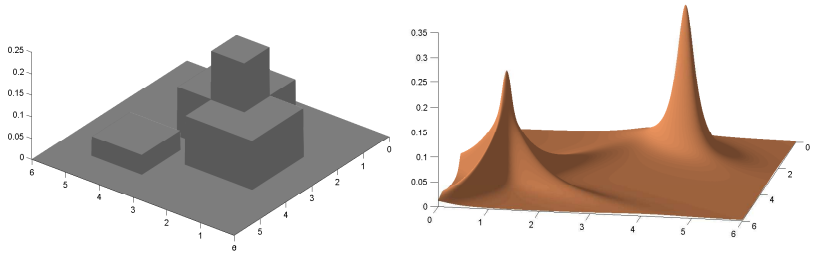
1.1.2 Exemples de mélanges

Vocation et usage des mélanges

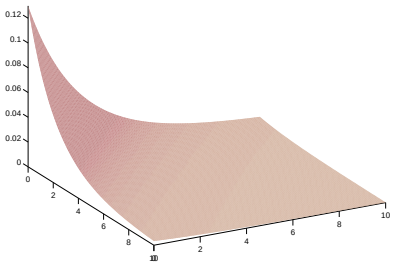
La vocation d'un mélange est de modéliser des données hétérogènes (ou supposées telles). Cette fonction est explicite dans [90] où Pearson fonde l'usage des mélanges : il justifie l'asymétrie d'un histogramme (représentant les poids de crabes) en supposant que l'échantillon qu'il étudie provient de deux populations gaussiennes distinctes et non d'une seule.

Lorsqu'on modélise des données hétérogènes par un mélange, il est souhaitable, que les composantes relatent une spécificité des données conditionnelles. Ainsi les mélanges de lois uniformes, de lois de Student ou de lois exponentielles (FIG. 1.1) ne modélisent pas le même type de données.

Mais en pratique, la spécification des composantes d'un mélange n'est pas toujours guidée par la singularité des données à modéliser. Les mélanges gaussiens, qui supposent les populations conditionnelles distribuées selon une loi normale, suscitent un intérêt important en raison (i) de leur flexibilité, (ii) de leur faculté à approcher une grande variété



(a) *Mélange de lois uniformes* (b) *Mélange de lois de Student*



(c) *Mélange de lois exponentielles*

FIGURE 1.1: *Mélanges de lois continues, à trois composantes : $K = 3$*

de densités comme le montrent Marron et Wand (1992) ([81]) dans le cas univarié, (iii) de leur usage mathématiquement simple et (iv) de la généralité de la loi normale qu'atteste le théorème central limite.

Ainsi l'usage des mélanges est subordonné au choix d'un paradigme. Le premier paradigme est orienté vers l'interprétation des classes. Il suppose (i) que les données conditionnelles sont homogènes et (ii) que la spécificité des composantes traduit une information intrinsèque aux données conditionnelles. Le second paradigme est orienté vers la modélisation au sens d'un modelage (ou d'une estimation de densité). Les composantes ne jouent qu'un rôle technique et ne sont pas destinées à être interprétées. Leur multiplication permet au modèle estimé d'épouser la forme de modèles complexes. L'usage de mélanges gaussiens est courant dans les deux cas.

Mélanges gaussiens

La densité d'un mélange gaussien défini sur $\mathbb{R}^d$, sécrit :

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k (2\pi)^{-d/2} |\boldsymbol{\Sigma}_k|^{-1/2} \exp\{-(\boldsymbol{x}-\boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)/2\} \ ; \ \boldsymbol{x} \in \mathbb{R}^d. \tag{1.3}$$

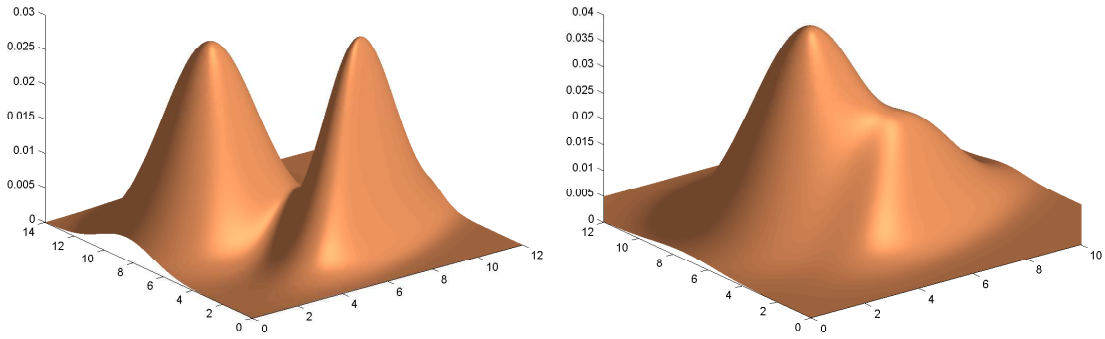
Le vecteur $\boldsymbol{\mu}_k$ de $\mathbb{R}^d$ désigne le centre de la composante k ; c'est la moyenne du vecteur conditionnel (gaussien) $(\boldsymbol{X}|Z_k = 1)$:

$$\boldsymbol{\mu}_k = E(\boldsymbol{X}|Z_k = 1). \tag{1.4}$$

La matrice $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$ est symétrique, définie, positive et désigne la matrice des covariances de la composante k :

$$\boldsymbol{\Sigma}_k = E\left[\{(\boldsymbol{X}|Z_k = 1) - \boldsymbol{\mu}_k\}\{(\boldsymbol{X}|Z_k = 1) - \boldsymbol{\mu}_k\}'\right]. \tag{1.5}$$

Les figures 1.2(a) et 1.2(b) représentent chacune un mélange gaussien de deux composantes défini sur $\mathbb{R}^2$. La valeur du paramètre de chaque mélange est indiquée dans le tableau 1.1.



(a) (b)

FIGURE 1.2: *Mélanges gaussiens de deux composantes*

Paramètre	π_1	π_2	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_1$	$\boldsymbol{\Sigma}_2$
FIG. 1.2(a)	0.6	0.4	$\begin{pmatrix}4\\9\end{pmatrix}$	$\begin{pmatrix}8\\5\end{pmatrix}$	$\boldsymbol{S}_{\pi/6}\begin{pmatrix}4&0\\0&3\end{pmatrix}\boldsymbol{S}'_{\pi/6}$	$\boldsymbol{S}_{-\pi/3}\begin{pmatrix}1&0\\0&5\end{pmatrix}\boldsymbol{S}'_{-\pi/3}$
FIG. 1.2(b)	0.8	0.2	$\begin{pmatrix}5\\8\end{pmatrix}$	$\begin{pmatrix}7.5\\5.5\end{pmatrix}$	$\boldsymbol{S}_{\pi/6}\begin{pmatrix}4&0\\0&3\end{pmatrix}\boldsymbol{S}'_{\pi/6}$	$\boldsymbol{S}_{-\pi/3}\begin{pmatrix}1&0\\0&5\end{pmatrix}\boldsymbol{S}'_{-\pi/3}$

TABLE 1.1: *Valeur du paramètre des mélanges gaussiens de* FIG. 1.2(a) *et* 1.2(b). $\boldsymbol{S}_\theta$ *désigne la matrice* 2×2 *associée à la rotation d'angle* θ.

Le mélange de FIG. 1.2(a) est bimodal. Ses composantes étant bien séparées, les données que génèrerait un tel mélange (ou des données qui conduiraient à l'inférence d'un tel modèle) seraient manifestement hétérogènes.
Le mélange de FIG. 1.2(b), lui, est unimodal et seule son asymétrie laisse à penser que s'il s'agit d'un mélange gaussien, il comporte au moins deux composantes. Pour une taille d'échantillon semblable, l'hétérogénéité de données générées par ce modèle serait plus difficile à déceler que dans le cas précédent.

Mélanges de mélanges

Les composantes d'un mélange sont parfois elles-mêmes, des mélanges. Le mélange de FIG. 1.3 comporte deux composantes ; chacune d'elles est un mélange de quatre composantes gaussiennes. On peut montrer qu'un tel mélange n'est pas identifiable (la même fonction de densité peut être obtenue pour une infinité de proportions mélange lorsqu'on ne les connaît pas).

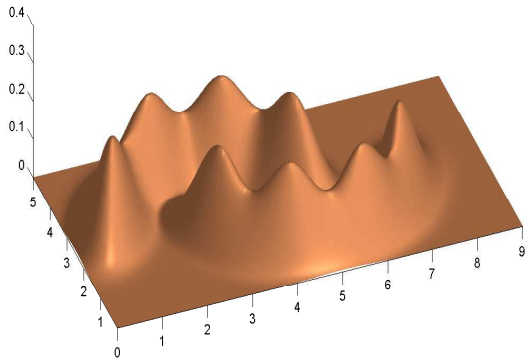


FIGURE 1.3: *Un mélange de mélanges : chacune des deux composantes est un mélange de (quatre) composantes gaussiennes.*

Ce type de mélange est utilisé par Hastie et Tibshirani (1996) ([59]) dans un contexte d'analyse discriminante, pour distinguer dans une collection de chiffres écrits à la main, les 3 des 8. Les chiffres 3 et les chiffres 8 correspondent aux deux composantes d'un mélange. Supposer que ces composantes sont elles-mêmes des mélanges, est dans leur cas, une hypothèse technique permettant de rendre maléable la frontière discriminante et ne donne lieu à aucune interprétation.

Vermunt et Magidson(2005) ([107]) modélisent par un mélange de mélanges (multinomiaux) les réponses à un questionnaire, données par des employés de plusieurs équipes dans différentes entreprises. Dans leur cas les deux variables catégorielles latentes (équipe, entreprise) que suppose le modèle, traduisent la structure des donnnées et les classes obtenues sont interprétables : une classe correspond à une équipe dans l'une des entreprises considérées.

En faveur de composantes interprétables

La pertinence d'un modèle tient au compromis réalisé entre son biais (par rapport au vrai modèle) et sa variabilité (aux fluctuations d'échantillonnage). Nous prétendons que la seule façon d'espérer diminuer à la fois le biais et la variabilité d'un mélange consiste, lorsque c'est possible, à (i) choisir ses composantes de façon réaliste (de sorte qu'elles traduisent une spécificité des données conditionnelles) et (ii) à imposer aux composantes choisies, une contrainte pleine de sens (que l'on puisse interpréter comme une propriété des données conditionnelles).

Lorsque l'échantillon est de grande taille, on peut toujours modéliser des données continues par un mélange gaussien, comme dans [59]. La multiplication du nombre des composantes, combinée à la flexibilité de la loi normale, permet d'approcher le vrai modèle quelle que soit sa complexité. On peut donc penser que des mélanges gaussiens suffisent à modéliser toutes les sortes de données continues (leur rôle est alors d'estimer une densité) et qu'il est inutile de définir ou d'employer à cet effet d'autres types de mélanges. Mais en pratique la taille des échantillons est souvent limitée, et l'on doit compenser par des hypothèses réalistes sur le modèle, le manque d'information apporté par les données.

D'abord, lorsqu'on modélise un échantillon de petite taille, il est exclu d'augmenter le nombre des composantes d'un mélange dans l'intention d'approcher au mieux le vrai modèle : il se pourrait en effet qu'une composante ne relate qu'une singularité de l'échantillon. On supposera donc dans ce qui suit, que les données conditionnelles générées par chacune des composantes d'un mélange, sont homogènes.

Traduire une spécificité des données conditionnelles par un choix réaliste de la famille de lois à laquelle appartiennent les composantes, permet alors de diminuer naturellement le biais du modèle. Des composantes exponentielles par exemple, semblent indiquées pour des données de durées de vie, et l'usage d'un autre type de composantes risque au contraire, d'augmenter le biais du modèle.

Remarque. Il arrive qu'aucune loi conditionnelle ne soit plus indiquée qu'une autre (lorsqu'on ignore la nature des données par exemple). Dans ce cas, on emploiera souvent des composantes gaussiennes, en raison de leur flexibilité. La multiplicité de leurs paramètres est en effet déterminante dans leur capacité à approcher beaucoup d'autres lois.

Quel que soit le type de mélange considéré, si la taille du paramètre est grande par rapport à la taille de la donnée, le modèle inféré dépend fortement des fluctuations d'échantillonnage. On peut réduire la variabilité du modèle en imposant une contrainte au paramètre, mais une contrainte aveugle, dépourvue de sens, augmentera certainement (elle aussi) le biais du modèle.
La seule façon d'espérer diminuer simultanément la variabilité et le biais du modèle estimé, est de contraindre le paramètre du mélange, de façon réaliste, c'est à dire en formalisant à travers cette contrainte, une information de bon sens que recèlent les données.
Un mélange est dit parcimonieux lorsqu'on impose une contrainte à certains de ses paramètres. Nous présenterons dans ce qui suit, deux familles de mélanges gaussiens dont la parcimonie porte sur des paramètres respectivement géométriques (Section 1.1.3) et statistiques (Section2.1.2). En particulier nous montrerons l'avantage manifeste des modèles statistiques qui permettent (contrairement aux modèles géométriques) d'interpréter la spécificité des composantes inférées comme une propriété des données conditionnelles.

Remarque. En imposant une contrainte au paramètre, on réduit la variabilité du modèle. Par ailleurs, le critère BIC (voir section 1.3) permet de détecter pour de grands

échantillons, une diminution du biais : entre deux modèles, il choisit asymptotiquement le plus proche du vrai modèle.

Ainsi le choix du type de mélange comme le choix des contraintes que l'on impose à ses composantes, sont guidés par la recherche d'un compromis entre le biais du modèle et sa variabilité. Or le compromis biais-variance recherché est d'autant meilleur que les composantes choisies et les contraintes imposées relatent une information véritable, portant sur les données conditionnelles.

1.1.3 Mélanges gaussiens parcimonieux d'interprétation géométrique

Flury et al. (1994) ([42]) définissent dans un contexte d'analyse discriminante, des modèles parcimonieux de mélanges gaussiens, basés sur une décomposition spectrale des matrices de covariances. Celeux et Govaert (1995) [28] généralisent l'usage de ce type de mélanges gaussiens et mettent en évidence leurs liens avec des méthodes de classification usuelles.

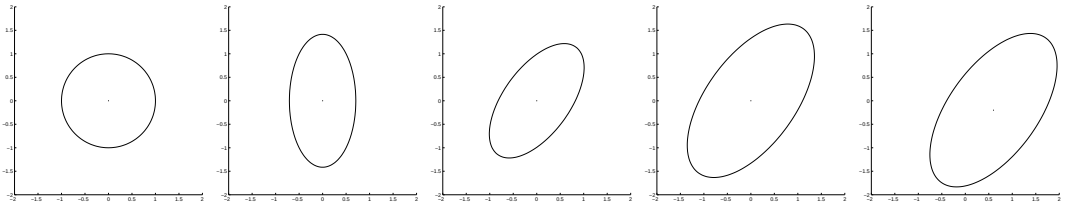
Une matrice symétrique, définie et positive est diagonalisable dans une base orthonormée (de vecteurs propres). Aussi chaque matrice des covariances $\mathbf{\Sigma}_k$ d'un mélange gaussien se décompose-t-elle en :

$$\mathbf{\Sigma}_k = \lambda_k \boldsymbol{S}_k \mathbf{\Lambda}_k \boldsymbol{S}_k', \tag{1.6}$$

où : (i) $\lambda_k = |\mathbf{\Sigma}_k|^{1/d}$, (ii) $\boldsymbol{S}_k$ est l'une des matrices orthogonales constituées des vecteurs propres de $\mathbf{\Sigma}_k$ et (iii) $\mathbf{\Lambda}_k$ est une matrice diagonale, définie et positive, de déterminant 1. Les coefficients diagonaux de $\mathbf{\Lambda}_k$ sont alors les valeurs propres normalisées de $\mathbf{\Sigma}_k$ c'est à dire ses valeurs propres divisées par λ_k. Si on leur impose d'être rangés par ordre décroissant, la décomposition (1.6) est unique.

Quel que soit le mélange gaussien (1.3) considéré, on peut représenter le paramètre $(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$ de la classe k, par l'ellipsoïde des points de $\mathbb{R}^d$ situés à la distance 1 de $\boldsymbol{\mu}_k$, pour la distance de Mahalanobis $\mathbf{\Sigma}_k^{-1}$: $\Gamma_k(1, \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) = \{\boldsymbol{x} \in \mathbb{R}^d; (\boldsymbol{x} - \boldsymbol{\mu}_k)' \mathbf{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) = 1\}$. Γ_k s'obtient, d'après la décomposition (1.6), par transformations de la sphère unité $\Gamma_k(1, \mathbf{0}, \boldsymbol{I}_d)$ selon, successivement : la composée des affinités orthogonales induites par la matrice $\mathbf{\Lambda}_k^{1/2}$, la rotation de matrice $\boldsymbol{S}_k$, l'homothétie de rapport $\lambda_k^{1/2}$ et la translation de vecteur $\boldsymbol{\mu}_k$. La figure 1.4 illustre dans $\mathbb{R}^2$, la composition des transformations affines qui mènent du cercle unité à l'ellipse Γ_k. Ainsi λ_k, $\boldsymbol{S}_k$, $\mathbf{\Lambda}_k$ et $\boldsymbol{\mu}_k$ apparaissent-ils comme des paramètres qui confèrent respectivement à l'ellipsoïde $\mathbf{\Gamma}_k$, son volume, son orientation, sa forme et sa position.

Remarque. La décomposition (1.6) permet également de voir le vecteur conditionnel $(\mathbf{X}|Z_k = 1)$ comme issu de la transormation stochastique successive d'un vecteur aléatoire normal centré réduit $\boldsymbol{U} \in \mathbb{R}^d$. Par affinités orthogonales $\boldsymbol{U}$ se transforme en



(a) $\Gamma_k(1, \mathbf{0}, \boldsymbol{I}_2)$ (b) *Forme* (c) *Orientation* (d) *Volume* (e) $\Gamma_k(1, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

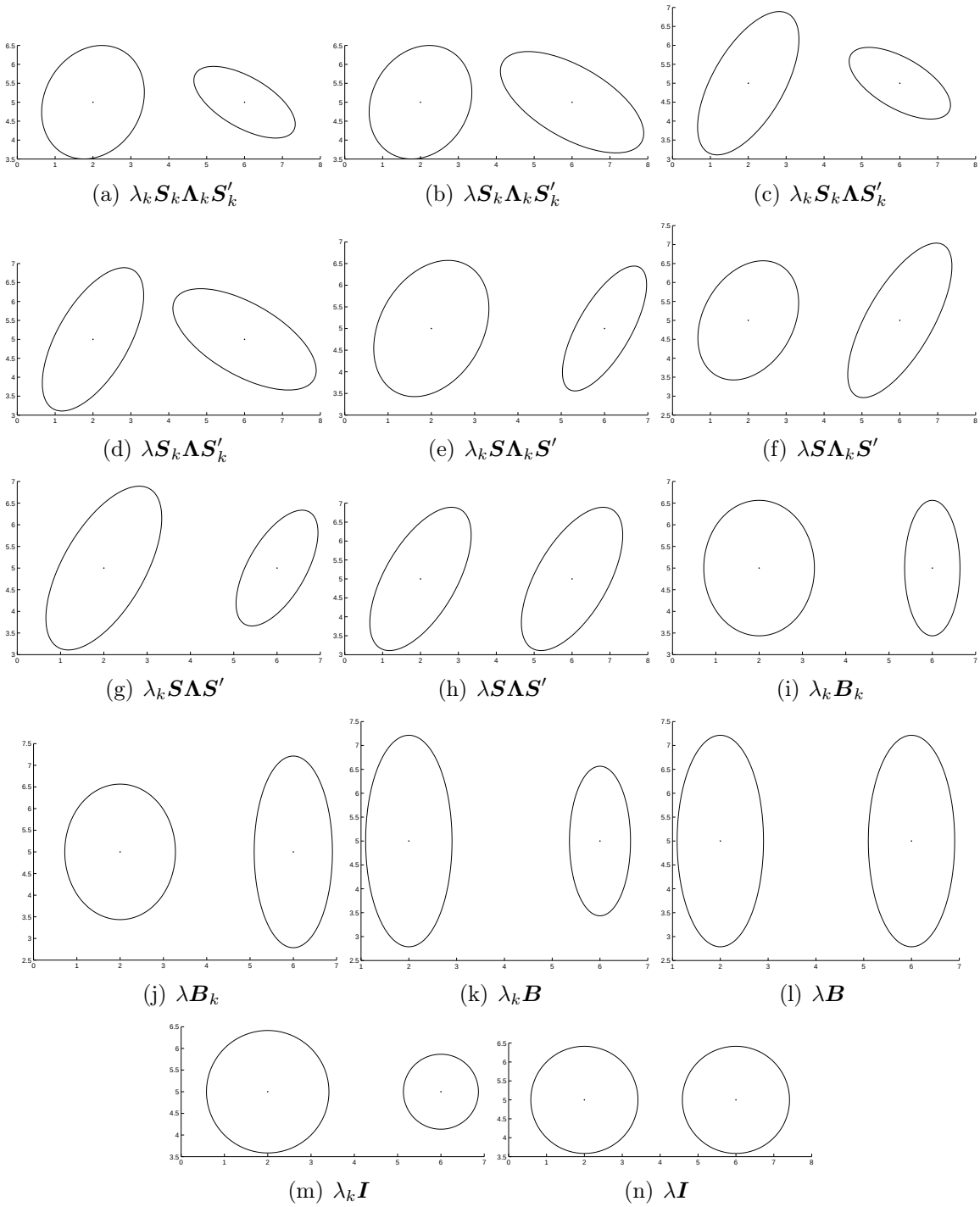
FIGURE 1.4: *Définition des paramètres de forme ($\boldsymbol{\Lambda}_k$), d'orientation ($\boldsymbol{S}_k$), de volume (λ_k) et de position ($\boldsymbol{\mu}_k$) de l'ellipsoïde Γ_k, par transformations (affines) successives de la sphére unité.*

$\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{U}$, un vecteur gaussien dont les covariables sont aussi indépendantes mais pas forcément réduites. La rotation de matrice $\boldsymbol{S}_k$ le transforme en $\boldsymbol{S}_k\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{U}$; elle fait perdre leur indépendance aux covariables. La dilatation de rapport $\lambda_k^{1/2}$ change la variance des covariables de façon isotropique (de la même façon dans toutes les directions) pour donner le vecteur $\lambda_k^{1/2}\boldsymbol{S}_k\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{U}$. La translation de vecteur $\boldsymbol{\mu}_k$ donne alors le vecteur $\lambda_k^{1/2}\boldsymbol{S}_k\boldsymbol{\Lambda}_k^{1/2}\boldsymbol{U} + \boldsymbol{\mu}_k$ de même loi que $(\mathbf{X}|Z_k = 1)$.

Flury et al. (1994) [42] considèrent dans un contexte d'analyse discriminante, quatre modèles d'intérêt qui supposent successivement : (i) des matrices de covariances libres, (ii) des classes ayant la même orientation (mais dont la forme et le volume sont libres), (iii) des classes ayant la même orientation, la même forme mais dont le volume est libre, et (iv) des matrices de covariances homogènes.

Celeux et Govaert (1995) [28] proposent de systématiser la combinaison des hypothèses entre l'homogénéité et la liberté de la forme, de l'orientation et du volume des classes. Ils définissent ainsi une famille, dite générale, de huit modèles. $[\lambda\boldsymbol{S}_k\boldsymbol{\Lambda}_k\boldsymbol{S}_k']$ par exemple désigne le modèle de cette famille qui suppose les volumes homogènes et les autres paramètres libres. Ils envisagent d'autre part que les covariables conditionnelles puissent être indépendantes c'est à dire que les matrices $\boldsymbol{S}_k$ soient toutes égales à l'identité (ou encore que les matrices $\boldsymbol{B}_k = \boldsymbol{S}_k\boldsymbol{\Lambda}_k\boldsymbol{S}_k'$ soient diagonales) et constituent ainsi une seconde famille, dite diagonale, de quatre modèles. Ils définissent enfin la famille sphérique (de deux modèles) en imposant dans chaque classe des covariables indépendantes et de même variance (matrices $\boldsymbol{B}_k$ égales à l'identité). Les figures 1.5(a) à 1.5(n) illustrent chacune, un des quatorze modèles de mélange définis par [28], sur $\mathbb{R}^2$.

La représentation géométrique des classes que permet la décomposition (1.6) est souvent commode. Elle permet par exemple à Bouveyron (2006) [21] d'imaginer des mélanges gaussiens spécifiques destinés à classifier des données placées dans un espace de grande dimension. La disproportion entre la taille de la donnée et la dimension de l'espace se traduit dans ce contexte, par une faible variabilité de certaines combinaisons de variables. En imaginant que les ellipsoïdes associés aux classes sont relativement aplatis dans certaines directions, Bouveyron impose des contraintes aux valeurs propres des matrices de covariances. Il définit ainsi deux classifieurs ad-hoc pour les données de grande dimension : HDDA (High Dimensional Discriminant Analysis) pour les données



(a) $\lambda_k \boldsymbol{S}_k \boldsymbol{\Lambda}_k \boldsymbol{S}'_k$ (b) $\lambda \boldsymbol{S}_k \boldsymbol{\Lambda}_k \boldsymbol{S}'_k$ (c) $\lambda_k \boldsymbol{S}_k \boldsymbol{\Lambda} \boldsymbol{S}'_k$

(d) $\lambda \boldsymbol{S}_k \boldsymbol{\Lambda} \boldsymbol{S}'_k$ (e) $\lambda_k \boldsymbol{S} \boldsymbol{\Lambda}_k \boldsymbol{S}'$ (f) $\lambda \boldsymbol{S} \boldsymbol{\Lambda}_k \boldsymbol{S}'$

(g) $\lambda_k \boldsymbol{S} \boldsymbol{\Lambda} \boldsymbol{S}'$ (h) $\lambda \boldsymbol{S} \boldsymbol{\Lambda} \boldsymbol{S}'$ (i) $\lambda_k \boldsymbol{B}_k$

(j) $\lambda \boldsymbol{B}_k$ (k) $\lambda_k \boldsymbol{B}$ (l) $\lambda \boldsymbol{B}$

(m) $\lambda_k \boldsymbol{I}$ (n) $\lambda \boldsymbol{I}$

FIGURE 1.5: *Quatorze modèles de mélanges gaussiens dont la parcimonie porte sur des paramètres géométriques.*

supervisées et HDDC (High Dimensional Data Clustering) pour les données non supervisées.

Remarque. Que les modèles parcimonieux de [28] permettent une représentation géométrique des classes n'est pas leur seul intérêt. G. Govaert (2009) ([49]) montre l'équivalence entre des méthodes de classification usuelles et la classification par des mélanges de [28]. Classifier des données sans label (contexte non supervisé) en supposant qu'elles

proviennent du modèle $[\lambda \boldsymbol{I}]$, avec des classes de même poids : $\pi_1 = \cdots = \pi_K$, et en inférant le paramètre par un algorithme CEM (Section 1.2.4), est strictement équivalent à une méthode de k-means basée sur une métrique euclidienne, ou encore à la méthode de partitionnement de Ward ([108]) basée sur la minimisation de l'inertie intra-classes. Cette interprétation n'est pas isolée : G. Celeux et G. Govaert établissent une correspondance systématique entre l'inférence d'un modèle de [28] par CEM, et l'optimisation d'un critère géométrique spécifique.

1.1.4 Mélanges gaussiens dédiés à la grande dimension

La modélisation et la classification de données situées dans un espace de grande dimension, connaissent un essor important depuis quelques années, sous l'effet (entre autre) (i) du développement de la reconnaissance d'objets dans le domaine de l'intelligence artificielle (voir [21]) et (ii) du développement des méthodes de séquençage de l'ADN et de l'intérêt (relativement) récent des biologistes pour la détermination de la fonction des gènes (annotation fonctionnelle du génome) (voir [5]).

Les données de grande dimension se caractérisent par une carence d'information dans certaines régions de l'espace. Dans un espace de dimension n par exemple, n points sont toujours contenus dans un hyperplan et n'apportent aucune information sur la dispersion de la population dans le supplémentaire orthogonal de cet hyperplan. Si l'on suppose que ces points proviennent d'une loi normale, l'estimation conduit à une loi dégénérée : comme la taille de l'échantillon est faible, la matrice des covariances estimée est singulière. La singularité du paramètre traduit alors non pas une spécificité de la donnée, mais une trop grande complexité du modèle. Ainsi la carence d'information dans un espace de grande dimension s'explique en partie par une disproportion entre la taille de la donnée et de la dimension de l'espace.

Mais le manque d'information dans certaines régions revêt également un caractère intrinsèque à la dimension de l'espace. Par exemple, à partir de la dimension 5, lorsque la dimension de l'espace augmente, le volume de la boule unité décroît jusqu'à valoir 0 dans un espace de dimension infinie. On peut également montrer que plus la dimension d'un espace est grande, plus des données réparties de façon uniforme dans sa boule unité, sont proches du bord. Ainsi dans un espace de dimension 20, près de 90% des points distribués de façon uniforme à l'intérieur de la boule unité, sont à une distance du centre supérieure à 0.9 (voir [21], p.38).

Les méthodes traditionnelles de classification en grande dimension, conditionnent les données avant de les classifier, de façon à les placer dans un espace de dimension réduite. On peut ranger ces méthodes en deux catégories : celles qui extraient des variables et celles qui sélectionnent des variables.

Extraction et sélection de variables, deux conditionnements classiques

Les méthodes d'extraction de variables comme PCA ([65]) (et les méthodes non linéaires associées : Kernel-PCA de [96] ou non-linear PCA de [48] et [58]) ainsi que les techniques liées aux réseaux de neurones ([36], [71], [95], [102]) proposent de construire à partir des variables canoniques, des descripteurs (rarement interprétables) qui portent l'essentiel de l'information. Ces méthodes sont passées en revue dans [43] et [24].
Les méthodes de sélection de variables dont on trouvera une revue dans [54], proposent, elles, de choisir parmi les variables canoniques, celles qui recèlent le plus d'information. Mais ces méthodes reposent sur deux idées contestables. Elles supposent que le sous-espace contenant le plus d'information (que ce soit un sous-espace canonique pour les méthodes de sélection, ou un sous-espace factoriel pour les méthodes d'extraction), est le plus approprié à la classification. Or on peut imaginer que l'information utile à la classification soit au contraire contenue dans le supplémentaire orthogonal de ce sous-espace.
D'autre part, lorsque la dimension de l'espace est supérieure à la taille de l'échantillon, les données vivent dans un espace de dimension inférieure, que l'on appelle le sous-espace intrinsèque des données. Les méthodes d'extraction comme les méthodes de sélection de variables, considèrent généralement que le sous-espace intrinsèque est le même pour toutes les données conditionnelles, hypothèse qui est, elle aussi, contestable.

Une alternative à l'extraction et à la sélection de variables consiste à modéliser les données de grande dimension par un mélange spécifique, dont les composantes traduisent le manque d'information dans certaines directions de l'espace. Nous citerons ici deux types de mélanges gaussiens dédiés à la grande dimension : les mélanges parcimonieux de C. Bouveyron (2006) ([21]) et les mélanges de Factor Analyzers de G. McLachlan et D. Peel (2000) ([85]).

Deux modèles gaussiens concurrents pour la grande dimension, extension à la classification simultanée

Un vecteur normal centré réduit $\boldsymbol{U} \in \mathbb{R}^b$ est un Factor Analyzer pour un vecteur gaussien $\mathbf{Y} \in \mathbb{R}^d$ $(d > b)$ de centre $\boldsymbol{\mu}$ et de matrice des covariances $\boldsymbol{\Sigma}$, s'il existe une matrice $\boldsymbol{B} \in \mathbb{R}^{d \times b}$ et un vecteur $\boldsymbol{e} \in \mathbb{R}^d$ normal, centré, de covariables indépendantes, indépendant de $\boldsymbol{U}$, tel que :

$$\mathbf{Y} \sim \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{U} + \boldsymbol{e}. \tag{1.7}$$

Ce modèle revient à postuler que l'aléa ne consiste qu'en un terme d'erreur dans le supplémentaire orthogonal du sous-espace (de $\mathbb{R}^d$) engendré par les vecteurs colonnes de $\boldsymbol{B}$. Il implique que la matrice des covariances de $\mathbf{Y}$ est de la forme $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Delta}$ ($\boldsymbol{\Delta}$ désigne la matrice diagonale définie positive des variances de $\boldsymbol{e}$.)

Supposer que chaque vecteur conditionnel $(\mathbf{X}|Z_k = 1)$ admet un Factor Analyzer :

$$(\mathbf{X}|Z_k = 1) \sim \boldsymbol{\mu}_k + \boldsymbol{B}_k\boldsymbol{U}_k + \boldsymbol{e}_k, \tag{1.8}$$

permet de modéliser composante par composante le défaut d'information dans certaines directions de l'espace, en donnant aux matrices de covariances conditionnelles la forme particulière $\boldsymbol{\Sigma}_k = \boldsymbol{B}_k\boldsymbol{B}_k' + \boldsymbol{\Delta}_k$. Les mélanges de Factor Analyzers sont donc des mélanges gaussiens hétéroscédastiques particuliers, dont la parcimonie porte sur la dimension com-

mune b des variables latentes $\boldsymbol{U}_k$.

Pour modéliser les vecteurs gaussiens $(\mathbf{X}|Z_k = 1)$ dans un espace de grande dimension, C. Bouveyron (2006) ([21]) s'inspire lui, de l'interprétation géométrique des classes vue à la section 1.1.3. Il part de la décomposition canonique de la matrice des covariances :

$$\boldsymbol{\Sigma}_k = \boldsymbol{S}_k \boldsymbol{\Lambda}_k \boldsymbol{S}_k', \tag{1.9}$$

où $\boldsymbol{\Lambda}_k$ désigne la matrice diagonale des valeurs propres de $\boldsymbol{\Sigma}_k$ rangées par ordre décroissant, et $\boldsymbol{S}_k$ une matrice orthogonale. Il formalise alors le manque d'information dans certaines directions de l'espace par un aplatissement isotropique des ellipsoïdes associés aux classes. Il suppose en effet pour chaque k, que les $d - d_k$ ($d_k \in \{1, \ldots, d-1\}$) derniers coefficients diagonaux de la matrice $\boldsymbol{\Lambda}_k$ valent tous $\beta_k \in \mathbb{R}_*^+$. Ce modèle décompose l'aléa dans la classe k, dans les différentes directions spectrales de $\boldsymbol{\Sigma}_k$. Il suppose que la donnée vit essentiellement dans l'espace engendré par les d_k premiers vecteurs colonnes de $\boldsymbol{S}_k$. La loi des vecteurs conditionnels dans [21], se décompose selon :

$$(\mathbf{X}|Z_k = 1) \sim \boldsymbol{\mu}_k + \boldsymbol{S}_k(\boldsymbol{\Lambda}_k - \beta_k \boldsymbol{I})^{1/2} \boldsymbol{U}_k + \sqrt{\beta_k} \boldsymbol{e}_k, \tag{1.10}$$

où $\boldsymbol{U}_k$ et $\boldsymbol{e}_k$ désignent deux vecteurs gaussiens de $\mathbb{R}^d$, centrés, réduits et indépendants.

(1.8) et (1.10) montrent que les mélanges gaussiens de Bouveyron et de McLachlan dédiés à la grande dimension, reposent sur une structure commune :

$$\textit{aléa}(k) = \boldsymbol{\mu}_k + \textit{aléa principal}(k) + \textit{aléa résiduel}(k). \tag{1.11}$$

Lorsque la dimension intrinsèque des données conditionnelles, d_k, est homogène et vaut δ, les mélanges gaussiens de [21] sont des mélanges de Factor Analyzers particuliers. En effet, dans ce cas, chaque matrice $\boldsymbol{\Lambda}_k$ peut s'écrire $\boldsymbol{\Lambda}_k = \mathrm{diag}(\alpha_{1,k} + \beta_k, \ldots, \alpha_{\delta,k} + \beta_k, \beta_k, \ldots, \beta_k)$ avec $\alpha_{j,k} > 0$ ($j \in \{1, \ldots, \delta\}$). La matrice des covariances dans la classe k s'écrit alors : $\boldsymbol{\Sigma}_k = \boldsymbol{B}_k \boldsymbol{B}_k' + \beta_k \boldsymbol{I}$, où $\boldsymbol{B}_k$ est la matrice $d \times \delta$ dont le j^{e} vecteur colonne est le j^{e} vecteur colonne de $\boldsymbol{S}_k$ multiplié par $\sqrt{\alpha_{j,k}}$. Un modèle de [21] qui suppose d_k homogène, est donc strictement équivalent à un mélange de Factor Analyzers dont les erreurs conditionnelles sont isotropiques. Ce mélange particulier de Factor Analysers est appelé dans la littérature, un mélange de Principal Component Analyzers. Il est étudié en tant que tel par S. Roweis (1998) ([94]) et utilisé pour la compression d'images par C.M. Bishop (1999) ([19]).

Les mélanges gaussiens de [21] et les mélanges de Factor Analyzers sont très proches. D'une part ils reposent sur la même structure (1.11), et d'autre part ils possèdent un sous-modèle commun : le mélange de Principal Component Analyzers. Mais en dépit de leurs ressemblances, ces deux types de mélanges gaussiens, dédié chacun aux données de grande dimension, recèlent une différence importante. Dans un mélange de Factor Analyzers le modèle gaussien conditionnel est invariant à la modification des unités de mesure des données. Dans les mélanges de C. Bouveyron ([21]), le modèle gaussien conditionnel n'est pas invariant à la modification des unités de mesure.
Le modèle d'un vecteur aléatoire $\mathbf{X}$ de $\mathbb{R}^d$ est invariant à la modification des unités de mesure si pour toute matrice $\boldsymbol{D} \in \mathbb{R}^{d \times d}$ diagonale, définie, positive, le modèle de $\boldsymbol{D}\mathbf{X}$

est le même que celui de $\mathbf{X}$. Si le vecteur gaussien $(\mathbf{X}|Z_k = 1)$ admet un Factor Analyzer $\boldsymbol{U}_k$ alors $\boldsymbol{U}_k$ est aussi un Factor Analyzer de $(\boldsymbol{D}\mathbf{X}|Z_k = 1)$. En revanche si les dernières valeurs propres de la matrice $\boldsymbol{\Sigma}_k$ sont égales, les dernières valeurs propres de la matrice des covariances du vecteur $(\boldsymbol{D}\mathbf{X}|Z_k = 1)$ ne le sont pas forcément. En effet, la dilatation covariable par covariable d'un vecteur gaussien modifie le spectre de sa matrice des covariances, ainsi que la dimension des sous-espaces propres de cette matrice.

Au chapitre 4, nous verrons qu'une des formes de la classification simultanée repose sur un lien affine entre populations. Or cet avatar de la classification simultanée requiert que le modèle des populations conditionnelles soit invariant à la modification des unités de mesure. Donc, il est exclu d'étendre la classification simultanée aux modèles de C. Bouveyron. La classification simultanée de données de grande dimension, basée sur un lien affine entre populations, ne peut être envisagée si les populations sont des modèles gaussiens de [21].
En revanche, il est possible de définir un classifieur simultané de grande dimension, basé sur des mélanges de Factor Analyzers.

Remarque. Parmi les modèles dédiés à la grande dimension, les mélanges de Factor Analyzers ne sont peut-être pas les seuls à pouvoir s'étendre à la classification simultanée. C. Maugis (2008) ([82]) ainsi que C. Maugis et al. (2008) [83] proposent une procédure originale de sélection de variables, basée sur un choix de modèle. Il ne s'agit pas, comme dans les méthodes traditionnelles de sélection de variables, d'éliminer les variables que l'on considère comme les moins informatives, mais d'assigner à chacune des variables disponibles un rôle dans la classification. Les variables sont classées en trois catégories : informative, redondante ou indépendante, et le rôle joué par chaque variable (c'est à dire la répartition des variables dans la partition précédente) est déterminé comme un choix de modèle. Peut-être serait-il intéressant, lorsqu'on classifie plusieurs échantillons dans un espace de grande dimension, de combiner un lien stochastique entre covariables de même sens (Chapître 4) et partition des variables comme dans [82].

1.1.5 Quelques écueils de la modélisation par des mélanges

Identifiabilité

Lorsqu'on infère le paramètre d'un mélange (et de n'importe quel modèle en général) il doit être identifiable, c'est à dire que deux valeurs distinctes dans l'espace $\boldsymbol{\Theta}$ du paramètre ne doivent pas conduire à la même fonction de densité :

$$\forall(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \boldsymbol{\Theta}, [f(\bullet; \boldsymbol{\theta}_1) = f(\bullet; \boldsymbol{\theta}_2)] \Rightarrow [\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2]\,. \tag{1.12}$$

Nous estimerons le paramètre d'un mélange dans les paragraphes suivants, par maximum de vraisemblance. Or les propriétés classiques des estimateurs du maximum de vraisemblance (invariance fonctionnelle, convergence (normalité asymptotique), etc.) ne sont valables que pour des modèles identifiables (L'identifiablilité du paramètre n'est pas suffisante à ces propriétés mais elle nécessaire.).

Un mélange de lois uniformes par exemple n'est pas identifiable en général. Il suffit pour s'en convaincre, d'observer qu'en dimension 1, le mélange des deux composantes uniformes $\mathscr{U}_{[0,\theta]}(\bullet)$ (densité de la loi uniforme sur l'intervalle $[0,\theta]$) et $\mathscr{U}_{[\theta,1]}(\bullet)$, affectées respectivement des poids θ et $1-\theta$, ne dépend pas de $\theta \in]0,1[$:

$$\forall(\theta_1,\theta_2) \in]0,1[^2 : \theta_1 \mathscr{U}_{[0,\theta_1]}(\bullet) + (1-\theta_1)\mathscr{U}_{[\theta_1,1]}(\bullet) = \theta_2 \mathscr{U}_{[0,\theta_2]}(\bullet) + (1-\theta_2)\mathscr{U}_{[\theta_2,1]}(\bullet). \quad (1.13)$$

Tout mélange gaussien est identifiable à une permutation près du label de ses composantes. Considérons deux mélanges gaussiens de K composantes chacun, notons $\boldsymbol{\theta}_{j,k} = (\pi_{j,k}, \boldsymbol{\mu}_{j,k}, \boldsymbol{\Sigma}_{j,k})$ le paramètre (le poids, le centre et la matrice des covariances) de la composante k dans le mélange j ($k \in \{1,\dots,K\}, j \in \{1,2\}$), et supposons que :

$$\sum_{k=1}^{K} \pi_{1,k}\phi_d(\bullet; \boldsymbol{\mu}_{1,k}, \boldsymbol{\Sigma}_{1,k}) = \sum_{k=1}^{K} \pi_{2,k}\phi_d(\bullet; \boldsymbol{\mu}_{2,k}, \boldsymbol{\Sigma}_{2,k}) \quad (1.14)$$

($\phi_d(\bullet; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ désigne la fonction de densité de la loi normale d-dimensionnelle, centrée en $\boldsymbol{\mu}$ et de matrice des covariances $\boldsymbol{\Sigma}$.). Alors il existe une permutation $\boldsymbol{s}$ de $\mathscr{S}_K$ (groupe symétrique sur $\{1,\dots,K\}$) telle que, pour tout $k \in \{1,\dots,K\}$: $\boldsymbol{\theta}_{1,k} = \boldsymbol{\theta}_{2,\boldsymbol{s}(k)}$.

De façon générale, un mélange (gaussien ou non) ayant au moins deux composantes, est toujours, au mieux, identifiable à une permutation près du label de ses composantes.

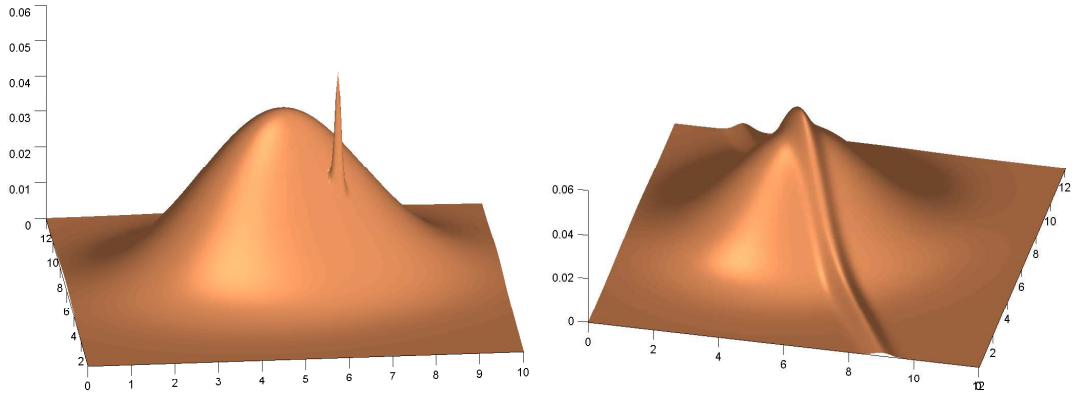
Composantes dues à un effet d'échantillonnage

Les figures 1.6(a) et 1.6(b) représentent chacune un mélange gaussien de deux composantes (sur $\mathbb{R}^2$), dont la valeur du paramètre est indiquée dans le tableau 1.2. Dans chacun de ces mélanges, le centre des populations conditionnelles est proche, et les composantes se distinguent essentiellement par le volume c'est à dire $|\boldsymbol{\Sigma}_1|^{1/2}$ et $|\boldsymbol{\Sigma}_2|^{1/2}$. Les composantes de tels mélanges sont difficilement interprétables. L'hétérogénéité qu'ils modélisent est peu commune et même peu plausible dans la plupart des jeux de données réelles. Ces figures illustrent la possibilité, quand on modélise des données par un mélange (pas forcément gaussien), de 'fausse composante' (appelée 'spurious local maximizer' par [85]) : le modèle est satisfaisant du point de vue du critère de choix de modèle employé, mais certaines composantes sont dues à un effet d'échantillonnage et ne relatent pas une structure interprétable de la donnée.

Dégénérescence (dans les mélanges gaussiens)

Lorsqu'on estime le paramètre d'un modèle (quel qu'il soit) par maximum de vraisemblance, il est souhaitable que sa fonction de vraisemblance soit majorée. Or ce n'est pas toujours le cas. La dégénérescence se caractérise par une valeur infinie de la vraisemblance en un point de l'espace du paramètre.

La fonction de vraisemblance d'un mélange gaussien par exemple n'est pas majorée. Elle tend vers l'infini lorsqu'une des composantes est centrée en une donnée, et que le



(a) (b)

FIGURE 1.6: *Deux cas de 'fausse composante' dans un mélange gaussien.*

Paramètre	π_1	π_2	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_1$	$\boldsymbol{\Sigma}_2$
FIG. 1.6(a)	0.999	0.001	$\begin{pmatrix}5\\7\end{pmatrix}$	$\begin{pmatrix}6.2\\5.8\end{pmatrix}$	$\boldsymbol{S}_{\pi/6}\begin{pmatrix}4&0\\0&3\end{pmatrix}\boldsymbol{S}'_{\pi/6}$	$\begin{pmatrix}1/200&0\\0&1/200\end{pmatrix}$
FIG. 1.6(b)	0.95	0.05	$\begin{pmatrix}5\\7\end{pmatrix}$	$\begin{pmatrix}5\\7\end{pmatrix}$	$\boldsymbol{S}_{\pi/6}\begin{pmatrix}4&0\\0&3\end{pmatrix}\boldsymbol{S}'_{\pi/6}$	$\boldsymbol{S}_{2\pi/3}\begin{pmatrix}20&0\\0&0.05\end{pmatrix}\boldsymbol{S}'_{2\pi/3}$

TABLE 1.2: *Valeur du paramètre des mélanges gaussiens de* FIG. 1.6(a) *et* 1.6(b). $\boldsymbol{S}_\theta$ *désigne la matrice* 2×2 *de la rotation d'angle* θ.

volume de cette composante tend vers 0.

Kiefer et Wolfowitz (1956) ([69]) indiquent des conditions suffisantes pour éviter ce phénomène dans le cas général. Mais on se focalise surtout, depuis, sur la dégénérescence dans les mélanges gaussiens (à cause de l'intérêt que suscite ce type de modèle).

Jusqu'à une époque très récente les solutions qui étaient apportées à la dégénérescence dans les mélanges gaussiens reposaient, comme dans [62], sur une majoration imposée à la fonction de vraisemblance, par une restriction directe (arbitraire) de l'espace du paramètre. C. Biernacki (2010) ([14]) change de point de vue en présentant la dégénérescence dans les mélanges gaussiens non plus comme un défaut de la fonction de vraisemblance, mais en formalisant le fait que certaines partitions de la donnée sont inaptes à estimer le paramètre des composantes. Il propose alors une stratégie destinée à éviter une explosion de la fonction de vraisemblance, basée sur le calcul d'un risque de dégénérescence en tout point de l'espace du paramètre.

1.1.6 Les modèles de mélange en classification

Méthodologie de la classification automatique

L'objectif

La classification automatique (appelée également classification non supervisée ou clustering) se donne pour objectif d'estimer une partition en K groupes, sous jacente à un échantillon $\mathbf{x}_1, \ldots, \mathbf{x}_n$ d'un espace $\mathcal{X}$. La partition recherchée peut être représentée par un tableau disjonctif complet $n \times K$ défini par : $z_{i,k} = 1$ si l'individu $\mathbf{x}_i$ est dans le groupe k et $z_{i,k} = 0$ sinon.

Le modèle

On suppose que la donnée $\mathbf{x}_1, \ldots, \mathbf{x}_n$ provient d'un mélange tel que (1.1) et on assimile les groupes recherchés aux composantes du mélange : le vecteur $\mathbf{z}_i$ indique la composante du mélange ayant généré $\mathbf{x}_i$ (voir 1.1).

Cette hypothèse semble à la fois naturelle et anodine mais elle marque une rupture entre les méthodes traditionnelles et historiques de classification et la classification à base de mélanges. Elle donne en effet une définition probabiliste aux groupes recherchés (puisque l'on fait une hypothèse sur la loi de la variable catégorielle $\mathbf{Z}$ dont les réalisations $\mathbf{z}_i$ sont issues) alors que dans les autres méthodes de classification, la partition estimée n'est (au mieux) que le résultat de l'optimisation d'un critère.

Les couples $(\mathbf{z}_i, \mathbf{x}_i)$ sont donc des réalisations de vecteurs aléatoires $(\mathbf{Z}_i, \mathbf{X}_i)$ indépendants et identiquement distribués au couple $(\mathbf{Z}, \mathbf{X})$ défini en 1.1.1.
Déterminer une partition de l'échantillon revient alors à estimer les vecteurs conditionnels $(\mathbf{Z}_i | \mathbf{X}_i = \mathbf{x}_i)$.

La méthode

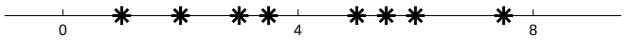
Dans un contexte paramétrique c'est à dire quand on suppose l'échantillon issu de (1.2), une estimation $\hat{\boldsymbol{\theta}}$ du paramètre (obtenue généralement par maximum de vraisemblance, mais d'autres estimateurs sont envisageables) conduit à une partition probabiliste $\boldsymbol{t} = (t_{i,k})_{\substack{i=1,\ldots,n \\ k=1,\ldots,K}}$ de la donnée, dont le coefficient :

$$t_{i,k} = \hat{\pi}_k f_k(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k) / \sum_{j=1}^{K} \hat{\pi}_j f_j(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_j) \tag{1.15}$$

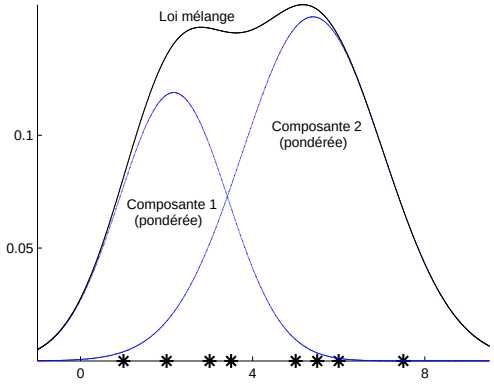
représente la probabilité conditionnelle pour le point $\mathbf{x}_i$, de provenir de la composante k. On alloue ensuite chaque point $\mathbf{x}_i$ par Maximum A Posteriori (MAP) au groupe correspondant à la plus grande des probabilités conditionnelles $t_{i,k}$ $(k = 1, \ldots, K)$:

$$\hat{z}_{i,k} = 1 \Leftrightarrow t_{i,k} \geq t_{i,j}; j = 1, \ldots, K. \tag{1.16}$$

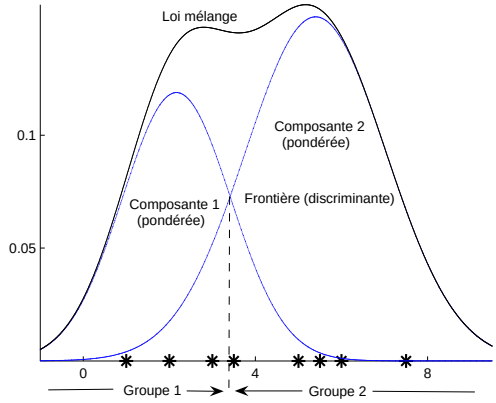
La figure 1.7 représente les deux étapes (inférence d'un modèle et classement des données) de la classification par modèle de mélange.



(a) Données à classifier



(b) Inférence d'un mélange (EM)



(c) Classement des données (MAP)

FIGURE 1.7: *Classification par modèle de mélange dans un contexte non supervisé.*

L'estimation de la partition, dans cette procédure, est subordonnée à celle du paramètre (Le paramètre est affecté par plug-in au classement par MAP). Estimer le paramètre par maximum de vraisemblance - on parle dans ce cas d'une approche mélange de la classification - revient alors à considérer l'adéquation du modèle comme primordiale, et la classification des données comme subalterne.

G. Celeux et G. Govaert (1992) ([27]) proposent de façon alternative, d'estimer conjointement le paramètre et la partition, de manière à considérer l'objectif de classification comme aussi important que la modélisation - on parle alors d'approche classification. Ils définissent ainsi les algorithmes CEM et SEM (sur lesquels nous reviendrons à la section 1.2.4) qui intégrent la partition recherchée au critère à optimiser.

Règle de classement et frontière discriminante

Généralités

La règle du MAP affecte chaque point de la donnée à la composante dont il a le plus de chances (a posteriori) de provenir. Son principe semble donc intuitivement satisfaisant, mais sa justification mathématique est plus convaincante encore. En effet, le MAP détermine (conditionnellement au paramètre estimé) l'application $r : \mathcal{X} \to \{1, \dots, K\}$ qui minimise l'erreur moyenne de classement :

$$1 - E_{(\mathbf{X},\mathbf{Z})} \left[\mathbf{Z}_{r(\mathbf{X})} \right], \tag{1.17}$$

et classe les données selon cette règle (appelée la règle de Bayes).

La règle de Bayes $r_{\mathcal{B}}$ affecte un point $\boldsymbol{x} \in \mathcal{X}$ au groupe k dont la densité (pondérée

du poids de la classe) est la plus grande en $\boldsymbol{x}$:

$$[r_{\mathcal{B}}(\boldsymbol{x}) = k] \Rightarrow [\forall j \in \{1, \ldots, K\}, \pi_k f_k(\boldsymbol{x}; \boldsymbol{\alpha}_k) \geq \pi_j f_j(\boldsymbol{x}; \boldsymbol{\alpha}_j)]\,. \tag{1.18}$$

Lorsque $\mathcal{X} = \mathbb{R}^d$ la fontière discriminante est constituée du bord (au sens de l'intersection des adhérences) des sous-ensembles $r_{\mathcal{B}}^{-1}(k)$, $(k = 1, \ldots, K)$ de $\mathbb{R}^d$.

Frontière discriminante dans le cas gaussien

Dans le cas de deux classes gausiennes par exemple, la frontière discriminante est constituée des points $\boldsymbol{x}$ de $\mathbb{R}^d$ tels que :

$$(\boldsymbol{x} - \boldsymbol{m})' \left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right)(\boldsymbol{x} - \boldsymbol{m}) - 2\boldsymbol{v}' \left(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\right)(\boldsymbol{x} - \boldsymbol{m}) - \boldsymbol{v}' \left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right)\boldsymbol{v} = -2\ln\left[(\pi_2/\pi_1)(|\boldsymbol{\Sigma}_1|/|\boldsymbol{\Sigma}_2|)^{1/2}\right], \tag{1.19}$$

où $\boldsymbol{m} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ et $\boldsymbol{v} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/2$.

La frontière entre les deux classes gaussiennes est donc une quadrique dont la nature dépend à la fois du rang de la matrice symétrique $\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}$ et de sa signature.

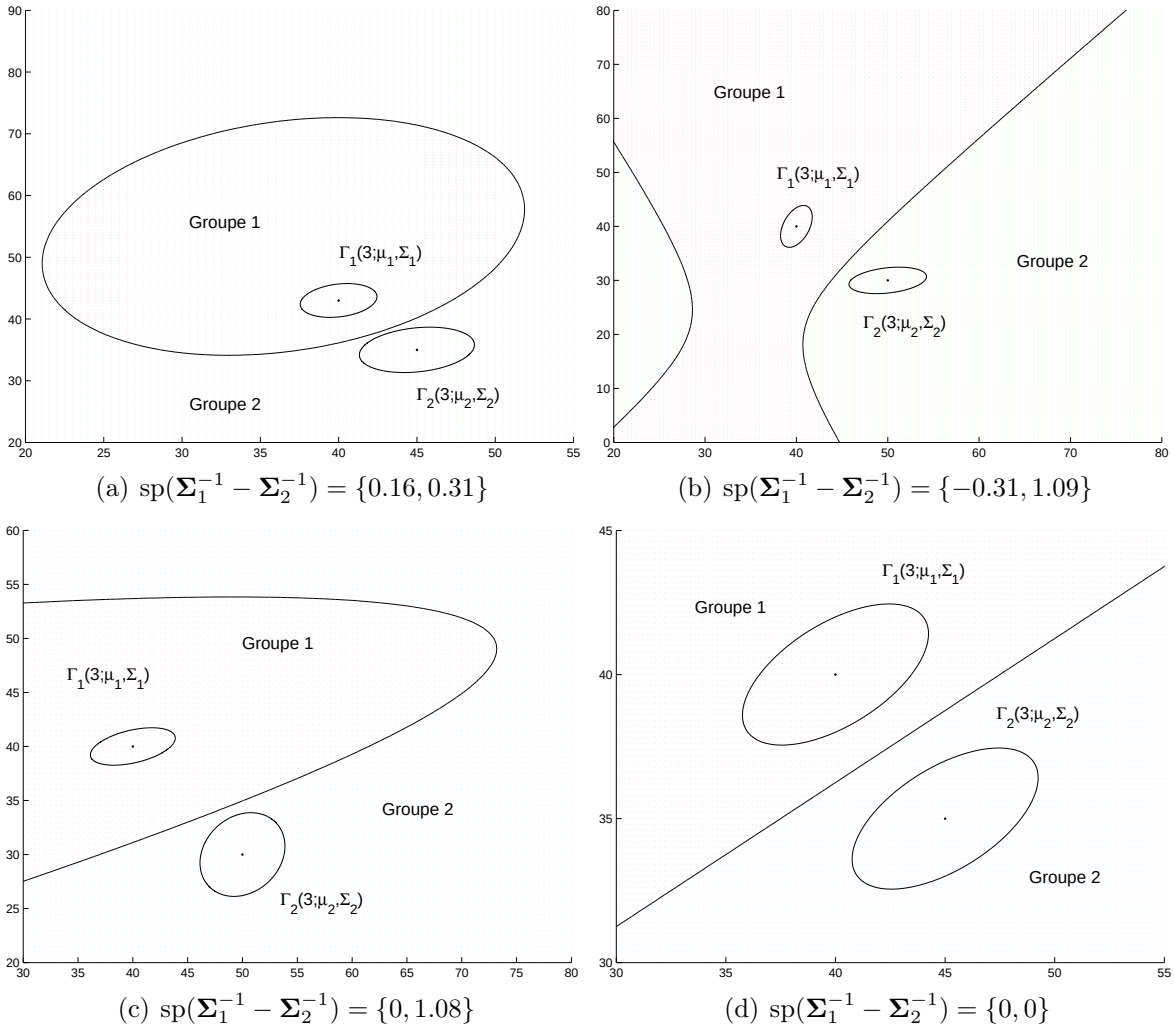
En dimension 2 par exemple, quatre cas sont à envisager.
Lorsque $\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}$ est régulière et possède deux valeurs propres de même signe, la frontière discriminante est une ellipse (FIG. 1.8(a)).
Quand $\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}$ est régulière avec deux valeurs propres de signes différents, la frontière discriminante est une hyperbole (FIG. 1.8(b)).
Lorsque $\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}$ est dégénérée mais possède une valeur propre non nulle, la frontière discriminante est une parabole (FIG. 1.8(c)).
Enfin lorsque les matrices de covariances $\boldsymbol{\Sigma}_1$ et $\boldsymbol{\Sigma}_2$ sont égales (le mélange est homoscédastique), la quadrique (1.19) est une forme linéaire et la frontière discriminante est une droite (FIG. 1.8(d)).

En dimension supérieure à 2, la nature de la frontière discriminante ne dépend toujours que de $\mathrm{sp}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})$, le spectre de la matrice symétrique $\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}$. On pourra se reporter à [52] pour une classification exhaustive des quadriques en dimension 3.

Lorsque le nombre de classes est supérieur à 2 (et toujours dans un mélange gaussien) la frontière discriminante est composée de sections de quadriques, comme le montre la figure 1.9 en dimension 2.

Nombre de groupes

Dans le contexte général de la classification automatique, on ignore tout des groupes recherchés, jusqu'à leur interprétation et leur nombre.



(a) $\mathrm{sp}(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1}) = \{0.16, 0.31\}$

(b) $\mathrm{sp}(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1}) = \{-0.31, 1.09\}$

(c) $\mathrm{sp}(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1}) = \{0, 1.08\}$

(d) $\mathrm{sp}(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1}) = \{0, 0\}$

FIGURE 1.8: *Frontière discriminante dans un mélange gaussien à deux composantes*

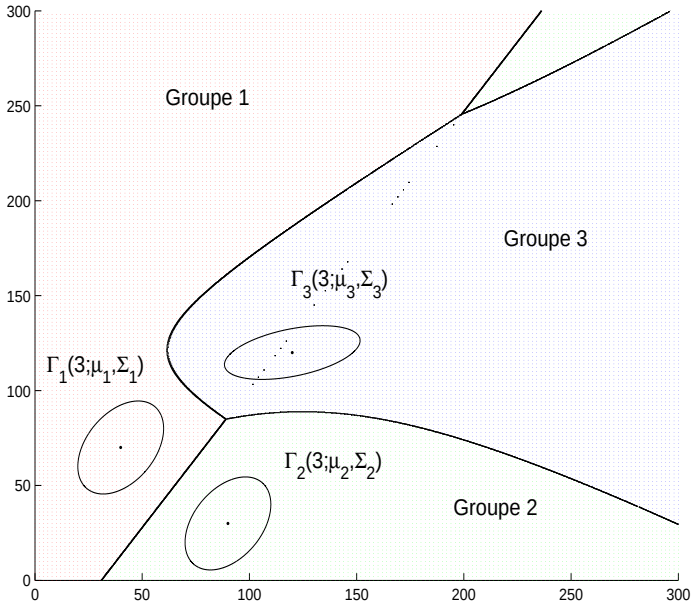


FIGURE 1.9: *Frontière discriminante dans un mélange gaussien à trois composantes.*

Les méthodes traditionnelles de classification peinent autant à définir ce qu'est un groupe qu'à justifier les procédures qui permettent d'en déterminer le nombre. Les groupes obtenus dans une méthode de k-means par exemple, résultent de la minimisation de l'inertie intra-classe. Or ce critère d'inertie est d'emblée minimal si l'on suppose autant de groupes que de points. On peut émettre une objection de même nature lorsque A. Hardy (1996) ([56]) et A. Hardy et N. Kasoro (2009) [57] proposent de déterminer le nombre des groupes en minimisant la mesure globale de leur enveloppe convexe.

L'objectif de ce paragraphe est de montrer que lorsqu'on classifie des données à l'aide de mélanges, on bénéficie pour déterminer le nombre de groupes, des outils puissants de l'inférence paramétrique et, en particulier, que les critères de choix de modèle constituent à cet effet un outil adapté et largement répandu.

Jusqu'ici nous avons admis implicitement une correspondance biunivoque entre les groupes recherchés, les classes déterminées par la règle de Bayes, et les composantes du mélange.
La règle de Bayes établit une partition de l'espace et nous avons supposé que chaque composante du mélange donnait lieu à une classe de la partition. Or il se peut que la partition compte strictement moins de classes que le mélange ne comporte de composantes. En effet lorsque le poids d'une composante est petit, les points qu'elle génère sont répartis par la règle de classement dans des classes issues de composantes de poids plus important. La notion probabiliste de composante est donc plus fine que la notion ensembliste de classe.
Nous avons admis d'autre part que chaque classe de la partition déterminait un groupe dans la donnée. Nous verrons que la correspondance un-à-un entre les classes de la partition et les groupes dans la donnée, bien que classique, est remise en cause par J.P. Baudry (2009) ([9]) au profit d'une définition plus subtile des groupes.
Pour notre part, dans ce travail, nous continuerons de supposer que chaque composante dans la donnée correspond à une classe de la partition, et que chaque classe de la partition correspond à une composante du mélange. Ainsi déterminer le nombre de groupes dans la donnée revient à choisir l'ordre du mélange, c'est à dire le nombre de ses composantes.

Une première approche très répandue (voir [40], [86], [99]) consiste en un test d'hypothèse : il s'agit de tester l'hypothèse nulle d'un nombre de groupes K_0, contre l'hypothèse alternative d'un nombre de groupes supérieur, K_1. La statistique de test est le log-rapport de la vraisemblance maximale du paramètre du modèle sous chacune des hypothèses (que l'on note $LRTS$ pour Likelihood Ratio Test Statistic). Cette méthode connaît un succès important, mais elle présente plusieurs inconvénients.
D'une part, dans ce contexte particulier - la détermination de l'ordre d'un mélange -, on ne connait pas (en général) la distribution asymptotique de la $LRTS$. Le paramètre sous l'hypothèse nulle, correspond en effet à une valeur particulière et non identifiable du paramètre sous l'hypothèse alternative. La $LRTS$ ne suit donc pas, asymtotiquement, une loi de χ^2, comme c'est le cas dans les conditions habituelles de régularité.
Puisqu'on ne connait pas la loi de la $LRTS$, on ne dispose pas de seuil de décision associé au test. McLachlan ([85] p. 193, ou [86]) propose d'obtenir par re-échantillonnage, la p-valeur du test. Mais rien ne garantit que le modèle postulé (le modèle des données

sous l'hypothèse nulle) soit proche du vrai modèle et la p-valeur obtenue par bootstrap peut être subordonnée de façon excessive au modèle postulé sous l'hypothèse nulle.
Enfin, dans l'esprit, le LRT sert essentiellement à tester s'il existe ou non une structure dans la donnée ($K_0 = 1$ groupe contre $K_1 = 2$ groupes). Lorsque le nombre de groupes possibles dans la donnée, est plus important, cette méthode oblige à combiner les tests d'hypothèses ce qui induit une incertitude sur le test global.
Dans tous les cas le LRT dissocie la détermination de l'ordre d'un mélange du choix de la spécificité de ses composantes. Pour cette raison nous lui préférerons dans ce travail l'usage d'un critère de choix de modèle, qui combine la détermination de l'ordre et le choix d'une hypothèse de parcimonie.

Une seconde approche consiste à voir la détermination de l'ordre d'un mélange comme faisant partie du choix d'un modèle.
Envisager pour chaque mélange en lice, différentes valeurs de K, élargit la famille de modèles disponibles. Choisir l'un de ces modèles selon un critère, c'est choisir l'ordre du mélange et, à travers cet ordre, le nombre de groupes sous-jacents à la donnée. Ainsi la détermination du nombre de groupes dans la donnée est un choix de modèle. La détermination du nombre de groupes dépend donc, comme l'estimation du paramètre, du paradigme choisi : classification ou modélisation.

Si l'on accorde une importance prépondérante à la modélisation des données (approche mélange) on choisit un mélange selon un critère comme BIC ou AIC qui ne prend pas en compte l'objectif de classification et juge uniquement de l'adéquation (en un certain sens) du modèle à la donnée. Si le critère est BIC, le nombre de groupes retenu correspond alors au nombre le plus probable de sous-populations homogènes dans la donnée. Cette approche orientée vers la recherche de sous populations homogènes est sujette aux fluctuations d'échantillonnage et peut conduire à surestimer le nombre de groupes dans des échantillons de petite taille. Il se peut en effet, dans un cas de fausse composante décrit à la section 1.1.5, que BIC marque une préférence pour un mélange à K composantes alors que le vrai modèle en comporte $K-1$. D'autre part, si les composantes du mélange ne modélisent pas convenablement les données conditionnelles, BIC aura tendance à surestimer l'ordre du mélange.

L'approche classification permet au nombre de groupes retenu, d'être robuste à la fois aux fluctuations d'échantillonnage et à une modélisation inappropriée des données conditionnelles.
Les critères de choix de modèle relatifs à cette approche (on en trouvera plusieurs exemples dans [12]) ont tous la même structure : un terme de vraisemblance permet de juger de la qualité d'un modèle, et un terme pénalisant un mélange aux classes imbriquées, permet de juger le pouvoir de classification du modèle. Le critère ICL ([16]) par exemple, est un des critères de l'approche classification. Il s'agit de BIC pénalisé par le logarithme de la probabilité de la partition :

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_{i,k} \ln t_{i,k}. \tag{1.20}$$

La pénalité (1.20) est d'autant plus proche de 0 que les probabilités conditionnelles $t_{i,k}$

sont proches de 0 ou de 1, c'est à dire que les classes de la partition sont séparées.
Ainsi un mélange comportant une 'fausse composante' (voir section 1.1.2) pourrait être écarté par ICL s'il est retenu par BIC.

Les critères de choix de modèle relatifs à une approche classification souffrent tous du même inconvénient : ils ne sont pas consistants pour le nombre de groupes (contrairement à BIC). Lorsque la donnée comporte deux groupes par exemple, mais que les groupes sont imbriqués, ces critères ont tendance, asymtotiquement, à ne déceler aucune structure dans la donnée. Les critères classification sont donc utiles pour déterminer le nombre de groupes dans la donnée dans certains cas particuliers (lorsqu'on classifie des échantillons de petite taille par exemple, ou lorsqu'on doute de la normalité des données conditionnelles). Mais il ne peuvent pas faire mieux que BIC dans les autres cas. Nous verrons en effet à la section 1.3 que BIC est consistant pour le nombre de groupes, sous certaines conditions.

Notons l'existence d'autres méthodes, plus heuristiques, de choix de l'ordre d'un mélange, qui ne se rapportent à aucune des deux catégories précédentes. Celle du coude de vraisemblance développée par Cutler et Windham ([29]) par exemple, repose sur une observation : la vraisemblance maximale est une fonction croissante de l'ordre d'un mélange mais on observe une rupture du taux d'augmentation de la vraisemblance maximale au voisinage du vrai nombre de groupes.
Le taux d'augmentation de la vraisemblance maximale en deçà duquel on juge avoir atteint le bon nombre de composantes, est arbitraire chez [29]. C. Biernacki (1999) ([13]) propose de le déterminer dans le cas de mélanges gaussiens de façon plus objective grâce à la notion probabiliste de précision sur les données.

J.P. Baudry et al. (2008) ([9]) et J.P. Baudry ([8]) proposent une méthode heuristique de choix du nombre de groupes, qui tire parti à la fois des avantages de BIC et d'ICL, mais qui ne consiste pas, à proprement parler, en un choix de modèle.
Leur méthode comporte deux étapes. Dans une première étape, ils choisissent un modèle et l'ordre K d'un mélange, grâce à BIC, selon une approche mélange traditionnelle. Dans une seconde étape ils fusionnent les classes k et k' pour lesquelles le critère :

$$\sum_{i=1}^{n}(t_{i,k}+t_{i,k'})\ln(t_{i,k}+t_{i,k'})-(t_{i,k}\ln t_{i,k}+t_{i,k'}\ln t_{i,k'}) \tag{1.21}$$

est maximal. Comme la quantité $t_{i,k}+t_{i,k'}$ représente pour chaque point $\mathbf{x}_i$, la probabilité qu'il provienne des classes k ou k', la seconde étape revient à choisir les groupes à fusionner de manière à assurer un gain d'entropie maximal, à la partition floue des données.
Cette procédure invite à reconsidérer la notion de groupe. Une classe reste une région de l'espace obtenue par application de la règle de classement. Mais un groupe est alors le résulat d'une procédure de fusion des classes, basée sur l'optimisation d'un critère d'entropie.

Généralement en classification automatique, on ignore le nombre de groupes à déterminer. Il arrive cependant qu'une procédure de clustering cherche à retrouver dans

l'échantillon, une classification externe à la donnée. Dans ce cas le nombre de groupes comme leur interprétation, sont guidés par la partition externe. A la section 4.1.5.1 par exemple, les variables biométriques mesurées sur des échantillons de Puffins cendrés, ont été choisies dans l'intention de déterminer le sexe des oiseaux. Cette perspective de sexage incite à fixer $K = 2$, puis à interpréter les groupes obtenus comme des mâles et des femelles. Mais une telle interprétation est toujours sujette à caution : les groupes obtenus pourraient, en effet, correspondre à un clivage des oiseaux selon une autre variable catégorielle que le sexe.

De façon générale en classification non supervisée, il faut être prudent dans l'interprétation des groupes inférés, même lorsque le choix du nombre de groupes est guidé par une partition externe à la donnée.
Notons à ce propos la proposition récente de J.P. Baudry pour rationaliser l'interprétation de la partition estimée. L'usage d'un critère de choix de modèle est toujours guidé, nous l'avons vu, par un objectif. Prenant appui sur ce principe, J.P. Baudry ([8], chapitre 8) propose d'incorporer l'interprétabilité des classes inférées, à l'objectif du choix de modèle. Il définit un critère heuristique $SICL$ (Supervised Integrated Completed Likelihood) qui relate à la fois l'adéquation du modèle, et la proximité des groupes estimés à une partition connue des données.

1.2 Estimation du paramètre

1.2.1 L'algorithme EM (généralités)

Objectif de l'algorithme EM

On considère un vecteur aléatoire $\mathbf{X}$ à valeurs dans un espace probabilisé $\mathcal{X}$, distribué selon une loi $f(\bullet\,;\boldsymbol{\theta})$ et dont le paramètre $\boldsymbol{\theta}$ appartient à un espace $\boldsymbol{\Theta}$. On souhaite estimer par maximum de vraisemblance le paramètre de cette loi.

On suppose qu'un échantillon $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ provient de façon i.i.d de cette loi c'est à dire que les $\mathbf{x}_i$ sont des réalisations de variables aléatoires $\mathbf{X}_i$ indépendantes et identiquement distribuées à $\mathbf{X}$.

Dans de nombreux cas - lorsque $f(\bullet\,;\boldsymbol{\theta})$ est une fonction de densité de la famille exponentielle par exemple - la log-vraisemblance du paramètre :

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(\mathbf{x}_i;\boldsymbol{\theta}), \tag{1.22}$$

admet un maximum dont l'argument est explicite et constitue l'estimateur $\hat{\boldsymbol{\theta}}$ recherché.

Mais le plus souvent (dans les modèles de mélange en particulier) les équations de vraisemblance $\dfrac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \ln L(\boldsymbol{\theta}) = 0$ n'ont pas de solution explicite.

L'algorithme EM (Espérance-Maximisation) permet alors - dans certains cas seulement, qui dépendent de la structure de la donnée - d'augmenter (1.22) de façon itérative et de converger vers l'un de ses maxima locaux.

La structure de donnée manquante

Considérons (i) un vecteur aléatoire $\mathbf{Y}$ à valeurs dans un espace probabilisé $\mathcal{Y}$ dont la loi de probabilité $g(\bullet\,;\boldsymbol{\theta})$ ne dépend que du paramètre $\boldsymbol{\theta}$ de la loi de $\mathbf{X}$ et (ii) une application surjective et explicite $\phi : \mathcal{Y} \to \mathcal{X}$.

Notons que les lois de $\mathbf{X}$ et $\mathbf{Y}$ sont alors liées par la relation :

$$\forall \boldsymbol{x} \ \in \mathcal{X} : f(\boldsymbol{x};\boldsymbol{\theta}) = \int_{\mathcal{Y}_{\boldsymbol{x}}} g(\boldsymbol{y};\boldsymbol{\theta})\mathrm{d}\boldsymbol{y}, \text{ où } \mathcal{Y}_{\boldsymbol{x}} = \{\boldsymbol{y} \in \mathcal{Y}; \phi(\boldsymbol{y}) = \boldsymbol{x}\}. \tag{1.23}$$

Supposons qu'un échantillon $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ de $\mathcal{Y}$ vérifie $\phi(\mathbf{y}_i) = \mathbf{x}_i$ pour tout i et que les $\mathbf{y}_i$ soient des réalisations de variables aléatoires $\mathbf{Y}_i$ indépendantes identiquement distribuées à $\mathbf{Y}$.

La log-vraisemblance de $\boldsymbol{\theta}$ peut être calculée :

– sur la donnée observée $\mathbf{x}$ selon (1.22) (log-vraisemblance observée de $\boldsymbol{\theta}$),

– ou sur la donnée complète $\mathbf{y}$ selon :

$$L_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln g(\mathbf{y}_i;\boldsymbol{\theta}). \tag{1.24}$$

(1.24) désigne alors la log-vraisemblance complétée de $\boldsymbol{\theta}$.

Pour tout $\boldsymbol{x} \in \mathcal{X}$ et tout $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ on définit la loi conditionnelle $k(\bullet\,;\boldsymbol{x},\boldsymbol{\theta}) = g(\bullet\,;\boldsymbol{\theta})/f(\boldsymbol{x};\boldsymbol{\theta})$. En pratique la donnée de l'échantillon $\mathbf{y}$ n'est pas totalement connue (l'étape E de EM revient à en estimer la partie manquante) mais chaque variable aléatoire $(\mathbf{Y}_i|\mathbf{X}_i = \mathbf{x}_i)$ est distribuée selon $k(\bullet\,;\mathbf{x}_i,\boldsymbol{\theta})$.

L'essence d'EM

L'algorithme EM repose sur une propriété de l'écart entropique :

$$H(\boldsymbol{\theta}',\boldsymbol{\theta}) = E\left[\sum_{i=1}^{n} \ln k(\mathbf{Y}_i;\mathbf{X}_i,\boldsymbol{\theta}')|\mathbf{x}_i,\boldsymbol{\theta}\right], \tag{1.25}$$

lorsque $\boldsymbol{\theta}$ et $\boldsymbol{\theta}'$ sont deux valeurs quelconques du paramètre.

Pour tout $\boldsymbol{x} \in \mathcal{X}$, l'écart de Kullback entre les deux lois de probabilité $k(\bullet\,; \boldsymbol{x}, \boldsymbol{\theta})$ et $k(\bullet\,; \boldsymbol{x}, \boldsymbol{\theta}')$ est positif :

$$KL\left[k(\bullet\,; \boldsymbol{x}, \boldsymbol{\theta}); k(\bullet\,; \boldsymbol{x}, \boldsymbol{\theta}')\right] \geq 0. \tag{1.26}$$

De (1.26) on déduit que pour tout i :

$$E(\ln k(\mathbf{Y}_i; \mathbf{X}_i, \boldsymbol{\theta}')|\mathbf{x}_i, \boldsymbol{\theta}) \leq E(\ln k(\mathbf{Y}_i; \mathbf{X}_i, \boldsymbol{\theta})|\mathbf{x}_i, \boldsymbol{\theta}) \tag{1.27}$$

et que :

$$H(\boldsymbol{\theta}', \boldsymbol{\theta}) \leq H(\boldsymbol{\theta}, \boldsymbol{\theta}), \tag{1.28}$$

avec égalité lorsque $\boldsymbol{\theta}' = \boldsymbol{\theta}$ (et uniquement dans ce cas).

Or (1.25) mesure la différence entre la log-vraisemblance (complétée) attendue du paramètre $\boldsymbol{\theta}'$ lorsqu'on suppose la donnée complète générée par $\boldsymbol{\theta}$:

$$Q(\boldsymbol{\theta}', \boldsymbol{\theta}) = E\left[\sum_{i=1}^{n} \ln g(\mathbf{Y}_i; \boldsymbol{\theta}')|\mathbf{x}_i, \boldsymbol{\theta}\right], \tag{1.29}$$

et sa log-vraisemblance observée $L(\boldsymbol{\theta}')$:

$$H(\boldsymbol{\theta}', \boldsymbol{\theta}) = Q(\boldsymbol{\theta}', \boldsymbol{\theta}) - L(\boldsymbol{\theta}'), \tag{1.30}$$

On en déduit :

$$L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}) = \underbrace{[Q(\boldsymbol{\theta}', \boldsymbol{\theta}) - Q(\boldsymbol{\theta}, \boldsymbol{\theta})]}_{\Delta_Q(\boldsymbol{\theta}', \boldsymbol{\theta})} - \underbrace{[H(\boldsymbol{\theta}', \boldsymbol{\theta}) - H(\boldsymbol{\theta}, \boldsymbol{\theta})]}_{\Delta_H(\boldsymbol{\theta}', \boldsymbol{\theta})}. \tag{1.31}$$

Pour augmenter la vraisemblance observée (1.22) d'une valeur $\boldsymbol{\theta}$ du paramètre, il suffit donc (en supposant que la donnée a été générée par $\boldsymbol{\theta}$) d'augmenter sa vraisemblance attendue (1.29).

En effet : $\Delta_H(\boldsymbol{\theta}', \boldsymbol{\theta}) < 0$ si $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$ et $\Delta_H(\boldsymbol{\theta}', \boldsymbol{\theta}) = 0$ si $\boldsymbol{\theta}' = \boldsymbol{\theta}$ (d'après (1.28)). Donc si $\Delta_Q(\boldsymbol{\theta}', \boldsymbol{\theta}) \geq 0$ alors $L(\boldsymbol{\theta}') \geq L(\boldsymbol{\theta})$.

Structure de l'algorithme EM

EM est un algorithme itératif qui, en partant d'une valeur initiale $\boldsymbol{\theta}^0$ du paramètre, définit dans $\boldsymbol{\Theta}$, une suite finie de valeurs du paramètre $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^M$ correspondant à une suite croissante de valeurs de la vraisemblance : pour tout $0 \leq m \leq M-1$: $L(\boldsymbol{\theta}^m) \leq L(\boldsymbol{\theta}^{m+1})$.

L'itération m ($0 \leq m \leq M-1$) de EM est constituée dans cet ordre, des deux étapes suivantes :

- Etape E : On calcule $Q(\boldsymbol{\theta}', \boldsymbol{\theta}^m)$, la log-vraisemblance attendue de $\boldsymbol{\theta}'$, en supposant la donnée générée par $\boldsymbol{\theta}^m$, le paramètre en cours à l'itération m.

- Etape M : On détermine $\boldsymbol{\theta}^{m+1} = \text{argmax}_{\boldsymbol{\theta}'} Q(\boldsymbol{\theta}', \boldsymbol{\theta}^m)$.

Pour tout m, la vraisemblance attendue du paramètre $\boldsymbol{\theta}^{m+1}$ est donc supérieure à celle de $\boldsymbol{\theta}^m$ (en supposant la donnée générée par $\boldsymbol{\theta}^m$ dans les deux cas) :

$$Q(\boldsymbol{\theta}^{m+1}, \boldsymbol{\theta}^m) \geq Q(\boldsymbol{\theta}^m, \boldsymbol{\theta}^m) \tag{1.32}$$

La vraisemblance de $\boldsymbol{\theta}^{m+1}$ est donc supérieure à celle de $\boldsymbol{\theta}^m$ puisque : $\Delta_H(\boldsymbol{\theta}^{m+1}, \boldsymbol{\theta}^m) \leq 0$.

Ainsi $L(\boldsymbol{\theta}^m)$ $(m = 0, \ldots, M)$ est une suite croissante de valeurs de la log-vraisemblance.

Points fixes de l'algorithme EM et maxima de la vraisemblance

L'algorithme EM permet de définir dans l'espace du paramètre, une suite croissante de valeurs de la vraisemblance.
Donc si la vraisemblance admet un maximum global unique en $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, alors $\boldsymbol{\theta}^*$ est un point fixe d'EM.
Mais si la vraisemblance admet deux maxima globaux (ou plus) $\boldsymbol{\theta}_1^*$ et $\boldsymbol{\theta}_2^*$, il se peut qu'EM ne converge pas : les points $\boldsymbol{\theta}^m$ $(m = 0, \ldots, M)$ peuvent prendre, à partir d'un certain rang, les valeurs $\boldsymbol{\theta}_1^*$ et $\boldsymbol{\theta}_2^*$ alternativement.
Aussi, quand l'unicité du maximum de la vraisemblance n'est pas démontrée, rien n'assure que l'algorithme EM converge.

Dempster et al. (1977) [37] montrent que si la vraisemblance est bornée et qu'il existe un scalaire $\lambda > 0$ tel que pour tout m :

$$|Q(\boldsymbol{\theta}^{m+1}, \boldsymbol{\theta}^m) - Q(\boldsymbol{\theta}^{m+1}, \boldsymbol{\theta}^m)| \geq \lambda \|\boldsymbol{\theta}^{m+1} - \boldsymbol{\theta}^m\|_2^{\,2}, \tag{1.33}$$

alors EM converge vers un point fixe situé dans l'adhérence de $\boldsymbol{\Theta}$.
Mais en pratique la condition (1.33) est difficile à établir, et généralement, la convergence d'EM vers un point fixe $\boldsymbol{\theta}_0$, est admise.

Une croissance lente de la vraisemblance dans les denières itérations d'EM fera penser que la dernière valeur du paramètre estimée, $\boldsymbol{\theta}^M$, est proche de $\boldsymbol{\theta}_0$. Mais cette interprétation est source d'erreur. Il se peut en effet que $\boldsymbol{\theta}^M$ soit proche d'un point selle de la vraisemblance et non de $\boldsymbol{\theta}_0$.

Dempster et al. (1977) ([37]) montrent que si les matrices hessiennes $\mathscr{H}_{Q(\boldsymbol{\theta},\boldsymbol{\theta}^m)|\boldsymbol{\theta}=\boldsymbol{\theta}^{m+1}}$ $(m = 0, \ldots, M-1)$ sont définies négatives et que leurs spectres sont dans un domaine (indépendant de m) contenant 0, alors $\boldsymbol{\theta}_0$ est bien un maximum de la vraisemblance.

Mais rien n'assure alors que ce maximum soit global et non local.
La stratégie habituelle pour déterminer le maximum de la vraisemblance, consiste à lancer l'algorithme EM en partant de différentes valeurs initiales du paramètre et à retenir parmi les valeurs estimées du paramètre, celle de plus grande vraisemblance.

1.2.2 Application au paramètre d'un mélange

L'algorithme EM permet en particulier d'estimer le paramètre d'un mélange.

Considérons un mélange de K lois paramétriques définies sur un espace probabilisé $\mathcal{X}$: $f(\bullet\,;\boldsymbol{\theta})$ conforme à (1.2). Lorsqu'aucune contrainte ne lie les composantes $\boldsymbol{\alpha}_k$ $(k = 1, \ldots, K)$ du paramètre (c'est à dire lorsque les éventuelles contraintes portant sur chaque $\boldsymbol{\alpha}_k$ sont intrinsèques à la composante k), estimer le paramètre $\boldsymbol{\theta}$ du mélange est possible dès lors que l'estimation (par MV) de chacun des paramètres $\boldsymbol{\alpha}_k$ l'est.

Supposons comme en 1.1.6 qu'un échantillon $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ de $\mathcal{X}$ provient de façon i.i.d de (1.2), et notons $\mathbf{z}_i$ $(i \in \{1, \ldots, n\})$ la donnée manquante indiquant la composante d'origine du point $\mathbf{x}_i$. On rappelle que les couples aléatoires $(\mathbf{Z}_i, \mathbf{X}_i)$ sont supposés indépendants, que pour tout i la k^{e} composante $Z_{i,k}$ du vecteur $\mathbf{Z}_i$ vaut 1 (et les autres 0) avec probabilité π_k et que : $(\mathbf{X}_i | Z_{i,k} = 1) \sim f_k(\bullet\,;\boldsymbol{\alpha}_k)$.

La contribution d'un point $(\mathbf{z}_i, \mathbf{x}_i)$ de la donnée complète, à la vraisemblance de $\boldsymbol{\theta}$, s'écrit dans ce contexte :

$$g(\mathbf{z}_i, \mathbf{x}_i; \boldsymbol{\theta}) = \prod_{k=1}^{K} \left[\pi_k f_k(\mathbf{x}_i; \boldsymbol{\alpha}_k)\right]^{z_{i,k}}, \tag{1.34}$$

et la log-vraisemblance calculée sur les données complètes (1.24) s'écrit :

$$L_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbf{z}_{i,k} \{\ln \pi_k + \ln f_k(\mathbf{x}_i; \boldsymbol{\alpha}_k)\}. \tag{1.35}$$

En considérant les données manquantes $\mathbf{z}_i$ dans leur version aléatoire et la donnée comme étant générée par $\boldsymbol{\theta}$, le calcul de la log-vraisemblance attendue de $\boldsymbol{\theta}'$, (1.29), revient à celui des coefficients :

$$t_{i,k} = E[Z_{i,k} | \mathbf{x}_i, \boldsymbol{\theta}] = \pi_k f_k(\mathbf{x}_i; \boldsymbol{\alpha}_k) / \sum_{j=1}^{K} \pi_j f_j(\mathbf{x}_i; \boldsymbol{\alpha}_j). \tag{1.36}$$

Afin de simplifier les notations on écrit $t_{i,k}$ au lieu de $t_{i,k}(\boldsymbol{\theta})$. Il faut cependant garder à l'esprit que les coefficients $t_{i,k}$ sont toujours subordonnés à une valeur $\boldsymbol{\theta}$ du paramètre

(dont on suppose qu'elle a généré la donnée).

La log-vraisemblance attendue du paramètre $\boldsymbol{\theta}'$ est alors :

$$Q(\boldsymbol{\theta}', \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{i,k} \{\ln \pi'_k + \ln f_k(\mathbf{x}_i; \boldsymbol{\alpha}'_k)\}, \tag{1.37}$$

et si aucune contrainte ne lie les paramètres $\boldsymbol{\alpha}'_k$ entre eux, l'optimum de (1.37) est atteint pour $\pi'_k = \sum_{i=1}^{n} t_{i,k}/n$ et par des solutions de :

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}'_k} \sum_{i=1}^{n} t_{i,k} \ln f_k(\mathbf{x}_i; \boldsymbol{\alpha}'_k) = 0 \ ; \ k = 1, \ldots, K. \tag{1.38}$$

Un coefficient $t_{i,k}$ représente la probabilité (conditionnelle) pour le point $\mathbf{x}_i$ de provenir de la composante k du mélange (en supposant la donnée générée par $\boldsymbol{\theta}$).
Les coefficients de la matrice $\boldsymbol{t} = (t_{i,k})_{\substack{i=1,\ldots,n \\ k=1,\ldots,K}}$ sont compris entre 0 et 1 ($0 \leq t_{i,k} \leq 1$) et leur somme par ligne vaut toujours 1 ($\sum_{k=1}^{K} t_{i,k} = 1$). La matrice $\boldsymbol{t}$ (comme toute autre matrice dont les coefficients vérifieraient les deux propriétés précédentes) est une partition floue de la donnée. C'est la partition attendue de l'échantillon lorsqu'on le suppose généré par $\boldsymbol{\theta}$. Le coefficient $t_{i,k}$ peut également être vu dans (1.37) comme la contribution du point $\mathbf{x}_i$ à la vraisemblance du paramètre de la classe k.

Ainsi les étapes E et M de l'algorithme EM sont spécifiques dans les modèles de mélanges. L'étape E de l'itération $m+1$ consiste à calculer les probabilités conditionnelles $t_{i,k}$ selon (1.36) en supposant la donnée générée par le paramètre en cours $\boldsymbol{\theta}^m$. L'étape M qui suit consiste à déterminer la valeur $\boldsymbol{\theta}^{m+1}$ du paramètre qui optimise (1.37) par rapport à $\boldsymbol{\theta}'$.

Remarque. Estimer le paramètre d'un mélange (dont les composantes sont algébriquement indépendantes) requiert uniquement de savoir estimer le paramètre de ses composantes dans le cas de données labellées (coefficients $z_{i,k}$ connus). L'étape M d'EM revient en effet à estimer le paramètre de chaque composante k ($k = 1, \ldots, K$) en considérant que chaque point $\mathbf{x}_i$ contribue à la vraisemblance du paramètre $\boldsymbol{\alpha}_k$ proportionnellement à $t_{i,k}$.
Ainsi, un mélange dont on peut estimer le paramètre en analyse discriminante, pourra toujours être inféré dans un contexte non supervisé.

D'après cette présentation d'EM dans les mélanges, la partition floue $\boldsymbol{t}$ calculée à chaque étape E, joue un rôle technique et subalterne du point de vue de l'interprétation : c'est une béquille sur laquelle s'appuie l'étape M pour optimiser la vraisemblance complétée du paramètre et augmenter la vraisemblance. Le paragraphe 1.2.3 dissocie la partition floue, de la partition probabiliste $\boldsymbol{t}$ induite par le paramètre $\boldsymbol{\theta}$; il confère ainsi à la partition floue, une importance comparable à celle du paramètre dans l'algorithme EM.

1.2.3 Une interprétation d'EM pour les mélanges : la formule d'Hathaway

R.J. Hathaway (1986) ([60]) montre que l'on peut interpréter EM dans le cas particulier des mélanges, comme un algorithme optimisant un critère qui porte à la fois sur une partition floue des données et le paramètre du mélange. Nous verrons que ce critère est très utile pour (i) interpréter l'objectif sous-jacent à des méthodes de classification courantes et (ii) envisager de nouvelles procédures de classification.

Rappelons qu'une partition floue (on parle aussi de partition probabiliste) $\boldsymbol{w} = (w_{i,k})_{\substack{i=1,\dots,n \\ k=1,\dots,K}}$ relate pour chaque point d'un échantillon $\mathbf{x}_1, \dots, \mathbf{x}_n$, sa probabilité d'appartenir à chacun des K groupes d'une partition. Elle est caractérisée par :

- chaque coefficient $\boldsymbol{w}_{i,k}$ est compris entre 0 et 1.
- la somme de ses coefficients par colonne vaut 1 : pour tout i, $\sum_{k=1}^{K} \boldsymbol{w}_{i,k} = 1$.
- la somme de ses coefficients par ligne est non nulle : pour tout k, $\sum_{i=1}^{n} \boldsymbol{w}_{i,k} \neq 0$.

Son entropie :

$$E(\boldsymbol{w}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} w_{i,k} \ln w_{i,k} \tag{1.39}$$

est une mesure de la séparation des classes probabilistes. Elle est positive et majorée par $n \ln K$ ($0 \leq E(\boldsymbol{w}) \leq n \ln K$). Les classes (probabilistes) déterminées par $\boldsymbol{w}$ sont séparées lorsque $E(\boldsymbol{w})$ est proche de 0 ; elles sont mélangées lorsqu'au contraire $E(\boldsymbol{w})$ est proche de $n \ln K$.

Lorsque l'échantillon $\mathbf{x}_1, \dots, \mathbf{x}_n$ est issu (ou supposé tel) d'un mélange paramétrique (1.2), à toute valeur $\boldsymbol{\theta}$ du paramètre correspond (via la procédure de MAP) une partition floue $\boldsymbol{t}$, c'est à dire un objet de même nature que $\boldsymbol{w}$. Il semble donc naturel de chercher à mesurer la proximité du paramètre $\boldsymbol{\theta}$ et de la partition floue $\boldsymbol{w}$.

A cet effet l'écart de Kullback :

$$KL(\boldsymbol{w}, \boldsymbol{t}) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{i,k} \ln w_{i,k} - w_{i,k} \ln t_{i,k}, \tag{1.40}$$

entre les deux partitions probabilistes $\boldsymbol{w}$ et $\boldsymbol{t}$ semble tout indiqué.

L'interprétation d'EM que propose Hathaway, repose sur la dissociation de la partition floue $\boldsymbol{w}$ de la donnée, et de la partition $\boldsymbol{t}$ relative au paramètre $\boldsymbol{\theta}$.

On définit la log-vraisemblance d'une valeur $\boldsymbol{\theta}$ du paramètre, complétée de la partition $\boldsymbol{w}$ de la donnée par :

$$LC(\boldsymbol{\theta}, \boldsymbol{w}) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{i,k}\{\ln \pi_k + \ln f_k(\mathbf{x}_i; \boldsymbol{\alpha}_k)\} \tag{1.41}$$

et le critère :

$$C(\boldsymbol{\theta}, \boldsymbol{w}) = LC(\boldsymbol{\theta}, \boldsymbol{w}) + E(\boldsymbol{w}). \tag{1.42}$$

On remarque que si $\boldsymbol{w}$ est la vraie partition $(z_{i,k})_{\substack{i=1,\dots,n \\ k=1,\dots,K}}$ de la donnée, (1.41) coïncide avec la log-vraisemblance complétée de $\boldsymbol{\theta}$ (1.35) et que que si $\boldsymbol{w}$ est la partition floue $(t_{i,k})_{\substack{i=1,\dots,n \\ k=1,\dots,K}}$ induite par une valeur $\boldsymbol{\theta}$ du paramètre, $LC(\boldsymbol{\theta}', \boldsymbol{w})$ est la log-vraisemblance attendue de $\boldsymbol{\theta}'$ (1.37).

Ainsi l'étape M d'EM consiste à optimiser $C(\boldsymbol{\theta}', \boldsymbol{t})$ par rapport à $\boldsymbol{\theta}'$ (lorsque $\boldsymbol{t}$ est la partition floue induite par la valeur courante $\boldsymbol{\theta}$ du paramètre) .

Or on montre facilement que :

$$C(\boldsymbol{\theta}, \boldsymbol{w}) = L(\boldsymbol{\theta}) - KL(\boldsymbol{w}, \boldsymbol{t}) \tag{1.43}$$

où :

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i; \boldsymbol{\alpha}_k) \right) \tag{1.44}$$

est la log-vraisemblance du paramètre $\boldsymbol{\theta}$ (voir [60]).

L'égalité (1.43) montre que l'étape E d'EM détermine la partition $\boldsymbol{w}$ la plus proche au sens de l'écart de Kullback (1.40), de celle, $\boldsymbol{t}$, induite par la valeur courante $\boldsymbol{\theta}$ du paramètre. En l'absence de contrainte sur la partition floue $\boldsymbol{w}$, la solution est $\boldsymbol{w} = \boldsymbol{t}$.

La fomule d'Hathaway (qui consiste en l'égalité de (1.42) et (1.43)) montre que l'algorithme EM alternativement (i) détermine la partition floue $\boldsymbol{w}$ la plus proche au sens de l'écart de Kullback, de celle induite par la valeur $\boldsymbol{\theta}$ du paramètre et (ii) augmente la vraisemblance du paramètre, complétée de cette partition floue.

Mais la formule d'Hathaway est d'une portée plus large encore que l'interprétation de l'algorithme EM. Elle permet par exemple d'envisager d'autres procédures de classification.

1.2.4 Algorithmes dérivés d'EM

Depuis l'article initial de Dempster, Laird et Rubin (1977) ([37]), l'algorithme EM donne lieu à une littérature foisonnante, y compris dans le domaine particulier des mélanges. Nous ne nous attardons pas ici, sur les nombreux algorithmes EM propres à des mélanges spécifiques, décrits par [85], ni sur les algorithmes dérivés comme ECM ou ECME (voir [88]) dont la vocation est d'accélerer EM. Nous avons montré que dans EM, la partition probabiliste de la donnée joue un rôle auxiliaire : c'est un point d'appui qui sert à augmenter itérativement la vraisemblance en optimisant la vraisemblance complétée attendue. Or il existe une famille d'algorithmes dérivés d'EM, qui donnent à la partition floue et au paramètre, une place de même importance. L'algorithme $\tilde{\text{E}}$M que nous définirons à la section 5.3 fait partie des algorithmes dérivés d'EM.

L'algorithme Classification EM (CEM)

EM recherche alternativement l'optimum de (1.42) par rapport à $\boldsymbol{\theta}$, le paramètre du mélange, et $\boldsymbol{w}$, une partition probabiliste de la donnée. Nous avons vu que le paramètre $\hat{\boldsymbol{\theta}}$ estimé en un point stationnaire de l'algorithme, correspond au maximum de la vraisemblance du modèle. La partition de la donnée se déduit alors par MAP de la partition floue $\hat{\boldsymbol{w}}$. G. Celeux et G. Govaert (1992) ([27]) proposent de modifier l'étape E d'EM en cherchant $\boldsymbol{w}$ non plus dans $\mathscr{W}$, l'ensemble de toutes les partitions probabilistes possibles, mais sur le bord de $\mathscr{W}$. Il s'agit donc à la place de l'étape E, d'estimer la partition de la donnée, par MAP de la partition floue relative au paramètre en cours. Cette procédure a pour effet de réduire l'importance de l'adéquation du modèle dans la classification, et au contraire d'augmenter celle de la partition. Le paramètre ainsi estimé n'est plus le maximum de la vraisemblance du modèle. Mais le couple $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{w}})$ vers lequel converge l'algorithme, est un point stationnaire (un maximum sous certaines conditions d'ordre deux) du critère de vraisemblance classifiante (1.42) (voir [27]).
G. Govaert montre ([49], p.269) que CEM permet d'interpréter de façon probabiliste, une grande variété d'algorithmes de classification reposant sur l'optimisation d'un critère géométrique. Un algorithme de k-means basé sur une métrique euclidienne par exemple, est équivalent à CEM dans un mélange gaussien homoscédastique aux matrices de covariances sphériques ([49], p.273).

L'algorithme Stochastic EM (SEM)

Il arrive qu'EM converge lentement, qu'il converge vers un point selle de la vraisemblance, ou encore que le paramètre estimé dépende beaucoup du paramètre initial (notamment lorsque les composantes du mélange se chevauchent fortement). Pour éviter ces inconvénients, G. Celeux et J. Diebolt (1985) ([25]) proposent de rendre stochastique l'étape E d'EM. La partition $\boldsymbol{w}$ est choisie de sorte que chaque point $\mathbf{x}_i$ appartienne au groupe k avec probabilité $t_{i,k}$. Le vecteur $\boldsymbol{w}_i = (w_{i,1}, \ldots, w_{i,K})$ est tiré au sort lors de l'étape SE (pour Stochastic E), selon une loi multinomiale d'ordre 1 et de paramètre $\boldsymbol{t}_i = (t_{i,1}, \ldots, t_{i,K})$. Ainsi la partition $\boldsymbol{t}$ relative à la valeur en cours du paramètre, tient lieu d'un paramètre qui permet de re-probabiliser à chaque étape SE, l'espace $\mathscr{W}$ de la partition. La suite des valeurs du paramètre ainsi définie, est une chaîne de Markov dont G. Celeux et J. Diebolt montrent dans [26] qu'elle est homogène, ergodique et convergente. Le paramètre est estimé comme moyenne des paramètres inférés par SEM, ce qui permet par exemple d'éviter les points selles de la vraisemblance.

Ainsi l'interprétation d'EM selon la formule d'Hathaway rend à la partition probabiliste des données un rôle égal en importance à celui du paramètre. Elle permet d'envisager de nouveaux algorithmes de classification dérivés d'EM. Ces algorithmes reposent tous sur une définition nouvelle de l'espace $\mathscr{W}$ dans lequel on recherche la partition floue pour optimiser le critère (1.43). L'algorithme $\tilde{\text{E}}$M que nous proposons à la section 5.3 fait partie de cette famille d'algorithmes : dans un contexte de classification simultanée, il permettra de transférer une contrainte, du paramètre à la partition floue.

1.3 Choix d'un modèle

Pourquoi plusieurs modèles ?

Modéliser des données hétérogènes par un mélange paramétrique permet de placer la classification dans le contexte mathématiquement très riche de l'inférence statistique. Un mélange paramétrique permet par ailleurs (i) d'interpréter les groupes de façon générative et (ii) d'interpréter les propriétés des composantes comme des qualités intrinsèques des données conditionnelles.
Mais en contrepartie, l'hypothèse paramétrique possède un coût : le biais du modèle estimé d'une part, et sa sensibilité aux fluctuations d'échantillonnage d'autre part.

Faire concourir plusieurs modèles de mélanges $\mathscr{M}_j$ $(j = 1, \ldots, m)$ qui diffèrent par leur complexité, par le nombre de leurs composantes, par la spécificité de leurs composantes, permet de :

- trouver un compromis entre le biais du modèle estimé et sa variabilité,
- mettre en concurrence plusieurs hypothèses quant au nombre de groupes sous-jacents à la donnée,
- mettre en concurrence plusieurs hypothèses quant aux propriétés de chaque groupe.

Le choix de l'un de ces modèles est alors subordonné à l'objectif visé par la classification.

Critères de choix de modèles liés à un objectif

Comme nous l'avons vu à la section 1.1.6, la classification de données basée sur un mélange, procède en deux étapes : l'inférence du modèle et le classement des données.
Dans certains cas, on accorde une importance prépondérante à la modélisation des données et l'on considére que leur classification est un sous-produit de leur modélisation. Cette position repose sur l'idée que le meilleur classifieur moyen est le modèle le plus proche du vrai modèle. Le critère de choix de modèle employé sera alors un critère général comme AIC ou BIC. Ces critères jugent de l'adéquation du modèle à la donnée, et ne rendent compte de la qualité de la partition obtenue que pour un nombre infini de données.
Or à taille d'échantillon finie, il est possible d'obtenir une partition acceptable des données avec un mélange éloigné du vrai modèle. C'est le cas lorsque les donnnées conditionnelles sont bien séparées par exemple. On peut alors, sans abandonner l'objectif de modélisation, prendre en compte l'objectif de classification dans le choix du modèle. On emploiera dans ce cas un critère spécifique comme ICL. Ce critère évalue la qualité du modèle d'une part. Mais il rend compte également de l'interprétabilité des classes, en favorisant les modèles conduisant à des groupes bien séparés.

Choix d'un critère général

AIC, un critère fréquentiste

Le critère AIC de H. Akaïke ([2], [3]) cherche à minimiser l'écart de Kullback moyen entre le vrai modèle et le modèle postulé. Il repose plus précisément sur la minimisation de la déviance moyenne :

$$2E_{\mathbf{x},\mathbf{x}'}\left[\ln p(\mathbf{x}') - \ln p(\mathbf{x}';\hat{\boldsymbol{\theta}})\right], \tag{1.45}$$

où $\mathbf{x}$ et $\mathbf{x}'$ désignent, dans leur version aléatoire, deux échantillons indépendants et issus du même modèle, $p(\mathbf{x}')$ la vraie loi jointe des données de $\mathbf{x}'$, et $p(\mathbf{x}',\hat{\boldsymbol{\theta}})$ la loi jointe postulée des données de $\mathbf{x}'$, dont le paramètre $\hat{\boldsymbol{\theta}}$ est estimé par maximum de vraisemblance sur les données de $\mathbf{x}$.
Lorsque la vraie loi est dans la famille de modèles candidats, une approximation de (1.45) est donnée par :

$$AIC = \ell(\hat{\boldsymbol{\theta}};\mathbf{x}) - \nu, \tag{1.46}$$

où $\boldsymbol{\nu}$ désigne la taille du paramètre $\boldsymbol{\theta}$ et $\ell(\hat{\boldsymbol{\theta}};\mathbf{x})$, sa log-vraisemblance maximale.

Mais d'une part l'usage d'AIC suppose que l'un des modèles de la famille considérée est le vrai modèle, hypothèse qui est contestable. D'autre part AIC ne permet de choisir que parmi une collection de modèles emboités. Enfin, AIC n'est pas consistant puisqu' il choisit asymptotiquement parmi des modèles emboités, le plus complexe avec une probabilité non nulle.

BIC, un critère bayésien

Le critère BIC cherche, lui, le modèle $\mathscr{M}$ dont la probabilité conditionnelle $p(\mathscr{M};\mathbf{x})$ est la plus grande. Lorsqu'on suppose équiprobables les modèles en compétition (on dit que la loi a priori des modèles est non informative), choisir un modèle selon BIC revient à optimiser la vraisemblance intégrée :

$$p(\mathbf{x};\mathscr{M}) = \int_{\boldsymbol{\Theta}} p(\mathbf{x};\mathscr{M},\boldsymbol{\theta})p(\boldsymbol{\theta};\mathscr{M})\mathrm{d}\boldsymbol{\theta}. \tag{1.47}$$

Mais le calcul de cette intégrale n'est pas explicite en général et les méthodes permettant d'approcher numériquement (1.47) (voir [67] par exemple) ont une efficacité limitée en grande dimension.

En supposant que la loi conditionnelle du paramètre $\boldsymbol{\theta}$ est normale, on montre sous certaines conditions de régularité que la log-vraisemblance intégrée s'écrit :

$$\ln p(\mathbf{x};\mathscr{M}) = \ln p(\hat{\boldsymbol{\theta}};\mathbf{x},\mathscr{M}) - (\nu/2)\ln n + \mathscr{O}_p(1/\sqrt{n}). \tag{1.48}$$

Cette égalité repose sur un développement à l'odre 2 de la fonction à intégrer dans (1.47), au voisinage du maximum de la vraisemblance $\hat{\boldsymbol{\theta}}$.

L'égalité (1.48) justifie alors de choisir un modèle en minimisant le critère :

$$BIC = -\ln p(\hat{\boldsymbol{\theta}}; \mathbf{x}, \mathscr{M}) + (\nu/2)\ln n. \tag{1.49}$$

Il n'est pas nécessaire lorsqu'on choisit un modèle selon BIC, de supposer que le vrai modèle est dans la famille considérée (comme c'était le cas pour AIC). D'autre part, les modèles candidats $\mathscr{M}_j$ $(j = 1, \ldots, m)$ peuvent ne pas être emboîtés (ils devaient l'être pour AIC). Enfin BIC est un critère consistant asymptotiquement.

Lorsque le paramètre est proche du bord de son espace, (1.48) n'est plus valable. Or choisir le nombre de composantes d'un mélange revient précisément à supposer à partir du modèle le plus complexe que le poids de certaines composantes est nul. B.G. Leroux (1992) ([74]) montre que BIC, asymptotiquement, ne sous-estime pas le nombre de groupes et C. Keribin (2000) ([68]) montre sous certaines conditions, que BIC ne surestime pas non plus le nombre de groupes.

Choix d'un critère spécifique

ICL, un critère robuste

Paradoxalement, dans un contexte de classification, il n'est pas toujours souhaitable qu'un critère de choix de modèle soit consistant. BIC est un bon critère pour choisir un modèle parmi des mélanges parcimonieux, à condition que (i) la donnée soit suffisamment importante et que (ii) les composantes des mélanges permettent de modéliser convenablement les données conditionnelles. Or dans beaucoup de situations pratiques, les conditions (i) et (ii) ne sont pas satisfaites, et lorsque (ii) n'est pas vérifiée par exemple, BIC a tendance à surestimer l'ordre du mélange.

Le critère ICL proposé par C. Biernacki (2000) ([16]) est une alternative robuste à BIC lorsque l'on doute du modèle des données conditionnelles. Sans perdre de vue la modélisation des données, il intègre un objectif de classication (ce que ne fait pas BIC). Il repose sur une approximation non plus de la log-vraisemblance intégrée du paramètre mais de sa log-vraisemblance complétée et intégrée :

$$\ln p(\mathbf{x}, \mathbf{z}; \mathscr{M}) = \ln \int_{\boldsymbol{\Theta}} p(\mathbf{x}, \mathbf{z}; \mathscr{M}, \boldsymbol{\theta}) p(\boldsymbol{\theta}; \mathscr{M}) \mathrm{d}\boldsymbol{\theta}, \tag{1.50}$$

$\mathbf{z} = (z_{i,k})_{\substack{i=1,\ldots,n \\ k=1,\ldots,K}}$ désignant une partition déterministe de $\mathbf{x}$. En imitant l'élaboration de BIC, C. Biernacki (2000) approche (1.50) par l'opposé de :

$$ICL = -\ln p(\hat{\boldsymbol{\theta}}; \mathbf{x}, \mathscr{M}) + (\nu/2)\ln n - \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_{i,k} \ln t_{i,k}, \tag{1.51}$$

où $\hat{\boldsymbol{\theta}}$ et ν désignent respectivement l'estimateur du maximum de vraisemblance et la taille du paramètre $\boldsymbol{\theta}$, et $\hat{z}_{i,k}$, l'estimation de $z_{i,k}$ obtenue par MAP. Le meilleur des modèles est alors celui qui minimise (1.51).

L'équation(1.51) montre que :

$$ICL = BIC + ENT(\hat{\boldsymbol{\theta}}), \tag{1.52}$$

où $ENT(\hat{\boldsymbol{\theta}}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_{i,k} \ln t_{i,k}$ est un terme entropique qui juge de la séparation des classes dans le modèle estimé. $ENT(\hat{\boldsymbol{\theta}})$ est d'autant plus proche de 0 que les groupes sont séparés, et d'autant plus proche de $n \ln K$ que les groupes sont mélangés.

Il existe plusieurs autres versions d'ICL. L'une d'elles consiste par exemple à remplacer $\hat{z}_{i,k}$ par $t_{i,k}$ dans (1.51). Cette variante a pour effet d'augmenter la pénalisation imposée aux composantes qui se recouvrent.

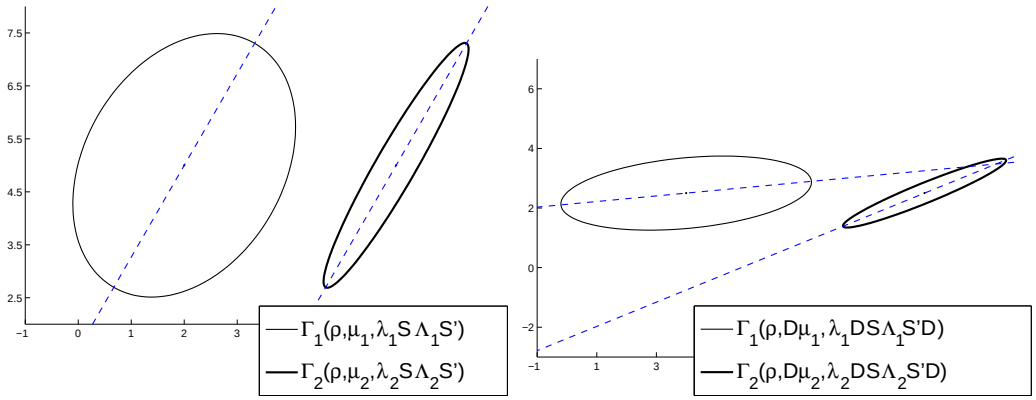
Chapitre 2

Mélanges gaussiens parcimonieux d'interprétation statistique

2.1 Présentation

2.1.1 Incohérences des modèles d'interprétation géométrique

Les mélanges gaussiens de [28] présentés à la section 1.1.3, reposent sur la décomposition spectrale (1.6) des matrices de covariances. Ils permettent d'inférer de façon parcimonieuse, la forme, l'orientation et le volume, trois paramètres géométriques des populations conditionnelles. Mais cette famille de modèles - que nous qualifierons par la suite de géométrique - souffre de plusieurs inconvénients. Par exemple, certains modèles de la famille ne sont pas stables par dilatation des covariables comme l'illustre FIG. 2.1. Le choix d'un modèle géométrique n'est donc pas indifférent au choix des unités de mesure des données. En particulier, le modèle géométrique estimé sur les données, ou sur les données réduites, n'est pas toujours le même.



(a) Des composantes ayant la même orientation ...

(b) ... qui deviennent pour d'autres unités, des composantes d'orientation distincte

FIGURE 2.1: *Un modèle géométrique dont l'orientation est homogène et dont les autres paramètres (volume et forme) sont libres, n'est pas stable par modification des unités de mesure.*

Par ailleurs, certains modèles de [28] ne sont pas stables par projection dans les plans canoniques comme le montre FIG. 2.2. Cette singularité (par rapport à la plupart des modèles probabilistes) empêche la représentation en dimension 2, d'un modèle géométrique de dimension supérieure.

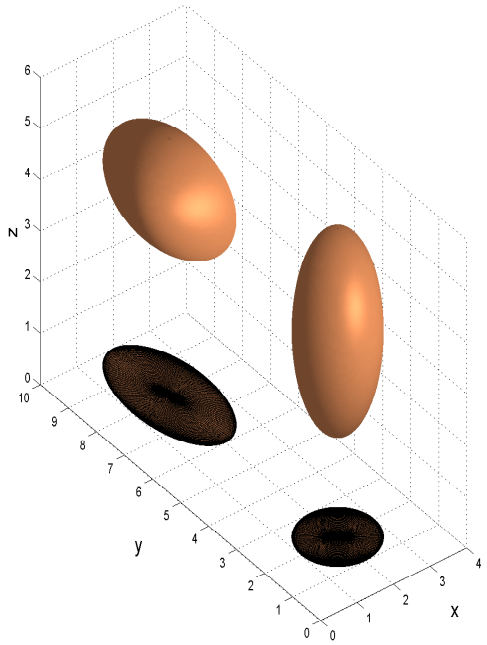


FIGURE 2.2: *Un modèle géométrique dont la forme et de volume sont homogènes mais dont l'orientation est libre, n'est pas stable par projection dans les plans canoniques.*

Le tableau 2.1 indique les modèles parmi ceux de [28], qui ne sont pas invariants à la réduction des données, ainsi que les modèles qui ne sont pas stables par projection dans les différents plans canoniques.

Dans ce chapitre nous définissons de nouveaux modèles de mélanges gaussiens, basés sur une décomposition variance-corrélation des matrices de covariances. La parcimonie de ces modèles porte sur des paramètres conditionnels d'interprétation statistique : l'écart-type et le coefficient de variation d'une variable, la corrélation de deux covariables. Chacun des nouveaux modèles est stable par dilatation des covariables (même si la dilatation n'est pas isotropique). Le modèle inféré est donc indifférent au choix des unités de mesure des données, et indifférent à la réduction des covariables. Ainsi la spécificité des composantes inférées relate davantage une propriété intrinsèque des données conditionnelles qu'un choix particulier des unités de mesure. Par ailleurs, chacun des nouveaux modèles est stable par projection dans n'importe quel plan canonique, ce qui permet de le représenter de façon fidèle en dimension réduite.

modèle	projection	réduction
$[\lambda_k \boldsymbol{S}_k \boldsymbol{\Lambda}_k \boldsymbol{S}_k']$	s	s
$[\lambda \boldsymbol{S}_k \boldsymbol{\Lambda}_k \boldsymbol{S}_k']$	×	s
$[\lambda_k \boldsymbol{S}_k \boldsymbol{\Lambda} \boldsymbol{S}_k']$	×	×
$[\lambda \boldsymbol{S}_k \boldsymbol{\Lambda} \boldsymbol{S}_k']$	×	×
$[\lambda_k \boldsymbol{S} \boldsymbol{\Lambda}_k \boldsymbol{S}']$	×	×
$[\lambda \boldsymbol{S} \boldsymbol{\Lambda}_k \boldsymbol{S}']$	×	×
$[\lambda_k \boldsymbol{S} \boldsymbol{\Lambda} \boldsymbol{S}']$	s	s
$[\lambda \boldsymbol{S} \boldsymbol{\Lambda} \boldsymbol{S}']$	s	s
$[\lambda_k \boldsymbol{B}_k]$	s	s
$[\lambda \boldsymbol{B}_k]$	×	s
$[\lambda_k \boldsymbol{B}]$	s	s
$[\lambda \boldsymbol{B}]$	s	s
$[\lambda_k \boldsymbol{I}]$	s	×
$[\lambda \boldsymbol{I}]$	s	×

TABLE 2.1: *Stabilité des modèles de la famille géométrique par reduction des covariables et par projection dans les plans canoniques. La stabilité du modèle est indiquée par 's' et la non-stabilité par '×'.*

2.1.2 Définition de nouveaux modèles d'interprétation statistique

Décomposition (canonique) variance-corrélation des matrices de covariances

Lorsque les composantes d'un mélange gaussien sont non dégénérées, les matrices de covariances sont symétriques, définies et positives. Chacune d'elles se décompose de façon unique en :

$$\boldsymbol{\Sigma}_k = \boldsymbol{T}_k \boldsymbol{R}_k \boldsymbol{T}_k, \tag{2.1}$$

où $\boldsymbol{T}_k$ est la matrice diagonale des écarts-types conditionnels :

$$\boldsymbol{T}_k(i,j) = \sqrt{\boldsymbol{\Sigma}_k(i,j)} \text{ si } i = j \text{ et } 0 \text{ sinon},$$

et $\boldsymbol{R}_k$, la matrice des corrélations conditionnelles : $\boldsymbol{R}_k = (\boldsymbol{T}_k)^{-1} \boldsymbol{\Sigma}_k (\boldsymbol{T}_k)^{-1}$.

La décomposition (2.1) permet d'envisager plusieurs modèles consistant pour chacun

d'eux, en une combinaison de contraintes significatives et interprétables, sur les matrices $\boldsymbol{T}_k$, sur les matrices $\boldsymbol{R}_k$ et sur les centres $\boldsymbol{\mu}_k$.

Modèle général et modèles parcimonieux

Dans le cas général les matrices $\boldsymbol{T}_k$ $(k = 1, \ldots, K)$ sont juste diagonales, définies, positives. Les écarts-types sont alors supposés libres d'une classe à l'autre. On envisagera que les matrices $\boldsymbol{T}_k$ puissent vérifier la relation : $\forall\ (k, k') : \boldsymbol{T}_{k'} = a_{k,k'}\boldsymbol{T}_k; a_{k,k'} \in \mathbb{R}_+^*$. On fait dans ce cas l'hypothèse que les écarts-types sont transformés de façon isotropique d'une classe à l'autre. Enfin lorsqu'on suppose les matrices $\boldsymbol{T}_k$ égales entre elles $(\boldsymbol{T}_k = \boldsymbol{T})$, on considère les écarts-types conditionnels comme homogènes (invariants d'une classe à l'autre).

Les matrices de corrélations $\boldsymbol{R}_k$ $(k = 1, \ldots, K)$ sont symétriques, définies, positives et leurs coefficients diagonaux valent 1. On peut supposer de plus qu'elles sont égales entre elles $(\boldsymbol{R}_k = \boldsymbol{R})$ c'est à dire que deux covariables quelconques sont toujours identiquement corrélées d'une population conditionnelle à l'autre.

Remarque. Une contrainte sur les matrices $\boldsymbol{T}_k$ postule une propriété intrinsèque aux covariables tandis qu'une contrainte sur les matrices $\boldsymbol{R}_k$ constitue un modèle entre variables. Ainsi, choisir l'un des modèles précédents de matrices $\boldsymbol{T}_k$ et de matrices $\boldsymbol{R}_k$, revient à combiner une hypothèse sur chaque variable et une hypothèse entre covariables.

Le coefficient de corrélation de deux covariables est une normalisation de leur covariance par leurs écarts-types. De même, le coefficient de variation d'une variable est une normalisation de sa moyenne par son écart-type. Comme certains modèles postulent l'homogénéité des corrélations conditionnelles, on envisage également que les coefficients de variation des covariables conditionnelles - représentés par les composantes des vecteurs $\boldsymbol{V}_k = \boldsymbol{T}_k^{-1}\boldsymbol{\mu}_k$ - puissent être constants d'une classe à l'autre $(\boldsymbol{V}_k = \boldsymbol{V})$.

On considère la famille des modèles obtenus par combinaison de contraintes sur les coefficients de variation - libres $(\boldsymbol{V}_k)$ ou égaux $(\boldsymbol{V})$ - les écarts-types - libres $(\boldsymbol{T}_k)$, transformés de façon isotropique $(a_k\boldsymbol{T})$ ou égaux $(\boldsymbol{T})$ - et les corrélations conditionnelles - libres $(\boldsymbol{R}_k)$ ou égales $(\boldsymbol{R})$ - des covariables. Le modèle qui suppose homogènes les trois paramètres conditionnels (écarts-types, coefficients de variation et corrélations) est une loi normale. Nous ne considérons pas ce modèle comme un mélange de la famille ; il est inféré en tant que mélange d'ordre un, lorsqu'on détermine le nombre de groupes dans la donnée. Ainsi, les onze modèles obtenus par combinaison de contraintes sur les trois paramètres statistiques, forment une famille qu'on appellera désormais RTV. Le tableau 2.2 indique la dimension du paramètre gaussien de chacun de ces modèles.

Re-écriture du modèle standard homoscédastique et hétéroscédastique

Le modèle le plus général suppose libres, à la fois les corrélations conditionnelles,

Modèle	Dimension.
$[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$ (général)	$Kd + Kd(d+1)/2$
$[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}]$	$d + Kd(d+1)/2$
$[\boldsymbol{R}_k, a_k\boldsymbol{T}, \boldsymbol{V}_k]$	$Kd + d + (K-1) + Kd(d-1)/2$
$[\boldsymbol{R}_k, a_k\boldsymbol{T}, \boldsymbol{V}]$	$d + d + (K-1) + Kd(d-1)/2$
$[\boldsymbol{R}_k, \boldsymbol{T}, \boldsymbol{V}_k]$	$Kd + d + Kd(d-1)/2$
$[\boldsymbol{R}_k, \boldsymbol{T}, \boldsymbol{V}]$	$d + d + Kd(d-1)/2$
$[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	$Kd + Kd + d(d-1)/2$
$[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}]$	$Kd + d(d+1)/2$
$[\boldsymbol{R}, a_k\boldsymbol{T}, \boldsymbol{V}_k]$	$Kd + (K-1) + d(d+1)/2$
$[\boldsymbol{R}, a_k\boldsymbol{T}, \boldsymbol{V}]$	$d + (K-1) + d(d+1)/2$
$[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$ (homoscédastique)	$Kd + d(d+1)/2$

TABLE 2.2: *Dimension du paramètre gaussien des modèles RTV.*

les écarts-types et les coefficients de variation ; on le note $[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$. C'est le modèle standard de composantes hétéroscédastiques.

Dans le modèle standard de classes homoscédastiques, puisque les matrices de covariances $\boldsymbol{\Sigma}_k$ $(k = 1, \ldots, K)$ sont égales, les matrices de corrélations $\boldsymbol{R}_k$ et les matrices des écarts-types $\boldsymbol{T}_k$ le sont aussi (C'est une conséquence de l'unicité de la décomposition (2.1).). Il s'agit donc du modèle $[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$.

Un modèle aux composantes de centre commun

Quand on suppose simultanément les coefficients de variation et les écarts-types homogènes (modèle $[\boldsymbol{R}_k, \boldsymbol{T}, \boldsymbol{V}]$), les classes ont toutes le même centre. Il est rare de faire une telle hypothèse dans un mélange mais nous conservons ce modèle afin que la famille de modèles dont on dispose soit générée par une combinaison exhaustive de contraintes sur les matrices $\boldsymbol{R}_k$, $\boldsymbol{T}_k$ et sur les vecteurs $\boldsymbol{V}_k$.

2.2 Des modèles indifférents à la réduction des covariables

Les propriétés 1 et 3 (Section 2.2.1) établissent que les modèles RTV sont stables par transformation linéaire et que la moitié d'entre eux le sont par transformation affine. Elles permettent d'affirmer que le choix d'un modèle RTV est indépendant des unités de mesure des données (propriété 2), indifférent (donc) à la réduction des covariables,

et indifférent (pour ceux d'entre eux qui laissent libres les coefficients de variation) au centrage des données. Les propriétés énoncées à la section 2.2.1, sont illustrées à la section 2.2.2 sur un jeu de données géologiques bien connu. Nous verrons que ces propriétés confèrent à la famille de modèles RTV une stabilité qui fait défaut aux modèles géométriques de [28], et qu'elles permettent d'interpréter la spécificité des composantes inférées comme une propriété des données conditionnelles.

2.2.1 Invariance du choix d'un modèle au choix des unités de mesure

Propriété 1 (Stabilité de chaque modèle RTV par transformation linéaire)**.** $\mathbf{X}$ *est un vecteur aléatoire de* $\mathbb{R}^d$*, distribué selon un modèle RTV.* $\boldsymbol{D} \in \mathbb{R}^{d\times d}$ *est une matrice diagonale, définie, positive.*
Alors le vecteur aléatoire $\boldsymbol{D}\mathbf{X}$ *est distribué selon le même modèle RTV que* $\mathbf{X}$*.*

Preuve. On note $\boldsymbol{\mu}_k$, $\boldsymbol{R}_k$, $\boldsymbol{T}_k$ les paramètres conditionnels de $\mathbf{X}$ et $\tilde{\boldsymbol{\mu}}_k$, $\tilde{\boldsymbol{R}}_k$, $\tilde{\boldsymbol{T}}_k$ ceux de $\boldsymbol{D}\mathbf{X}$.
$\mathbf{X}$ et $\boldsymbol{D}\mathbf{X}$ ont les mêmes corrélations conditionnelles : $\tilde{\boldsymbol{R}}_k = \boldsymbol{R}_k$ (resp. les mêmes coefficients de variation conditionnels : $\tilde{\boldsymbol{T}}_k^{-1}\tilde{\boldsymbol{\mu}}_k = {\boldsymbol{T}_k}^{-1}\boldsymbol{\mu}_k$). Supposer que les corrélations conditionnelles de $\mathbf{X}$ sont homogènes ($\boldsymbol{R}_k = \boldsymbol{R}$) entraîne que celles de $\boldsymbol{D}\mathbf{X}$ le sont aussi ($\tilde{\boldsymbol{R}}_k = \boldsymbol{R}$). De même, supposer que les coefficients de variation conditionnels de $\mathbf{X}$ sont égaux (${\boldsymbol{T}_k}^{-1}\boldsymbol{\mu}_k = \boldsymbol{V}$) entraîne que ceux de $\boldsymbol{D}\mathbf{X}$ le sont aussi ($\tilde{\boldsymbol{T}}_k^{-1}\tilde{\boldsymbol{\mu}}_k = \boldsymbol{V}$).

D'autre part puisque $\tilde{\boldsymbol{T}}_k = \boldsymbol{D}\boldsymbol{T}_k$, si les écarts-types conditionnels de $\mathbf{X}$ sont transformés de façon isotropique ($\boldsymbol{T}_{k'} = a_{k,k'}\boldsymbol{T}_k$) (resp. sont égaux ($\boldsymbol{T}_k = \boldsymbol{T}$)) d'une classe à l'autre, ceux de $\boldsymbol{D}\mathbf{X}$ le sont aussi.

□

Propriété 2. *Le choix d'un modèle RTV par AIC, BIC ou ICL, est indépendant du choix des unités de mesure des données.*

Preuve. $\mathbf{x} = \{\mathbf{x}_i; i = 1, \dots, n\}$ est un échantillon de $\mathbb{R}^d$ et $\boldsymbol{D}\mathbf{x} = \{\boldsymbol{D}\mathbf{x}_i; i = 1, \dots, n\}$, où $\boldsymbol{D} \in \mathbb{R}^{d\times d}$ est une matrice diagonale, définie, positive (matrice de changement des unités de mesure des données).

On considère un modèle de la famille RTV dont on note $\boldsymbol{\psi}$ le paramètre.
La propriété 1 assure que la modification des unités de mesure équivaut à une reparamétrisation du modèle. On en déduit :

$$\ell(\hat{\boldsymbol{\psi}}; \mathbf{x}) - \ell(\hat{\boldsymbol{\psi}}; \boldsymbol{D}\mathbf{x}) = n \ln |\boldsymbol{D}|, \tag{2.2}$$

où $\ell(\hat{\boldsymbol{\psi}}; \mathbf{x})$ et $\ell(\hat{\boldsymbol{\psi}}; \boldsymbol{D}\mathbf{x})$ désignent la log-vraisemblance maximale du paramètre $\boldsymbol{\psi}$ calculée respectivement sur la donnée de $\mathbf{x}$ et de $\boldsymbol{D}\mathbf{x}$. La différence entre les deux log-

vraisemblances maximales dépend de la taille de l'échantillon, du volume lié à la transformation des unités, mais pas du modèle RTV considéré.
Ainsi, pour tout modèle $\mathscr{M}$ de la famille RTV, on a :

$$BIC(\mathscr{M};\mathbf{x}) - BIC(\mathscr{M};\boldsymbol{D}\mathbf{x}) = n\ln|\boldsymbol{D}|. \tag{2.3}$$

Changer les unités de mesure des données selon $\boldsymbol{D}$ a donc pour effet de translater les valeurs de BIC des modèles de la famille RTV d'un terme commun $n\ln|\boldsymbol{D}|$; mais cela ne modifie pas le rang (selon BIC) du modèle dans la famille.

L'égalité (2.3) vaut également pour AIC et sa démonstration est identique à celle écrite pour BIC.

Si $\hat{\boldsymbol{\psi}}(\mathbf{x})$ désigne le paramètre d'un modèle inféré (par maximum de vraisemblance) sur la donnée de $\mathbf{x}$, alors, pour n'importe lequel des modèles RTV, les estimateurs $\hat{\boldsymbol{\psi}}(\mathbf{x})$ et $\hat{\boldsymbol{\psi}}(\boldsymbol{D}\mathbf{x})$ sont liés par les relations :

$$\hat{\boldsymbol{\mu}}_k(\boldsymbol{D}\mathbf{x}) = \boldsymbol{D}\hat{\boldsymbol{\mu}}_k(\mathbf{x}) \ \text{ et } \ \hat{\boldsymbol{\Sigma}}_k(\boldsymbol{D}\mathbf{x}) = \boldsymbol{D}\hat{\boldsymbol{\Sigma}}_k(\mathbf{x})\boldsymbol{D}\ ;\ k=1,\dots,K. \tag{2.4}$$

On en déduit que pour tout $i\in\{1,\dots,n\}$ et tout $k\in\{1,\dots,K\}$:

$$\Phi_d(\boldsymbol{D}\mathbf{x}_i;\hat{\boldsymbol{\mu}}_k(\boldsymbol{D}\mathbf{x}),\hat{\boldsymbol{\Sigma}}_k(\boldsymbol{D}\mathbf{x})) = |\boldsymbol{D}|^{-1}\Phi_d(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_k(\mathbf{x}),\hat{\boldsymbol{\Sigma}}_k(\mathbf{x})), \tag{2.5}$$

où $\Phi_d(\bullet;\boldsymbol{\mu},\boldsymbol{\Sigma})$ désigne la fonction de densité de la loi normale d-dimensionnelle de centre $\boldsymbol{\mu}$ et de matrice des covariances $\boldsymbol{\Sigma}$. Donc les paramètres $\hat{\boldsymbol{\psi}}(\mathbf{x})$ et $\hat{\boldsymbol{\psi}}(\boldsymbol{D}\mathbf{x})$ déterminent des partitions floues identiques, sur les données $\mathbf{x}$ et $\boldsymbol{D}\mathbf{x}$ respectivement. Leur entropie (voir section 1.3) est la même et la propriété (2.3) peut être étendue au critère ICL :

$$ICL(\mathscr{M};\mathbf{x}) - ICL(\mathscr{M};\boldsymbol{D}\mathbf{x}) = n\ln|\boldsymbol{D}|. \tag{2.6}$$

Ainsi, le choix d'un modèle dans la famille RTV, qu'il soit basé sur BIC, AIC ou ICL, n'est pas subordonné aux choix des unités de mesure des données.

□

Conséquence . *Le choix d'un modèle RTV et la partition de la donnée que ce modèle détermine, sont indifférents à la réduction des covariables.*

Remarque. Une famille de mélanges qui ne possèdent pas la propriété 1 (la stabilité par transformation linéaire diagonale), subordonne le modèle choisi (et la partition estimée) aux unités de mesure des données. Le modèle $[\lambda_k\boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$ de [28] par exemple (un mélange gaussien dont les composantes ont la même orientation, mais dont les autres paramètres géométriques sont libres), n'est pas stable par modification des unités de mesure. C'est un inconvénient majeur des modèles de [28]. En effet, si l'on considère que la spécificité des composantes dans un mélange, traduit une information sur les données (c'est à dire qu'imposer des orientations homogènes dans un mélange gaussien par exemple, fait sens pour certaines données), et que ce n'est pas uniquement un artifice

technique permettant de réduire la variabilité des estimateurs, alors il est souhaitable que cette information puisse être relatée indépendamment des unités de mesure.

Propriété 3 (Stabilité par transformation affine de chaque modèle RTV qui laisse libre les coefficients de variation)**.** $\mathbf{X}$ *est un vecteur aléatoire de* $\mathbb{R}^d$*, distribué selon un modèle RTV et dont les coefficients de variation conditionnels sont libres.* $\boldsymbol{D} \in \mathbb{R}^{d\times d}$ *est une matrice diagonale, définie, positive et* $\boldsymbol{b}$*, un vecteur de* $\mathbb{R}^d$*.*
Alors le vecteur aléatoire $\boldsymbol{D}\mathbf{X} + \boldsymbol{b}$ *est distribué selon le même modèle RTV que* $\mathbf{X}$*.*

Preuve. La preuve de cette propriété est similaire à celle de la propriété 1. □

Conséquence . *Centrer et réduire (axe par axe) la donnée de* $\mathbf{x}$ *apparaît, pour les modèles RTV qui laissent libres les coefficients de variation, comme une simple reparamétrisation du modèle et n'affecte ni le modèle choisi, ni la partition estimée.*

2.2.2 Classification des éruptions d'Old Faithful (Illustration)

L'objectif de cette section est à la fois d'illustrer, sur un jeu de données réelles, les propriétés des modèles RTV énoncées à la section 2.2.1, et de mettre en évidence l'intérêt des modèles d'interprétation statistique par rapport aux modèles géométriques de [28].

Les données

Le geyser Old Faithful situé dans le parc national de Yellowstone (Wyoming, USA) fait l'objet d'investigations régulières. Des géologues, S. W. Kieffer (1984) ([70]) par exemple, comparent l'activité du magma dans certaines régions volcaniques à celle de l'eau dans les conduits d'Old Faithful, et tentent d'expliquer la dynamique des éruptions de lave grâce aux éruptions aqueuses du geyser. Les données dédiées à l'étude des éruptions d'Old Faithful sont nombreuses, et la littérature sur le sujet, abondante. Nous nous intéressons ici aux données les plus courantes, celles de la librairie MASS ([106]). Il s'agit de deux cent soixante-douze éruptions (FIG. 2.3) décrites par leur durée (en minutes) et l'intervalle (en minutes également) qui sépare chaque éruption de la précédente.

Détermination du nombre de groupes et interprétation de leur spécificité

Si la figure 2.3 semble indiquer deux groupes d'éruptions, les méthodes de classification habituelles sont plus hésitantes sur la structure de ces données. Les modèles gaussiens parcimonieux d'interprététation géométrique de [28] déterminent trois groupes d'éruptions lorsqu'ils sont choisis par BIC, et deux groupes lorsqu'ils sont choisis par ICL (voir [16]). A. Atkinson et M. Riani ([6]) proposent d'autres méthodes de partitionnement qui aboutissent également à deux ou trois groupes, et ils signalent qu'une méthode de k-means traditionnelle en détermine plus de dix.

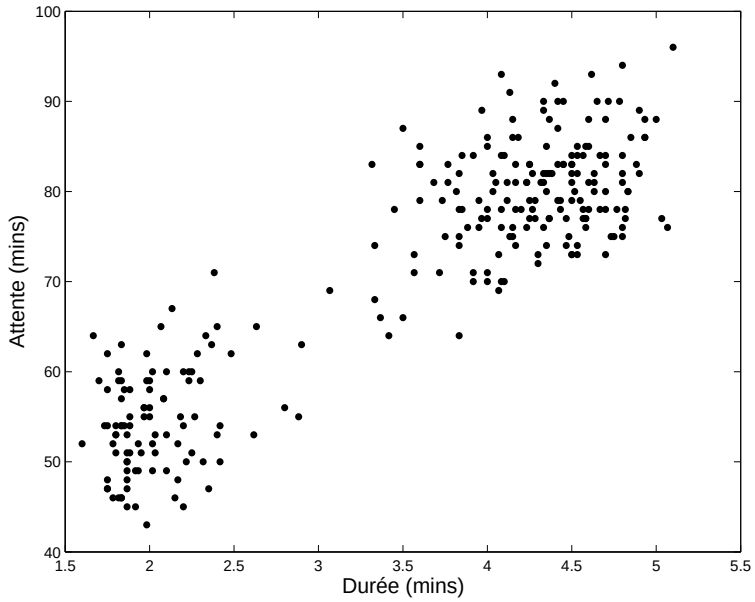


FIGURE 2.3: *Durée des éruptions du geyser Old Faithful et attente entre deux éruptions.*

Le tableaux 2.3 indique, pour un nombre de groupes variant de un à cinq, les meilleures valeurs de BIC et d'ICL obtenues par les modèles de la famille RTV.

K	1	2	3	4	5
meilleur BIC	1303.8	1158.6	**1156.3**	1160.1	1163.8
meilleur ICL	1303.8	**1158.8**	1163.5	1180.6	1185.9

TABLE 2.3: *Meilleures valeurs de BIC et d'ICL obtenues par les modèles RTV sur les données du geyser d'Old Faithful, pour un nombre variable de groupes*

Les modèles de la famille RTV, comme ceux de [28], déterminent trois groupes d'éruptions lorsqu'ils sont choisis par BIC, et deux groupes lorsqu'ils sont choisis avec ICL.

Les premières études sur Old Faithful (voir [70], [92]) se contentaient de l'histogramme bimodal des covariables pour affirmer que les éruptions du geyser se divisent en deux groupes : les éruptions longues et courtes. Si l'on admet, en s'appuyant (maintenant) sur un critère de choix de modèle comme ICL, qu'il existe bien deux groupes d'éruptions ($K = 2$), le meilleur modèle inféré dans la famille géométrique est $(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$ ($ICL = 1160.3$). Ce modèle permet, bien sûr, d'attribuer les éruptions d'Old Faithful à l'un des deux groupes inférés, mais il ne permet aucune interprétation satisfaisante de la spécificité des populations conditionnelles dans le domaine géologique. En effet, quel sens le géologue peut-il donner à des groupes d'éruptions distribués avec la même orientation ?
Le meilleur modèle de la famille RTV est $(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$ ($ICL = 1158.8$). Non seulement il est meilleur selon ICL, que le meilleur modèle de [28], mais sa spécificité s'interprète sans difficulté comme une propriété des éruptions d'Old Faithful. Le modèle

choisi postule en effet que la durée des éruptions et l'attente entre deux éruptions sont corrélées identiquement parmi les éruptions courtes et les éruptions longues.

Invariance du choix d'un modèle RTV au choix des unités de mesure

Comme l'attente entre deux éruptions d'Old Faithful est en moyenne, supérieure à une heure, et que la durée moyenne de chaque éruption est de trois minutes et demi, il arrive que la durée des éruptions soit donnée en secondes et l'attente en minutes.
Les tableaux TAB. 2.4 et TAB. 2.5 indiquent les quatre meilleurs modèles de chaque famille (géométrique et RTV) dans les deux cas de figure : lorsque la durée des éruptions est en secondes et l'attente entre deux éruptions, en minutes (TAB. 2.4), ou quand les deux covariables Durée et Attente sont exprimées en minutes (TAB. 2.5).

famille	modèle	ICL	rang
géométrique	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$	1160.3	1
	$(\pi_k)[\lambda_k \boldsymbol{S}_k\boldsymbol{\Lambda}_k\boldsymbol{S}_k']$	1161.4	2
	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}']$	1161.7	3
	$(\pi_k)[\lambda_k \boldsymbol{S}_k\boldsymbol{\Lambda}\boldsymbol{S}_k']$	1162.9	4
RTV	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	**1158.8**	1
	$(\pi_k)[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$	1161.4	2
	$(\pi_k)[\boldsymbol{R}, a_k\boldsymbol{T}, \boldsymbol{V}_k]$	1161.7	3
	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$	1163.4	4

TABLE 2.4: *Les meilleurs modèles de la famille géométrique et de la famille RTV inférés sur les données d'Old Faithful ($K = 2$ groupes), lorsque la durée des éruptions et l'attente entre les éruptions sont mesurées en minutes.*

Les tableaux TAB. 2.4 et TAB. 2.5 montrent que le rang des modèles géométriques de [28] change selon que la durée des éruptions est mesurée en minutes ou en secondes. Le rang des modèles de la famille RTV, lui, est invariant conformément à ce qu'annonçait la propriété 2.
On remarquera d'autre part que pour chaque modèle RTV, les valeurs d'ICL obtenues dans TAB. 2.4 et TAB. 2.5 diffèrent de $272 \times \ln(60)$, ce qui illustre la relation (2.6).

Comme nous l'avons mentionné à la section 1.1.2, les modèles parcimonieux (qu'ils soient gaussiens ou non) ont une vocation double. D'une part, ils permettent d'établir un compromis entre le biais du modèle estimé et sa variabilité aux fluctuations d'échantillonnage. D'autre part, ils permettent d'interpréter la spécificité des composantes comme une propriété des populations conditionnelles. Puisque les mélanges gaussiens de [28] ne sont pas invariants au choix des unités de mesure des données (en général) et puisque le choix des unités de mesure est toujours arbitraire, les modèles de [28] sont inaptes

famille	modèle	ICL	rang
géométrique	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$	**2272.4**	1
	$(\pi_k)[\lambda_k \boldsymbol{S}_k\boldsymbol{\Lambda}_k\boldsymbol{S}'_k]$	2275.0	2
	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$	2275.1	3
	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}']$	2275.4	4
RTV	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	2272.5	1
	$(\pi_k)[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$	2275.0	2
	$(\pi_k)[\boldsymbol{R}, a_k\boldsymbol{T}, \boldsymbol{V}_k]$	2275.4	3
	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$	2277.0	4

TABLE 2.5: *Les meilleurs modèles de la famille géométrique et de la famille RTV inférés sur les données d'Old Faithful ($K = 2$), lorsque la durée des éruptions est mesurée en secondes et l'attente entre les éruptions, en minutes.*

à traduire une propriété intrinsèque des données conditionnelles. Leur fonction est essentiellement technique et se limite à la recherche d'un compromis entre le biais et la variance du modèle inféré.
Le meilleur modèle géométrique, $(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$ ($ICL = 2272.4$), supplante le meilleur modèle RTV ($ICL = 2272.5$) lorsque les durées sont mesurées en secondes et les attentes, en minutes (voir TAB. 2.4 et TAB. 2.5). Mais la prééminence de ce modèle est alors liée aux unités de mesures choisies. En effet, si l'on réduit les covariables Durée et Attente ce modèle est détrôné non seulement par un modèle RTV, mais également par un modèle de sa propre famille (voir paragraphe suivant). L'homogénéité des orientations conditionnelles stipulée par ce modèle, ne traduit donc aucune propriété des deux groupes d'éruptions.

Invariance du choix d'un modèle RTV à la réduction des covariables

Puisque Durée et Attente changent d'unité d'une étude à l'autre, on peut envisager de s'affranchir du choix des unités, en réduisant les covariables.
Le tableau TAB. 2.6 indique les valeurs d'ICL obtenues par les quatre meilleurs modèles des deux familles (géométrique et RTV) sur les données de geyser réduites axe par axe. On observe d'une part que le rang des modèles RTV n'a pas changé par rapport aux tableaux 2.4 et 2.5 - ce que laissait prévoir la conséquence de la propriété 2 - alors que les quatre meilleurs modèles de la famille géométrique ont été modifiés. Le meilleur modèle des deux familles reste $(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$. Ainsi, l'hypothèse selon laquelle Durée et Attente sont identiquement corrélées parmi les deux groupes d'éruptions (courtes et longues), semble résister aux hypothèses parcimonieuses concurrentes des modèles géométriques.
On remarque d'autre part que pour chaque modèle RTV, les valeurs d'ICL dans TAB. 2.6 et TAB. 2.4 diffèrent de $272 \times \ln(a.b)$ où $a \approx 13.595$ et $b \approx 1.141$ désignent respective-

ment l'écart-type des covariables Durée et Attente exprimées en minutes. Ainsi, la valeur d'ICL d'un modèle RTV inféré sur des données aux covariables réduites, se déduit de la valeur d'ICL du même modèle inféré sur les données brutes. Ce résultat prévisible est une conséquence de (2.6).

famille	modèle	ICL	rang
géométrique	$(\pi_k)[\lambda_k \boldsymbol{S}_k \boldsymbol{\Lambda} \boldsymbol{S}_k']$	414.57	1
	$(\pi_k)[\lambda_k \boldsymbol{S}_k \boldsymbol{\Lambda}_k \boldsymbol{S}_k']$	415.55	2
	$(\pi_k)[\lambda_k \boldsymbol{S} \boldsymbol{\Lambda} \boldsymbol{S}']$	415.89	3
	$(\pi_k)[\lambda \boldsymbol{S}_k \boldsymbol{\Lambda} \boldsymbol{S}_k']$	417.02	4
RTV	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	**412.99**	1
	$(\pi_k)[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$	415.55	2
	$(\pi_k)[\boldsymbol{R}, a_k \boldsymbol{T}, \boldsymbol{V}_k]$	415.89	3
	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$	417.55	4

TABLE 2.6: *Les quatre meilleurs modèles de la famille géométrique et de la famille RTV inférés sur les données d'Old Faithful ($K = 2$), lorque la durée des éruptions et l'attente entre les éruptions sont réduites.*

Critique de la réduction des covariables

Il est fréquent dans la littérature, de réduire (et de centrer) les covariables avant de classifier des données. Mais l'objectif d'un tel conditionnement n'est pas clair. On présente parfois la réduction comme une façon d'éviter que le modèle inféré (et donc la partition estimée) ne dépende du choix des unités de mesure des données. Mais cette justification est spécieuse. La réduction des covariables n'abolit pas la référence aux unités de mesure. Au contraire, c'est un choix particulier d'un système d'unités. Réduire une covariable c'est l'exprimer dans une unité particulière : celle qui lui permet d'avoir pour variance 1. La réduction ne doit donc pas être vue comme une façon de s'affranchir des unités de mesure, mais comme une standardisation de l'ordre de grandeur de la dispersion des covariables.

Ainsi dans le contexte géologique précédent, croire qu'il est plus légitime de modéliser les durées (d'éruption et d'attente) lorsqu'elles sont réduites que lorsqu'elles sont exprimées en minutes ou en secondes, est une erreur. La réduction des covariables est un artifice dont la seule fonction est de masquer l'instabilité du modèle inféré (des modèles géométriques de [28] en l'occurrence) à la modification des unités de mesure. En effet, soit le modèle inféré est stable par modification des unités de mesure (comme le sont les modèles RTV) et alors il n'y a pas lieu de réduire les covariables pour modéliser les données ; soit le modèle inféré n'est pas stable par modification des unités de mesure et alors le modèle inféré sur les données réduites ne dit rien de la spécificité des populations

conditionnelles dans la donnée.

L'idée qui émerge ici est importante. Dans un modèle de mélange, pour que la spécificité des composantes puisse être interprétée comme une propriété des populations conditionnelles, il est indispensable que le modèle conditionnel soit stable par modification des unités de mesure. Dans le cas contraire les contraintes imposées aux composantes ne jouent qu'un rôle technique et servent uniquement à réduire la variabilité du modèle aux fluctuations d'échantillonnage.

Invariance du choix d'un modèle RTV au centrage et à la réduction des covariables

Centrer et réduire les covariables Durée et Attente conduit exactement au même tableau que Tab. 2.6. En particulier, les modèles de Tab. 2.6 (qu'il s'agisse des modèles RTV ou des modèles géométriques) gardent la même valeur d'ICL qu'ils soient inférés sur les covariables centrées réduites ou sur les covariables réduites seulement. Ce résultat ne doit pas surprendre. En effet lorsque des données sont modélisées par un modèle de [28] ou par un modèle RTV laissant libres les coefficients de variation, translater les données revient à translater le centres des composantes ; cela ne change ni la valeur maximale de la vraisemblance, ni la partition floue déterminée par le paramètre, ni la valeur d'ICL du modèle.

Le seul point qui pourrait étonner lorsqu'on infère les modèles RTV sur les données de geyser centrées et réduites est le suivant : les modèles RTV qui supposent libres les coefficients de variation ont tous une valeur d'ICL inférieure à 428.02 et les modèles RTV qui supposent au contraire, les coefficients de variation homogènes, ont tous une valeur d'ICL supérieure à 566.42. Supposer que les populations conditionnelles des données de geyser centrées et réduites ont les mêmes coefficients de variation ne semble donc pas réaliste. Nous allons voir dans le paragraphe suivant que l'hypothèse de coefficients de variation homogènes est inappropriée pour des données centrées, quelles qu'elles soient.

Critique du centrage

La critique de la réduction que nous avons entreprise dans un paragraphe précédent est générale : ce conditionnement est fréquent et pourtant il nous semble, dans tous les cas, manquer de fondements.

La critique du centrage que nous entreprenons ici est plus restreinte mais elle ne se limite pas, toutefois, aux données d'Old Faithful. Voici en quoi elle consiste. Si l'on infère un modèle RTV sur des données centrées, les modèles qui supposent homogènes les coefficients de variation, peuvent d'emblée être écartés. Nous allons voir en effet que l'homogénéité des coefficients de variation n'est pas une hypothèse réaliste pour des données centrées. Restent en lice les modèles RTV qui laissent libres les coefficients de variation. Or il est équivalent d'inférer ces modèles là, sur les données centrées et sur les données non centrées.

Les modèles RTV qui supposent homogènes les coefficients de variation sont inappropriés pour des données centrées. En effet supposer que les populations conditionnelles des données centrées ont les mêmes coefficients de variation revient à supposer que les populations conditionnelles des données non centrées ont les mêmes centres.

Les autres modèles RTV, ceux qui laissent libres les coefficients de variation, sont stables par translation et peuvent être inférés indifféremment sur les données ou sur les données centrées, sans que cela change ni leur valeur d'ICL ni la partition estimée. Le centrage des données, dans leur cas, est donc inutile.

Ce que nous avons appelé un peu sévèrement critique du centrage se résume ainsi. La moitié des modèles RTV (ceux qui supposent homogènes les coefficients de variation) sont inappropriés pour des données centrées. Par ailleurs, centrer des données est inutile pour l'autre moitié des modèles RTV (ceux qui laissent libres les coefficients de variation).

2.3 Des modèles stables par projection dans les plans canoniques

Les propriétés 4 et 5 (Section 2.3.1) établissent que les modèles RTV sont stables par projection dans le plan de n'importe quel couple de covariables. Elles permettent de représenter dans $\mathbb{R}^2$ un modèle RTV de $\mathbb{R}^d$ ($d \geq 2$) (Section 2.3.2). Elles montrent par ailleurs que les mélanges gaussiens de la famille RTV sont construits dans une cohérence mathématique que ne possèdent pas les modèles géométriques de [28].

2.3.1 Pérennité des contraintes en dimension réduite

Propriété 4 (Stabilité de chaque modèle RTV par projection dans un plan canonique)**.** *$\mathbf{X}$ est un vecteur aléatoire de $\mathbb{R}^d$ ($d \geq 2$) distribué selon un modèle RTV et $\tilde{\mathbf{X}}$ est un vecteur aléatoire de $\mathbb{R}^2$ dont les composantes sont deux covariables (distinctes) de $\mathbf{X}$. Alors le vecteur aléatoire $\tilde{\mathbf{X}}$ est distribué selon un modèle RTV de $\mathbf{R}^2$ de contraintes identiques à celles du modèle de $\mathbf{X}$ dans $\mathbb{R}^d$.*

Preuve. On note $\boldsymbol{\mu}_k$, $\boldsymbol{R}_k$, $\boldsymbol{T}_k$ les paramètres conditionnels de $\mathbf{X}$.
Il existe une matrice $\boldsymbol{P}$ de $\{0,1\}^{2\times d}$ comportant exactement un coefficient 1 par ligne, au plus un coefficient 1 par colonne, et telle que $\tilde{\mathbf{X}} = \boldsymbol{P}\mathbf{X}$.
Chaque vecteur conditionnel $(\tilde{\mathbf{X}}|Z_k = 1)$ est gaussien, son centre est : $\tilde{\boldsymbol{\mu}}_k = \boldsymbol{P}\boldsymbol{\mu}_k$ et sa matrice des covariances :

$$\tilde{\boldsymbol{\Sigma}}_k = \boldsymbol{P}\boldsymbol{T}_k\boldsymbol{R}_k\boldsymbol{T}_k\boldsymbol{P}'. \tag{2.7}$$

Or $\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{I}_2$ (identité de $\mathbb{R}^2$) donc (2.7) peut s'écrire :

$$\tilde{\boldsymbol{\Sigma}}_k = (\boldsymbol{P}\boldsymbol{T}_k\boldsymbol{P}')(\boldsymbol{P}\boldsymbol{R}_k\boldsymbol{P}')(\boldsymbol{P}\boldsymbol{T}_k\boldsymbol{P}'). \tag{2.8}$$

La matrice $\boldsymbol{P}\boldsymbol{T}_k\boldsymbol{P}'$ est diagonale et les coefficients diagonaux de $\boldsymbol{P}\boldsymbol{R}_k\boldsymbol{P}'$ valent 1. La première de ces deux matrices est donc la matrice des écarts-types de $(\tilde{\mathbf{X}}|Z_k = 1)$: $\tilde{\boldsymbol{T}}_k = \boldsymbol{P}\boldsymbol{T}_k\boldsymbol{P}'$ et la seconde est sa matrice des corrélations : $\tilde{\boldsymbol{R}}_k = \boldsymbol{P}\boldsymbol{R}_k\boldsymbol{P}'$.

Supposer que les corrélations conditionnelles de $\mathbf{X}$ sont homogènes ($\boldsymbol{R}_k = \boldsymbol{R}$) entraîne que celles de $\tilde{\mathbf{X}}$ le sont aussi ($\tilde{\boldsymbol{R}}_k = \boldsymbol{P}\boldsymbol{R}\boldsymbol{P}'$).

Une contrainte imposée aux écarts-types conditionnels de $\mathbf{X}$ implique une contrainte identique sur les écarts-types de $\tilde{\mathbf{X}}$. Si les écarts-types conditionnels de $\mathbf{X}$ sont transformés de façon isotropique : $\boldsymbol{T}_{k'} = a_{k,k'}\boldsymbol{T}_k$ ($a_{k,k'} \in \mathbb{R}_+^*$) (resp. sont égaux : $\boldsymbol{T}_k = \boldsymbol{T}$) alors ceux de $\tilde{\mathbf{X}}$ le sont aussi et : $\tilde{\boldsymbol{T}}_{k'} = a_{k,k'}\tilde{\boldsymbol{T}}_k$ (resp. $\tilde{\boldsymbol{T}}_k = \boldsymbol{P}\boldsymbol{T}\boldsymbol{P}'$).

Enfin si l'on suppose que les coefficients de variation conditionnels de $\mathbf{X}$ sont homogènes : $\boldsymbol{T}_k^{-1}\boldsymbol{\mu}_k = \boldsymbol{V}$, ceux de $\tilde{\mathbf{X}}$ le sont aussi et : $\tilde{\boldsymbol{T}}_k^{-1}\tilde{\boldsymbol{\mu}}_k = \boldsymbol{P}\boldsymbol{V}$ (Cette égalité se déduit de la propriété : $(\boldsymbol{P}\boldsymbol{T}_k\boldsymbol{P}')^{-1} = \boldsymbol{P}\boldsymbol{T}_k^{-1}\boldsymbol{P}'$.).

□

Propriété 5 (Caractérisation des modèles RTV par la dimension 2)**.** $\mathbf{X}$ *est un vecteur aléatoire de* $\mathbb{R}^d$ *(*$d \geq 2$*) dont toutes les projections dans un plan canonique de* $\mathbb{R}^2$ *sont distribuées selon le même modèle RTV (modèle RTV de même contrainte et non de même paramètre).*
Alors le vecteur aléatoire $\mathbf{X}$ *est distribué selon un modèle RTV de* $\mathbb{R}^d$ *de contraintes identiques à celles du modèle de ses projections.*

Preuve. Ecrire une preuve complète de cette propriété serait fastidieux mais l'argument sur lequel elle repose est particulièrement simple : il suffit pour que deux vecteurs de $\mathbb{R}^d$ soient identiquement corrélés (par exemple), que leurs projections dans n'importe quel plan canonique le soient.

□

Remarque. La famille des mélanges gaussiens de [28] ne vérifie pas les propriétés 4 ni 5. Un mélange gaussien sur $\mathbb{R}^d$ ($d > 2$) dont les composantes ont le même volume mais dont les autres paramètres géométriques sont libres, se projette sur un sous espace canonique de dimension 2 en un mélange gaussien dont le volume des composantes n'est pas homogène en général. Inversement un mélange gaussien peut avoir, dans un sous espace canonique de dimension 2, des composantes de volume homogène (les autres paramètres étant libres), sans qu'il s'agisse dans $\mathbb{R}^d$ du modèle $[\lambda\boldsymbol{S}_k\boldsymbol{\Lambda}_k\boldsymbol{S}_k']$ de [28]. Aussi doit-on s'abstenir de représenter (par projection) dans $\mathbb{R}^2$ les modèles gaussien de [28] définis sur $\mathbb{R}^d$ ($d > 2$). La figure FIG. 1.5(b) par exemple, ne convient pas, paradoxalement, pour représenter un mélange gaussien de $\mathbb{R}^3$ dont la forme et l'orientation des composantes sont libres et le volume homogène.

2.3.2 Représentation graphique des modèles

Dans cette section, nous proposons un mode de représentation des mélanges gaussiens, qui permet de mettre en évidence l'homogénéité (ou au contraire, l'hétérogénéité) des paramètres conditionnels d'interprétation statistique sur lesquels porte la parcimonie des modèles RTV.

La moyenne $\boldsymbol{\mu}$ et la matrice des covariances $\boldsymbol{\Sigma}$ d'un vecteur gaussien $\boldsymbol{Y}$ de $\mathbb{R}^2$, sont représentés, comme à la section 1.1.3, par une ellipse $\Gamma(\rho, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, dont les points sont à la distance ρ de $\boldsymbol{\mu}$ pour la distance de Mahalanobis $\boldsymbol{\Sigma}^{-1}$. Le plus petit rectangle contenant Γ (représenté en pointillés sur la figure 2.4) donne une idée de la dispersion des covariables de $\boldsymbol{Y}$, et permet (éventuellement) de la comparer à la dispersion des covariables d'un autre vecteur gaussien de $\mathbb{R}^2$. La corrélation des covariables de $\boldsymbol{Y}$ est représentée sur la figure 2.4, par le segment (bleu) centré en $\boldsymbol{\mu}$. L'angle que forme le segment avec l'horizontale, est proportionnel à la corrélation des covariables, et le coefficient de proportionnalité vaut $\pi/2$. Ainsi, les covariables de $\boldsymbol{Y}$ sont d'autant plus proches de l'indépendance que le segment bleu est proche de l'horizontale ; inversement, les covariables de $\boldsymbol{Y}$ sont d'autant plus corrélées que le segment bleu est proche de la verticale. $\boldsymbol{T}$ étant la matrice diagonale des écarts-types de $\boldsymbol{Y}$, le vecteur (rouge) $\boldsymbol{T}^{-1}\boldsymbol{\mu}$ d'origine $\boldsymbol{\mu}$, représente les coefficients de variation de $\boldsymbol{Y}$.

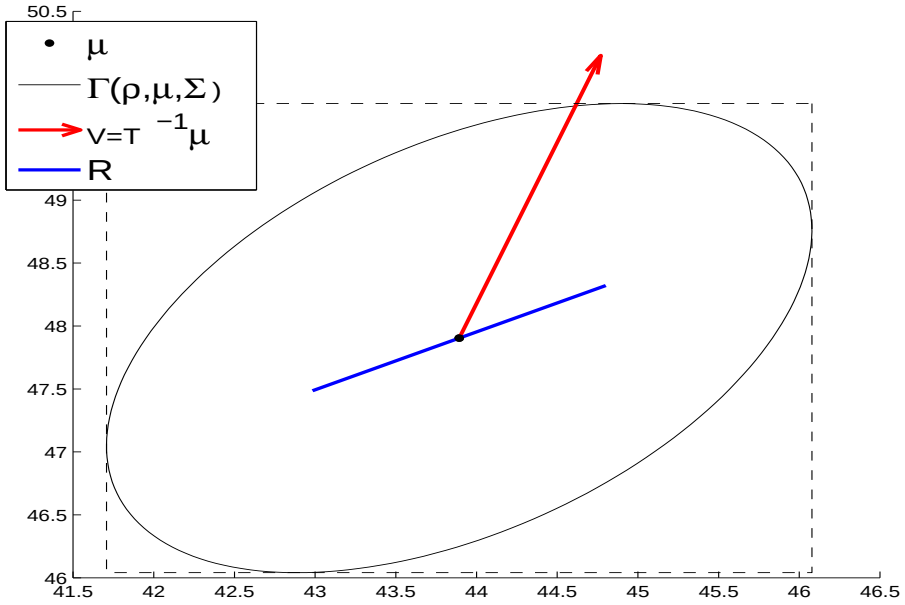
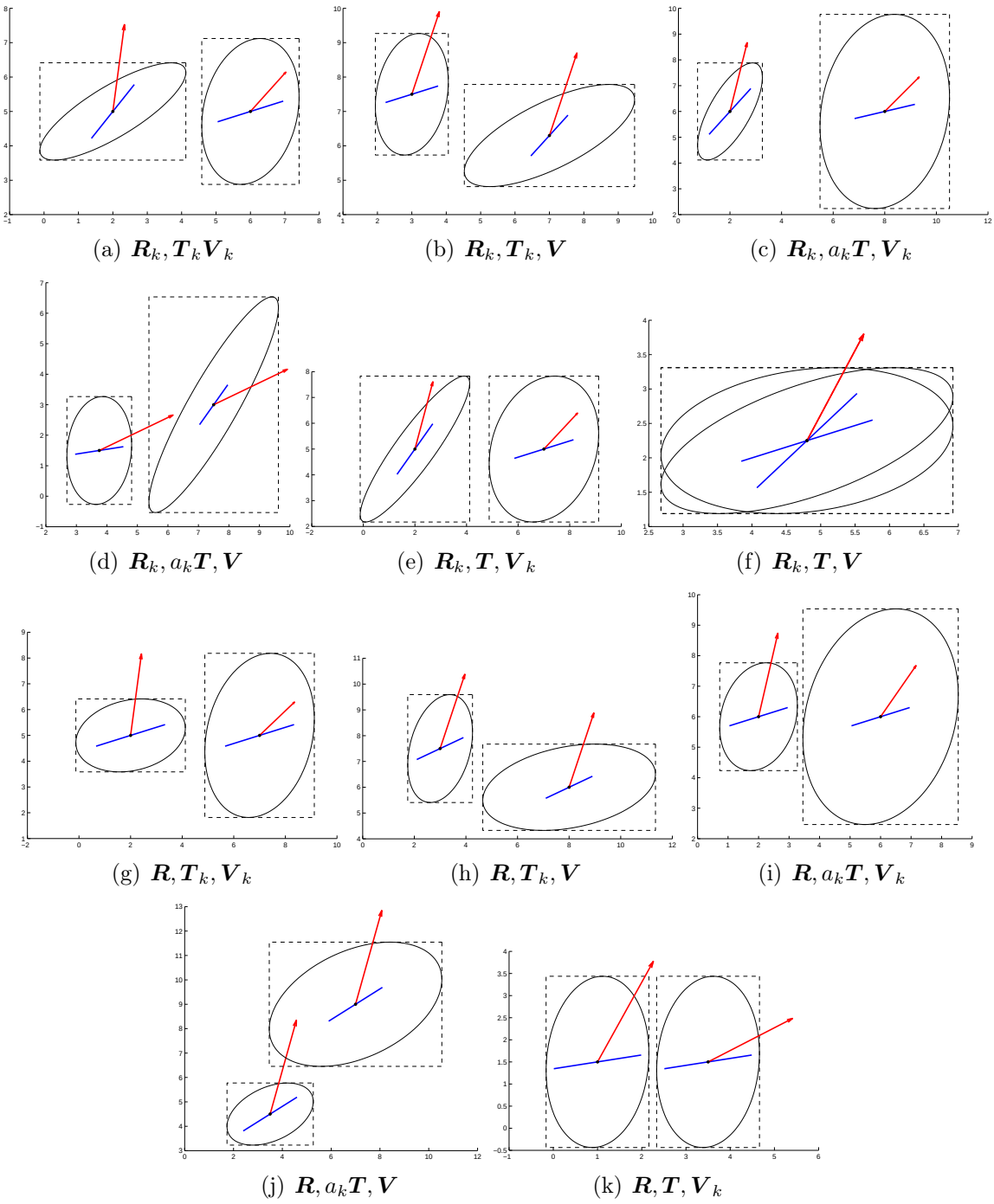


FIGURE 2.4: *Représentation des écarts-types, des coefficients de variation et de la corrélation des covariables d'un vecteur gaussien.*

Les figures 2.5(a) à 2.5(k) illustrent chacune un des onze modèles RTV (en dimension 2 et dans le cas de deux classes). Les rectangles noirs permettent de comparer les écarts-types conditionnels, les segments bleus, la corrélation des covariables et les vecteurs rouges, les coefficients de variation.



(a) $\boldsymbol{R}_k, \boldsymbol{T}_k \boldsymbol{V}_k$ (b) $\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}$ (c) $\boldsymbol{R}_k, a_k \boldsymbol{T}, \boldsymbol{V}_k$

(d) $\boldsymbol{R}_k, a_k \boldsymbol{T}, \boldsymbol{V}$ (e) $\boldsymbol{R}_k, \boldsymbol{T}, \boldsymbol{V}_k$ (f) $\boldsymbol{R}_k, \boldsymbol{T}, \boldsymbol{V}$

(g) $\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k$ (h) $\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}$ (i) $\boldsymbol{R}, a_k \boldsymbol{T}, \boldsymbol{V}_k$

(j) $\boldsymbol{R}, a_k \boldsymbol{T}, \boldsymbol{V}$ (k) $\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k$

FIGURE 2.5: *Onze modèles de mélanges gaussiens basés sur la parcimonie de paramètres d'interprétation statistique.*

Application aux données d'Old Faithful

La figure 2.6 représente le modèle $(\pi_k)[\boldsymbol{R}, \boldsymbol{T_k}, \boldsymbol{V_k}]$ inféré sur les données d'Old Faithful ainsi que la partition estimée des éruptions du geyser.

La comparaison des rectangles (en pointillés) montre que la variable catégorielle recherchée possède un effet plus important sur la dispersion des durées que sur celle des attentes.

Le modèle inféré postule que Durée et Attente sont corrélées de façon homogène parmi les éruptions courtes et longues, ce que traduit la pente identique des segments bleus. D'après le modèle retenu, les covariables n'ont pas le même coefficient de variation d'un groupe à l'autre. Cependant, la proximité des vecteurs rouges semble indiquer que les coefficients de variation conditionnels sont proches.

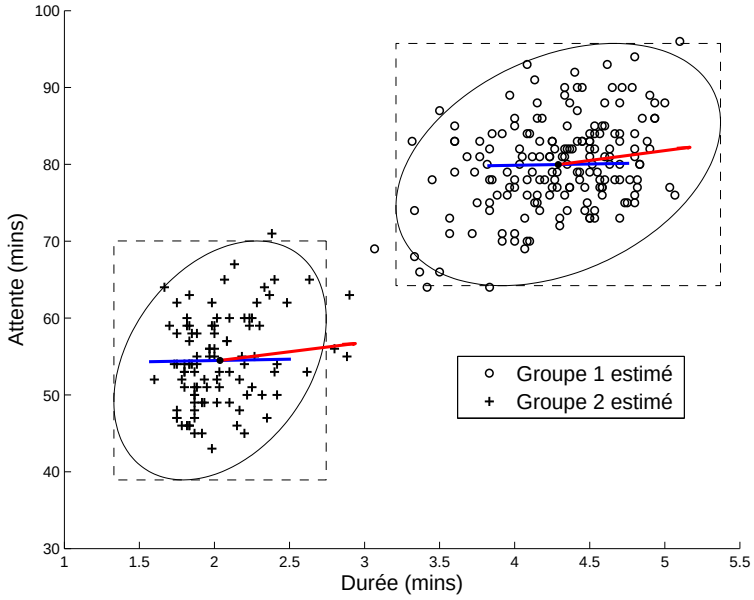


FIGURE 2.6: *La partition estimée des éruptions d'Old Faithful et le modèle choisi,* $(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$.

2.4 Interprétation dynamique des modèles de corrélations homogènes

La transformation stochastique d'une population gaussienne en une autre population gaussienne conduit, sous quelques hypothèses (réalistes dans beaucoup de cas pratiques), à une corrélation homogène des covariables dans les deux populations. Ainsi les modèles RTV de corrélations homogènes, se prêtent à une interprétation dynamique de la struture de la donnée. Nous montrons à la section 2.4.1 comment l'hypothèse d'une tranformation stochastique mutuelle des populations conditionnelles conduit à un modèle RTV de corrélations homogènes. Nous mettons en évidence à la section 2.4.2, l'intérêt de cette hypothèse dans un contexte biologique, en classifiant un échantillon d'oiseaux de l'espèce *Calonectris diomedea*. D'une part le modèle RTV choisi (et retenu parmi de nombreux autres) permet de retrouver la sous-espèce des oiseaux avec un taux d'erreur faible. D'autre part ce modèle permet une interprétation dynamique des populations conditionnelles : il suppose en effet que les sous-espèces de *Calonectris diomedea*, dérivent stochastiquement d'une population de référence.

2.4.1 Transformation stochastique des populations conditionnelles

Comme à la section 1.1.1, nous notons $\boldsymbol{X}$ le vecteur aléatoire de $\mathbb{R}^d$ qui modélise les données observées, et $\boldsymbol{Z} \in \{0;1\}^K$, la variables catégorielle qui détermine leur origine. (On note Z_k la k^{e} composante de $\boldsymbol{Z}$.)

Supposons qu'il existe pour tout couple $(k,k') \in \{1,\dots,K\}^2$, une application $\xi_k^{k'} : \mathbb{R}^d \to \mathbb{R}^d$ telle que :

$$(\boldsymbol{X}|Z_{k'}=1) \sim \xi_k^{k'}\,(\boldsymbol{X}|Z_k=1)\,. \tag{2.9}$$

Ainsi, selon cette hypothèse, les données conditionnelles du groupe k se transfoment stochastiquement en celles du groupe k'. Puisque les covariables qui décrivent les données possèdent la même signification d'un groupe à l'autre, on peut supposer que chaque descripteur dans un groupe dépend essentiellement du même descripteur dans un autre groupe. Nous formalisons cette hypothèse en supposant que la j^{e} ($j \in \{1,\dots,d\}$) composante $\left(\xi_k^{k'}\right)^{(j)}$ de l'application $\xi_k^{k'}$ ne dépend que de la j^{e} composante $\boldsymbol{x}^{(j)}$ de sa variable $\boldsymbol{x}$:

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

Autrement dit $\left(\xi_k^{k'}\right)^{(j)}$ correspond à une application de $\mathbb{R}$ dans $\mathbb{R}$ qui transforme en loi, la variable aléatoire normale $(\boldsymbol{X}|Z_k=1)^{(j)}$ en la variable $(\boldsymbol{X}|Z_{k'}=1)^{(j)}$ qui est, elle aussi, normale. Or les seules applications continûment dérivables, qui transforment une variable aléatoire normale en une autre, sont affines (voir Annexe B Conséquence du Théorème 1). Sous l'hypothèse - notée $\mathscr{H}_2$ - que pour chaque application $\xi_k^{k'}$, chacune des composantes $\left(\xi_k^{k'}\right)^j$ est continûment différentiable, les applications $\xi_k^{k'}$ sont donc affines. Pour tout couple $(k,k') \in \{1,\dots,K\}^2$, il existe alors une matrice $\boldsymbol{D}_k^{k'} \in \mathbb{R}^{d\times d}$ diagonale et un vecteur $\boldsymbol{b}_k^{k'} \in \mathbb{R}^d$ tels que :

$$(\boldsymbol{X}|Z_{k'}=1) \sim \boldsymbol{D}_k^{k'}(\boldsymbol{X}|Z_k=1) + \boldsymbol{b}_k^{k'}. \tag{2.10}$$

Puisque les populations conditionnelles sont non dégénérées, les matrices de tranformation $\boldsymbol{D}_k^{k'}$ sont régulières et les paramètres gaussiens des populations conditionnelles sont liés par :

$$\boldsymbol{\mu}_{k'} = \boldsymbol{D}_k^{k'}\boldsymbol{\mu}_k + \boldsymbol{b}_k^{k'} \ \text{ et } \ \boldsymbol{\Sigma}_{k'} = \boldsymbol{D}_k^{k'}\boldsymbol{\Sigma}_k\boldsymbol{D}_k^{k'}. \tag{2.11}$$

La relation (2.10) est la forme que prend la transformation stochastique mutuelle (2.9) des populations conditionnelles, sous les deux hypothèses $\mathscr{H}_1$ (séparabilité des variables) et $\mathscr{H}_2$ (régularité de la transformation). Supposer de plus - on note $\mathscr{H}_3$ cette hypothèse - que les matrices $\boldsymbol{D}_k^{k'}$ sont positives revient à postuler que la corrélation des covariables ne change pas de signe entre les populations conditionnelles.

La relation (2.11) entre les matrices de covariances ainsi que l'unicité de la décomposition (2.1) montrent alors que chaque matrice de transformation $\boldsymbol{D}_k^{k'}$ est liée aux matrices d'écarts-types $\boldsymbol{T}_k$ et $\boldsymbol{T}_{k'}$ selon :

$$\boldsymbol{D}_k^{k'} = \boldsymbol{T}_{k'}\boldsymbol{T}_k^{-1}, \tag{2.12}$$

et que les matrices de corrélations conditionnelles sont homogènes :

$$\boldsymbol{R}_{k'} = \boldsymbol{R}_k. \tag{2.13}$$

Ainsi un modèle RTV de corrélations homogènes s'interprète comme une transformation stochastique affine mutuelle (2.10) des populations conditionnelles. Cette transformation se déduit d'une transformation stochastique plus générale, (2.9), grâce à trois hypothèses $\mathscr{H}_1$, $\mathscr{H}_2$, $\mathscr{H}_3$, réalistes dans de nombreux cas.

Ainsi, d'après le modèle $[\pi_k, \boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$ retenu à la section 2.2.2 pour modéliser les données de geyser, il existerait une transformation stochastique affine et mutuelle entre les deux groupes d'éruptions (longues et courtes) d'Old Faithful.

Lorsqu'un modèle RTV suppose à la fois les corrélations homogènes et les coefficients de variation homogènes, la transformation stochastique mutuelle des populations conditionnelles est linéaire et non plus seulement affine : les vecteurs de translation $\boldsymbol{b}_k^{k'}$ dans la relation (2.10) sont nuls. D'après ce modèle, les populations conditionnelles dérivent stochastiquement d'une population de référence identifiable.

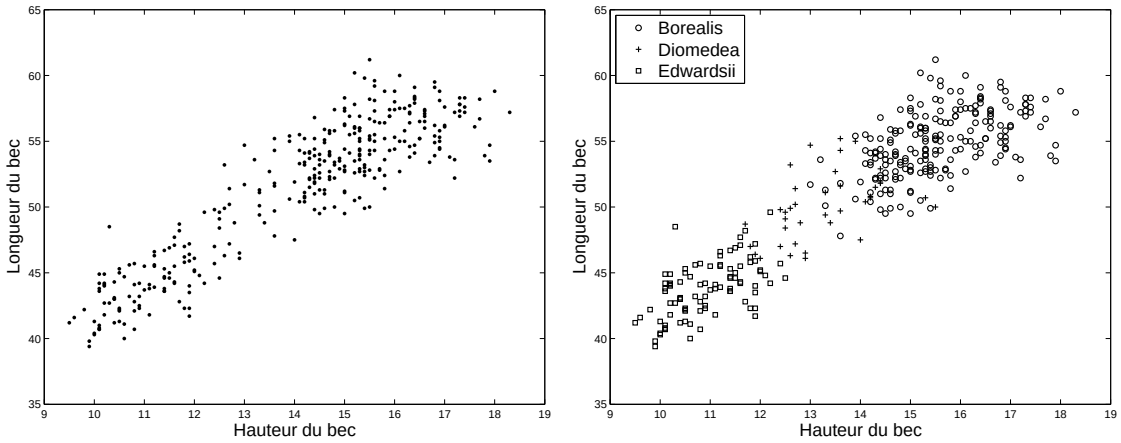
2.4.2 Classification des oiseaux de l'espèce *Calonectris diomedea*

Nous considérons un échantillon de trois cent trente six puffins cendrés ($n = 336$), des oiseaux de mer de l'espèce *Calonectris diomedea*, décrits par cinq variables morphométriques ($d = 5$) : la hauteur du bec, la longueur du bec, etc. Ces oiseaux, étudiés par Thibault et al. (1997) ([104]), se répartissent en trois sous-espèces : *borealis*, *diomedea* et *edwardsii*. Nous nous proposons de les classifier en feignant d'ignorer la sous-espèce à laquelle ils appartiennent.

La figure FIG. 2.7(a) représente l'échantillon d'oiseaux à classifier, décrit par la hauteur et la longueur du bec, et FIG. 2.7(b) indique la partition recherchée.

Nous allons voir que les modèles RTV permettent de retrouver la structure en trois groupes de cet échantillon d'oiseaux, alors que les modèles géométriques ne le permettent pas. Nous montrerons ensuite que la partition relative au modèle RTV inféré, peut être comparée avec un faible taux d'erreur, à la partition des oiseaux selon leur sous-espèce. Enfin nous verrons que le modèle retenu permet de supposer que les trois sous-espèces de puffins sont issues de la transformation stochastique d'une population de référence.

Le tableau 2.7 montre que les modèles RTV sélectionnés par BIC, retiennent bien trois groupes de puffins alors que les modèles géométriques de [28] en retiennent quatre.



(a) *La donnée* (b) *La partition recherchée*

FIGURE 2.7: *Un échantillon de puffins cendrés*

K	1	2	3	4	5
modèles RTV	4472.0	4356.6	**4335.2**	4347.8	4370.1
modèles géométriques	4472.0	4362.5	4344.2	**4341.7**	4355.8

TABLE 2.7: *Les meilleures valeurs de BIC obtenues par les modèles RTV et par les modèles géométriques, sur l'échantillon de puffins cendrés, pour un nombre variable de groupes*

Le tableau 2.8 indique, pour $K = 3$ groupes, le meilleur modèle de chaque famille (RTV et géométrique), la valeur de BIC correspondante, ainsi que le taux d'erreur de classement obtenu en comparant la partition estimée (par MAP) à la sous-espèce (connue) des oiseaux.
D'après TAB. 2.8, non seulement $(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}]$ est le meilleur des modèles parmi ceux des deux familles, mais ce modèle est un meilleur classifieur que le meilleur des modèles géométriques. Ainsi le modèle $(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}]$ permet de distinguer les *borealis*, les *diomedea* et les *edwardsii* mieux que ne le fait le meilleur modèle de la famille géométrique.

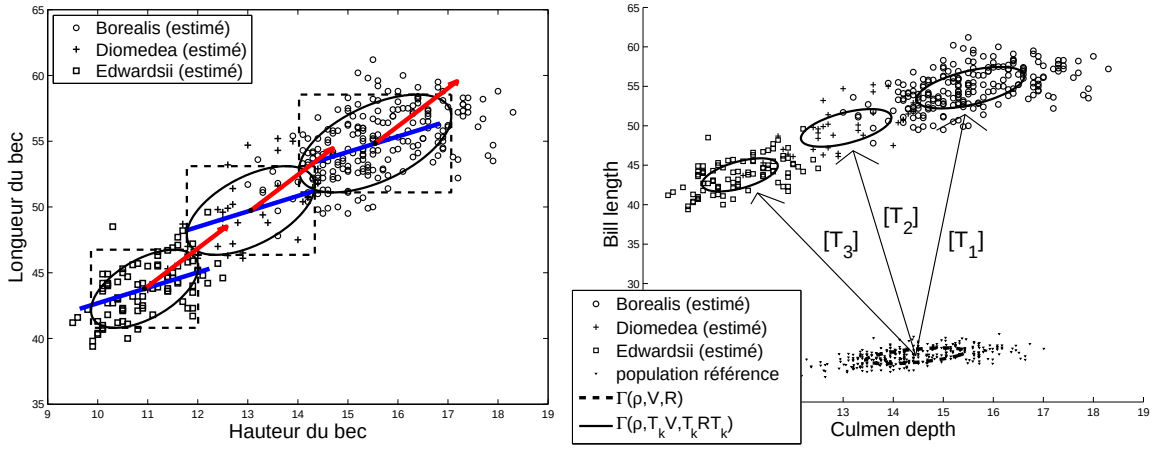
famille	meilleur modèle	BIC	taux d'erreur
RTV	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}]$	**4335.2**	**2.68**%
géométrique	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}']$	4344.2	2.98%

TABLE 2.8: *Taux d'erreur et BIC obtenus par le meilleur modèle de chaque famille (RTV et géométrique) en considérant $K = 3$ groupes.*

D'après le modèle $(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}]$ retenu, la longueur du bec, celle du tarse, de l'aile, de la queue, ainsi que la hauteur du bec, ont le même coefficient de variation parmi *borealis*, *diomedea* et *edwardsii*. D'autre part, les cinq covariables sont identiquement

corrélées d'une sous-espèce à l'autre. Selon ce modèle, les trois sous-espèces dérivent stochastiquement d'une espèce de référence modélisée par un vecteur gaussien $\mathbf{X}_0$, dont le centre est $\boldsymbol{V}$ (le vecteur des coefficients de variation des covariables, commun aux trois sous-espèces) et la matrice des covariances, $\boldsymbol{R}$ (la matrice de corrélation des covariables commune aux trois sous-espèces). En effet si $(\mathbf{X}|Z_k = 1)$ $(k = 1, 2, 3)$ désignent les vecteurs gaussiens modélisant les populations de *borealis*, *diomedea* et *edwardsii*, alors pour tout k : $(\mathbf{X}|Z_k = 1) \sim \boldsymbol{T}_k\mathbf{X}_0$, où $\boldsymbol{T}_k$ désigne la matrice diagonale des écarts-types dans la sous-espèce k.

La figure 2.8(a) représente le modèle choisi $(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}]$ et la partition estimée (interprétée comme un clivage des oiseaux selon leur sous-espèce). La figure 2.8(b) représente les trois sous-espèces (estimées) de puffins, et la population de référence dont elles sont issues stochastiquement d'après le modèle retenu.



(a) *Représentation du modèle inféré et partition estimée*

(b) *Interprétation dynamique du modèle choisi*

FIGURE 2.8: *Le modèle choisi pour les puffins cendrés,* $(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}]$

2.5 Estimation du paramètre

L'algorithme d'optimisation qui permet d'estimer $\boldsymbol{\psi}$ en maximisant sa vraisemblance dépend du modèle considéré. Il s'agit d'un algorithme EM ([37]) pour les modèles standards $[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$ et $[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$, et d'un algorithme GEM ([37]) dans les autres cas.

L'étape E commune à ces algorithmes, consiste à calculer les probabilités conditionnelles $t_{i,k}$ relatives à la valeur courante du paramètre $\boldsymbol{\psi}$, selon (1.15).

La log-vraisemblance attendue du paramètre $\boldsymbol{\psi}$ s'écrit alors :

$$\sum_{i=1}^{n}\sum_{k=1}^{K} t_{i,k}\{\ln \pi_k + \ln \Phi_d\left(\mathbf{x}_i; \boldsymbol{T}_k\boldsymbol{V}_k, \boldsymbol{T}_k\boldsymbol{R}_k\boldsymbol{T}_k\right)\} \tag{2.14}$$

(Rappelons que $\boldsymbol{V}_k = \boldsymbol{T}_k^{-1}\boldsymbol{\mu}_k$ désigne le vecteur dont les composantes sont les coefficients de variation des covariables de $(\mathbf{X}|Z_k = 1)$.). Comme (2.14) est additivement séparable par rapport aux composantes $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ d'une part, et $\boldsymbol{v} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_K)$, $\boldsymbol{\rho} = (\boldsymbol{R}_1, \ldots, \boldsymbol{R}_K)$ et $\boldsymbol{\tau} = (\boldsymbol{T}_1, \ldots, \boldsymbol{T}_K)$ d'autre part, l'étape M - ou GM suivant le modèle considéré - se décompose en deux étapes :

- Etape 1 de GM. Estimation des proportions mélange lorsqu'on les suppose libres, selon la formule classique : $\hat{\pi}_k = \hat{n}_k/n$ où $\hat{n}_k = \sum_{i=1}^n t_{i,k}$.
- Etape 2 de GM. Optimisation (modèles $[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$ et $[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$) ou augmentation (modèles restants) de (2.14) par rapport aux composantes $\boldsymbol{v}$, $\boldsymbol{\rho}$ et $\boldsymbol{\tau}$ du paramètre.

L'étape 2 de GM revient à minimiser (algorithmes EM) ou à diminuer (algorithmes GEM) le critère :

$$\sum_{k=1}^{K}\sum_{i=1}^{n} t_{i,k}\{2\ln|\boldsymbol{T}_k| + \ln|\boldsymbol{R}_k| + (\boldsymbol{T}_k^{-1}\mathbf{x}_i - \boldsymbol{V}_k)'\boldsymbol{R}_k^{-1}(\boldsymbol{T}_k^{-1}\mathbf{x}_i - \boldsymbol{V}_k)\}. \tag{2.15}$$

Voici pour les deux modèles standards de classes hétéroscédastiques et homoscédastiques, le détail de l'étape 2 de GM.

- modèle $[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$ (classes hétéroscédastiques). On note respectivement $\bar{\mathbf{x}}_k = (1/\hat{n}_k)\sum_{i=1}^n t_{i,k}\,\mathbf{x}_i$ et $\boldsymbol{S}_k = (1/\hat{n}_k)\sum_{i=1}^{n} t_{i,k}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)'$, la moyenne empirique attendue et la matrice empirique des covariances de la classe k.
  Le minimum de (2.15) est atteint lorsque $\boldsymbol{T}_k$ et $\boldsymbol{R}_k$ sont issus de la décomposition selon (2.1) de $\boldsymbol{S}_k$ et pour $\boldsymbol{V}_k = \boldsymbol{T}_k^{-1}\bar{\mathbf{x}}_k$.
- modèle $[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$ (classes homoscédastiques). Le critère (2.15) devient :

  $$\sum_{k=1}^{K}\sum_{i=1}^{n} t_{i,k}\{2\ln|\boldsymbol{T}| + \ln|\boldsymbol{R}| + (\boldsymbol{T}^{-1}\mathbf{x}_i - \boldsymbol{V}_k)'\boldsymbol{R}^{-1}(\boldsymbol{T}^{-1}\mathbf{x}_i - \boldsymbol{V}_k)\}. \tag{2.16}$$

  (2.16) est minimal lorsque $\boldsymbol{T}$ et $\boldsymbol{R}$ sont issus de la décomposition selon (2.1) de : $\boldsymbol{S} = (1/n)\sum_{k=1}^{K}\sum_{i=1}^{n} t_{i,k}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)'$, et pour $\boldsymbol{V}_k = \boldsymbol{T}^{-1}\bar{\mathbf{x}}_k$.

Dans les autres modèles le minimum de (2.15) n'est pas explicite. On peut cependant diminuer (2.15) de façon itérative par rapport aux composantes $\boldsymbol{v}$ (*Iter 1*), $\boldsymbol{\tau}$ (*Iter 2*) et $\boldsymbol{\rho}$ (*Iter 3*) du paramètre.

Voici le détail des trois itérations qui constituent l'étape GM de chaque algorithme GEM.

⊙ *Iter 1.* Les composantes $\boldsymbol{\tau}$ et $\boldsymbol{\rho}$ du paramètre étant fixées, (2.15) est minimale pour $\boldsymbol{V}_k = \boldsymbol{T}_k^{-1}\bar{\mathbf{x}}_k$ si les coefficients de variation conditionnels sont supposés libres, et pour :

$$\boldsymbol{V} = \left(\sum_{k=1}^{K} \hat{n}_k \boldsymbol{R}_k^{-1}\right)^{-1} \left(\sum_{k=1}^{K} \hat{n}_k \boldsymbol{R}_k^{-1}\boldsymbol{T}_k^{-1}\bar{\mathbf{x}}_k\right) \tag{2.17}$$

s'ils sont supposés égaux d'une classe à l'autre.

⊙ *Iter 2.* Les composantes $\boldsymbol{v}$ et $\boldsymbol{\rho}$ du paramètre sont fixées. Trois cas sont à considérer suivant que les écats-types sont libres, transformés de façon isotropique ou égaux d'une classe à l'autre.

– matrices $\boldsymbol{T}_k$ égales ($\boldsymbol{T}_k = \boldsymbol{T}$). Minimiser (2.15) revient à déterminer le minimum $\boldsymbol{x}_0$ de :

$$-2\ln|\text{diag } \boldsymbol{x}| - 2\boldsymbol{L}\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x};\ \boldsymbol{x} \in \left(\mathbb{R}_*^+\right)^d, \tag{2.18}$$

avec

$$\boldsymbol{Q} = (1/n)\sum_{k=1}^{K}\sum_{i=1}^{n} t_{i,k}(\text{diag } \mathbf{x}_i)\boldsymbol{R}_k^{-1}(\text{diag } \mathbf{x}_i) \tag{2.19}$$

et

$$\boldsymbol{L} = \sum_{k=1}^{K} (\hat{n}_k/n)\boldsymbol{V}_k'\boldsymbol{R}_k^{-1}(\text{diag } \bar{\mathbf{x}}_k). \tag{2.20}$$

$\boldsymbol{Q}$ étant symétrique, définie, positive, (2.18) est convexe par rapport à $\boldsymbol{x}$ et $\boldsymbol{x}_0$ peut être approché par un algorithme d'optimisation convexe. La matrice $\boldsymbol{T}$ recherchée est alors $\boldsymbol{T} = (\text{diag } \boldsymbol{x}_0)^{-1}$.

– $\boldsymbol{T}_k = a_{1,k}\boldsymbol{T}_1$ ($a_{1,k} > 0$). On note $a_k = a_{1,k}$ ($k = 1, \ldots, K$). On diminue le critère (2.15) de façon itérative (Le minimum de (2.15) par rapport à $\boldsymbol{\tau}$ n'est pas explixite dans ce modèle.) en le minimisant alternativement par rapport aux coefficients a_k et à la matrice $\boldsymbol{T}_1$. Lorsque les coefficients a_k ($k = 1, \ldots, K$) sont fixés, (2.15) est convexe en $\boldsymbol{T}_1^{-1}$; on détermine la valeur de $\boldsymbol{T}_1$ qui le minimise. Pour $\boldsymbol{T}_1$ fixée, (2.15) est minimal pour :

$$a_k = \frac{-\boldsymbol{L}_k(\text{diag } \boldsymbol{T}_1^{-1}) + \sqrt{\left[\boldsymbol{L}_k(\text{diag } \boldsymbol{T}_1^{-1})\right]^2 + 4d\left[(\text{diag } \boldsymbol{T}_1^{-1})'\boldsymbol{R}_k^{-1}(\text{diag } \boldsymbol{T}_1^{-1})\right]}}{2d}; \tag{2.21}$$

$k = 2, \ldots, K$.

– matrices $\boldsymbol{T}_k$ libres. Minimiser (2.15) par rapport à $\boldsymbol{\tau}$ revient à déterminer successivement (et indépendamment) les k minima de (2.18) obtenus pour des paramètres :

$$\boldsymbol{Q}_k = (1/\hat{n}_k)\sum_{i=1}^{n} t_{i,k}(\text{diag } \mathbf{x}_i)\boldsymbol{R}_k^{-1}(\text{diag } \mathbf{x}_i) \tag{2.22}$$

et

$$\boldsymbol{L}_k = \boldsymbol{V}_k'\boldsymbol{R}_k^{-1}(\text{diag } \bar{\mathbf{x}}_k). \tag{2.23}$$

Chaque minimum $\boldsymbol{x}_k$ peut être approché par un algorithme d'optimisation convexe puisque la matrice $\boldsymbol{Q}_k$ est symétrique, définie et positive.
Les matrices $\boldsymbol{T}_k$ recherchées sont alors données par : $\boldsymbol{T}_k = (\text{diag } \boldsymbol{x}_k)^{-1}$.

⊙ *Iter 3.* Les composantes $\boldsymbol{v}$ et $\boldsymbol{\tau}$ du paramètre sont fixées. Le critère (2.15) à minimiser devient :

$$\sum_{k=1}^{K} \hat{n}_k \left[\ln |\boldsymbol{R}_k| + \text{tr} \left(\boldsymbol{W}_k \boldsymbol{R}_k^{-1} \right) \right], \tag{2.24}$$

avec $\boldsymbol{W}_k = (1/\hat{n}_k) \sum_{i=1}^{n} t_{i,k} (\boldsymbol{T}_k^{-1} \mathbf{x}_i - \boldsymbol{V}_k)(\boldsymbol{T}_k^{-1} \mathbf{x}_i - \boldsymbol{V}_k)'$ si les corrélations conditionnelles sont supposées libres, et :

$$\ln |\boldsymbol{R}| + \text{tr} \left(\boldsymbol{W} \boldsymbol{R}^{-1} \right), \tag{2.25}$$

avec $\boldsymbol{W} = \sum_{k=1}^{K} (\hat{n}_k / n) \boldsymbol{W}_k$ si les corrélations conditionnelles sont supposées homogènes.

On peut diminuer (2.24) ou (2.25) d'après la proposition 5 (Annexe A), alternativement par rapport à chacune des corrélations de $\boldsymbol{R}_k$ ou de $\boldsymbol{R}$.

2.6 Bilan et perspectives

L'usage de mélanges parcimonieux (qu'ils soient gaussiens ou autres) vise à trouver un compromis entre le biais du modèle estimé et sa variabilité aux fluctuations d'échantillonnage. Mais le choix des paramètres sur lesquels porte la parcimonie d'une famille de mélanges, ne peut se réduire à cet objectif technique. Les contraintes envisagées doivent permettre aux modèles parcimonieux d'être à la fois représentés et interprétés.

Afin de répondre à ce nouvel objectif, nous avons défini dans ce chapitre, des modèles de mélanges gaussiens multidimensionnels dont la parcimonie porte sur l'écart-type, le coefficient de variation des variables, la corrélation des covariables, c'est à dire des paramètres d'interprétation statistique.
Nous avons montré que le choix de l'un de ces modèles est indifférent au choix des unités de mesure, indifférent à la réduction des covariables, et que pour ces modèles, le centrage des données est inutile. Ces nouveaux modèles - dits RTV - sont par ailleurs, stables par projection dans les plans canoniques, ce qui permet leur représentation fidèle en dimension réduite et assure la cohérence mathématique de leur élaboration.

L'invariance du choix d'un modèle RTV au changement des unités de mesure, et la stabilité du modèle choisi par projection dans les plans canoniques, sont déterminants dans l'intérêt de cette famille de mélanges. Les mélanges d'interprétation géométrique

de [28] ne possèdent aucune de ces propriétés. Nous avons présenté deux cas, l'un issu de la géologie, l'autre de la biologie, où les modèles géométriques sont supplantés par des modèles RTV. Dans ces deux exemples, la famille RTV fournit des modèles de meilleur qualité (pour un critère de choix de modèle), de meilleurs classifieurs et des modèles d'interprétation plus pertinente.

Certains modèles RTV permettent une interprétation dynamique des groupes inférés. Nous avons montré par exemple, que les modèles de corrélations homogènes peuvent être vus comme postulant une transformation stochastique mutuelle des populations conditionnelles, et nous avons illustré cette interprétation dans un contexte ornithologique.
L'interprétation dynamique des groupes inférés est une nouveauté en classification. Les méthodes traditionnelles de classification, qu'elles soient basées sur des mélanges ou sur l'optimisation d'un critère géométrique, considèrent toutes en effet, les classes inférées d'un point de vue statique. Il nous semble pertinent de définir (dans la mesure du possible) des modèles de mélanges qui, comme les mélanges de corrélations homogènes dans le cas gaussiens, permettent d'interpréter dynamiquement les classes. De tels mélanges établissent un lien interprétable entre populations conditionnelles et ne se caractérisent plus seulement par une contrainte statique sur leur paramètre.

Les multiples essais de classification que nous avons menés à l'aide des modèles RTV (dont ce chapitre ne relate qu'une partie), montrent une prédominance des modèles de corrélations homogènes. Or l'hypothèse de corrélations homogènes n'est pas universelle. Nous expliquons (en partie) le succès de ces modèles par le nombre important des corrélations conditionnelles à estimer dans un mélange multidimensionnel, et par le coût de leur estimation lorsqu'elles sont libres. Un mélange gaussien hétéroscédastique à trois composantes en dimension 7 par exemple, voit la taille de son paramètre diminuer de 40% lorsqu'on suppose les corrélations conditionnelles homogènes.
Nous observons cependant sur plusieurs jeux de données issus de la biologie (les Iris de Fisher, les vins italiens du jeu Wine disponible sur le site de l'UCI (`http://archive.ics.uci.edu/ml/datasets/Wine`), un échantillon d'oiseaux pyrénéens, des cincles plongeurs ([32])), que les modèles de corrélations homogènes restent prépondérants lorsqu'on les met en concurrence avec les modèles géométriques de [28]. Nous invitons donc les biologistes utilisateurs des modèles RTV, à envisager d'un point de vue d'expert, la possibilité de covariables biologiques corrélées de façon homogène dans des populations hétérogènes.

Nous avons insisté sur l'intérêt de modèles stables par projection et invariants au changement des unités de mesure. Ces propriétés revêtent-elles toujours autant d'importance dans le domaine spécifique de la grande dimension ? Autrement dit, parmi deux familles de mélanges dédiés à la grande dimension, laquelle faut-il choisir, si l'une possède ces propriété et l'autre non ? Les mélanges de Factors Analyzers évoqués à la section 1.1.4, par exemple, possèdent les deux propriétés. Les mélanges gaussiens de C. Bouveyron ([21]) eux, ne possèdent ni l'une ni l'autre. Cette différence ne suffit pas, pourtant, à privilégier les mélanges de Factors Analyzers aux modèles de C. Bouveyron ([21]). Quel que soit le domaine dont elles sont issues, les données de grande dimension ont un point commun : la carence d'information dans certaines directions de leur espace d'acquisition. La spécificité des modèles dédiés à la grande dimension doit donc

chercher à traduire cette carence plutôt que toute autre propriété (hypothétique) des données conditionnelles. Pour cette raison, mais également parce que l'inférence de leur paramètre est coûteuse en temps de calcul, nous déconseillons l'usage des modèles RTV dans le contexte de la grande dimension.

Nota Bene. Les mélanges parcimonieux de C. Bouveyron ([21]) sont issus, comme les mélanges géométriques de [28], d'une décomposition spectrale des matrices de covariances. A ce titre ils ne sont ni invariants au changement des unités de mesure (voir Section 1.1.4), ni stables par projection dans les plans canoniques. La plupart d'entre eux supposent même un aplatissement isotropique des classes ; cela réduit encore, la proportion de modèles de la famille possédant l'une de ces propriétés.

Les paramètres statistiques sur lesquels porte la parcimonie des modèles RTV ne sont pas propres aux mélanges gaussiens et l'on peut étendre les modèles RTV à tout mélange dont les lois conditionnelles admettent des moments d'ordre 1 et 2 finis. Ainsi par exemple, il sera possible de définir des mélanges parcimonieux de Student basés sur la décomposition (2.1) des matrices de covariances, dès lors que le degré de liberté des composantes est strictement supérieur à 2.

Appendices

Annexe A

Fondements de l'algorithme GEM pour les modèles RTV

L'algorithme GEM qui permet d'estimer le paramètre de neuf des onze modèles de la famille RTV, repose sur la proposition 5 ci-après.

Lemme 1. *La matrice $\boldsymbol{D} \in \mathbb{R}^{d\times d}$ diagonale, définie et positive qui minimise* $\mathrm{tr}(\boldsymbol{D}) - \ln|\boldsymbol{D}|$ *est l'identité ($\boldsymbol{D} = \boldsymbol{I}_d$) (et le minimum atteint vaut d).*

Preuve. L'application : $\boldsymbol{x} \mapsto (\boldsymbol{x} - \ln \boldsymbol{x})$ définie sur $\mathbb{R}_*^+$ est minimale en 1, et son minimum vaut 1. □

Proposition 1. *$\boldsymbol{W} \in \mathbb{R}^{d\times d}$ est une matrice de covariances (symétrique, définie, positive).*
La matrice de covariances $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ qui minimise :

$$\mathrm{tr}(\boldsymbol{W}\boldsymbol{\Sigma}^{-1}) - \ln|\boldsymbol{W}\boldsymbol{\Sigma}^{-1}|, \tag{A.1}$$

est $\boldsymbol{\Sigma} = \boldsymbol{W}$ (et le minimum atteint vaut d).

Preuve. $\boldsymbol{\Sigma}^{-1}$ est (comme $\boldsymbol{\Sigma}$) symétrique, définie, positive. On note $\boldsymbol{\Sigma}^{-1/2}$ l'unique matrice symétrique, définie et positive dont le carré vaut $\boldsymbol{\Sigma}^{-1}$. Puisque $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{W}\boldsymbol{\Sigma}^{-1/2}$ est symétrique, définie et positive, ses valeurs propres sont strictement positives. Comme $\boldsymbol{W}\boldsymbol{\Sigma}^{-1}$ et $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{W}\boldsymbol{\Sigma}^{-1/2}$ ont les mêmes valeurs propres, $\boldsymbol{W}\boldsymbol{\Sigma}^{-1}$ est semblable à une matrice $\boldsymbol{D}$ diagonale définie et positive : il existe une matrice $\boldsymbol{P} \in \mathbb{R}^{d\times d}$ inversible (matrice de changement de base) telle que $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{P}$. La trace et le déterminant étant invariants par changement de base, on a : $\mathrm{tr}(\boldsymbol{W}\boldsymbol{\Sigma}^{-1}) - \ln|\boldsymbol{W}\boldsymbol{\Sigma}^{-1}| = \mathrm{tr}(\boldsymbol{D}) - \ln|\boldsymbol{D}|$. D'après le lemme 1 le second membre est minimal lorsque $\boldsymbol{D} = \boldsymbol{I}_d$, c'est à dire quand $\boldsymbol{\Sigma} = \boldsymbol{W}$. □

Remarque . *La proposition 1 permet (accessoirement) d'établir que l'estimateur du maximum de vraisemblance de la matrice des covariances d'une loi normale est - lorsqu'on suppose connu le centre de cette loi - la matrice empirique des covariances de l'échantillon.*

Justification. $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ est un échantillon de $\mathbb{R}^d$. On suppose qu'il provient (de façon indépendante) d'une loi normale $d-$dimensionnelle $\mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ dont on connaît le centre $\boldsymbol{\mu}$. Estimer par maximum de vraisemblance la matrice des covariances $\boldsymbol{\Sigma}$ revient à minimiser : $\ln|\boldsymbol{\Sigma}| + \mathrm{tr}(\boldsymbol{W}\boldsymbol{\Sigma}^{-1})$ par rapport à $\boldsymbol{\Sigma}$, avec $\boldsymbol{W} = (1/n)\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'$. D'après la proposition 1 : $\hat{\boldsymbol{\Sigma}} = \boldsymbol{W}$.

□

Remarque. La proposition 1 est énoncée et démontrée dans [28]. La preuve que nous en proposons ici, basée sur des arguments d'algèbre linéaire, est une alternative à celle qu'en proposent G. Celeux et G. Govaert basée, elle, sur une optimisation sous contrainte.

Proposition 2. *$\boldsymbol{W} \in \mathbb{R}^{d\times d}$ est une matrice de covariances.*
Lorsqu'une matrice de corrélations $\boldsymbol{R} \in \mathbb{R}^{d\times d}$ ($\boldsymbol{R}$ est symétrique, définie et positive ; ses coefficients diagonaux valent 1 et ses autres coefficients sont strictement compris entre -1 et 1.) converge vers une matrice située sur le bord de son domaine, $\mathrm{tr}(\boldsymbol{W}\boldsymbol{R}^{-1}) - \ln|\boldsymbol{W}\boldsymbol{R}^{-1}|$ tend vers $+\infty$.

Preuve. Une matrice de corrélations est une matrice de covariances. Pour toute matrice de corrélations $\boldsymbol{R}$, il existe donc une matrice $\boldsymbol{P} \in \mathbb{R}^{d\times d}$ inversible telle que $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{W}\boldsymbol{R}^{-1}\boldsymbol{P}$ soit diagonale, définie, positive (voir la preuve de la proposition 1) et alors : $\mathrm{tr}(\boldsymbol{W}\boldsymbol{R}^{-1}) - \ln|\boldsymbol{W}\boldsymbol{R}^{-1}| = \mathrm{tr}(\boldsymbol{D}) - \ln|\boldsymbol{D}|$. Quand $\boldsymbol{R}$ converge vers une matrice située sur le bord de son domaine, $|\boldsymbol{R}|$ tend vers 0 et $|\boldsymbol{D}|$ vers $+\infty$. Comme $\boldsymbol{D}$ est diagonale, définie et positive : $\mathrm{tr}(\boldsymbol{D}) - \ln|\boldsymbol{D}| \geq (\ln|\boldsymbol{D}|)/2$. Donc $\mathrm{tr}(\boldsymbol{W}\boldsymbol{R}^{-1}) - \ln|\boldsymbol{W}\boldsymbol{R}^{-1}|$ tend aussi vers $+\infty$.

□

Proposition 3. *$\boldsymbol{W} \in \mathbb{R}^{d\times d}$ est une matrice de covariances (symétrique, définie et positive).*
Lorsque $\boldsymbol{R} \in \mathbb{R}^{d\times d}$ décrit l'ensemble $\mathscr{R}$ des matrices de corrélations, $\mathrm{tr}(\boldsymbol{W}\boldsymbol{R}^{-1}) - \ln|\boldsymbol{W}\boldsymbol{R}^{-1}|$ admet un minimum.

Preuve. D'après la proposition 1, $\mathrm{tr}(\boldsymbol{W}\boldsymbol{\Sigma}^{-1}) - \ln|\boldsymbol{W}\boldsymbol{\Sigma}^{-1}|$ admet un minimum lorsque $\boldsymbol{\Sigma}$ décrit l'ensemble des matrices de covariances. Or une matrice de corrélations est une matrice de covariances. On en déduit le résultat.

□

Proposition 4. *L'estimateur du maximum de vraisemblance des corrélations d'une loi normale - les autres paramètres de cette loi étant connus - existe. Les équations de vraisembance ont une solution et cette solution est à l'intérieur du domaine contenant les matrices de corrélations.*

Preuve. La matrice des covariances $\boldsymbol{\Sigma}$ d'une loi normale multi-dimensionnelle se décompose de façon unique en $\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{R}\boldsymbol{T}$ où $\boldsymbol{T}$ désigne la matrice diagonale des écarts-types

et $\boldsymbol{R}$, la matrice des corrélations (voir (2.1)).

$\mathbf{x}_1, \ldots, \mathbf{x}_n$ $(\in \mathbb{R}^d)$ sont des réalisations indépendantes d'une loi normale $d-$dimensionnelle dont on connaît le centre $\boldsymbol{\mu} \in \mathbb{R}^d$ et la matrice des écarts-types $\boldsymbol{T} \in \mathbb{R}^{d\times d}$, mais pas la matrice des corrélations $\boldsymbol{R} \in \mathbb{R}^{d\times d}$.

Optimiser la vraisemblance de $\boldsymbol{R}$ est équivalent à minimiser :

$$\ln|\boldsymbol{R}| + \mathrm{tr}(\boldsymbol{W}\boldsymbol{R}^{-1}), \tag{A.2}$$

avec : $\boldsymbol{W} = (1/n)\boldsymbol{T}^{-1}\left[\sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'\right]\boldsymbol{T}^{-1}$.

Le minimum de (A.2) existe d'après la proposition 3 et il est atteint à l'intérieur du domaine de $\boldsymbol{R}$ d'après la proposition 2.

□

Proposition 5. *$\boldsymbol{W} \in \mathbb{R}^{d\times d}$ est une matrice de covariances.*
Alors, en tout point de $\mathscr{R}$, les applications partielles de l'application :

$$\left\{\begin{array}{ccc} \mathscr{R} & \to & \mathbb{R} \\ \boldsymbol{R} & \mapsto & \ln|\boldsymbol{R}| + \mathrm{tr}(\boldsymbol{W}\boldsymbol{R}^{-1}) \end{array}\right. \tag{A.3}$$

admettent chacune soit 1 *soit* 2 *minima locaux.*

Preuve. $\boldsymbol{R}$ étant une matrice de corrélations (un point de $\mathscr{R}$), une application partielle de (A.3) en $\boldsymbol{R}$ peut s'écrire :

$$\boldsymbol{\xi} : \boldsymbol{x} \mapsto \ln|\boldsymbol{R} + \boldsymbol{x}\boldsymbol{H}| + \mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{R} + \boldsymbol{x}\boldsymbol{H})^{-1}\right) \ ; \ \boldsymbol{x} \in \mathbb{R}, \tag{A.4}$$

où $\boldsymbol{H}$ est une matrice dont tous les coefficients sont nuls sauf deux : $\boldsymbol{H}(i,j) = 1$ et $\boldsymbol{H}(j,i) = 1$ pour un certain couple $(i,j) \in \{1,\ldots,d\}^2$, $i \neq j$.

On montre que :

$$\boldsymbol{\xi}(\boldsymbol{x}) = \ln\left(-|\boldsymbol{\Delta}_{i,j}^{j,i}|\boldsymbol{x}^2 + |\boldsymbol{\Delta}_i^j|\boldsymbol{x} + |\boldsymbol{R}|\right) + \frac{\mathrm{tr}\left(\boldsymbol{W}.\mathrm{com}(\boldsymbol{R} + \boldsymbol{x}\boldsymbol{H})\right)}{-|\boldsymbol{\Delta}_{i,j}^{j,i}|\boldsymbol{x}^2 + |\boldsymbol{\Delta}_i^j|\boldsymbol{x} + |\boldsymbol{R}|}, \tag{A.5}$$

où $\boldsymbol{\Delta}_i^j$ (resp. $\boldsymbol{\Delta}_{i,j}^{j,i}$) désigne la matrice mineure de $\boldsymbol{R}$ obtenue en supprimant la ligne i (resp. les lignes i et j) et la colonne j (resp. les colonnes j et i).

Comme $\boldsymbol{R}$ est symétrique, définie et positive, les déterminants : $|\boldsymbol{\Delta}_{i,j}^{j,i}|$, $|\boldsymbol{\Delta}_i^j|$ et $|\boldsymbol{R}|$ sont strictements positifs et $\boldsymbol{\xi}$ est définie sur un intervalle $\boldsymbol{I}$ ouvert, contenant 0 (dont on peut déterminer les bornes).

On déduit de la proposition 1 que $\boldsymbol{\xi}$ tend vers $+\infty$ aux bornes de $\boldsymbol{I}$; elle a donc au moins un minimum sur $\boldsymbol{I}$.

Comme $\mathrm{tr}\,(\boldsymbol{W}.\mathrm{com}(\boldsymbol{R}+\boldsymbol{x}\boldsymbol{H}))$ est un polynôme de degré 2 en $\boldsymbol{x}$, $\boldsymbol{\xi}'(\boldsymbol{x})$ est du signe, sur $\boldsymbol{I}$, d'un polynôme de degré 3. Comme $\boldsymbol{\xi}'(\boldsymbol{x})$ s'annule au plus trois fois sur $\boldsymbol{I}$, $\boldsymbol{\xi}$ admet au plus 3 extrema locaux sur $\boldsymbol{I}$.

Des deux points précédents on déduit que $\boldsymbol{\xi}$ a soit un minimum global (et éventuellement un point d'inflexion), soit deux minima et un maximum local.

□

Conséquence . *La proposition 5 permet de définir un algorithme qui augmente de façon itérative la vraisemblance de la matrice des corrélations d'une loi normale - les autres paramètres de cette loi étant fixés - en l'optimisant alternativement par rapport à chacune des corrélations.*

Deuxième partie

Classification simultanée d'échantillons multiples

Sommaire

Chapitre 3

Introduction

3.1 Présentation de la méthode

Motivations

La classification simultanée est une méthode novatrice dédiée à la classification de plusieurs échantillons. Elle repose sur le principe d'un lien réaliste entre populations. Elle présume de façon raisonnée et étayée, que classifier plusieurs échantillons sous l'hypothèse de ce lien, améliore leur classification indépendante.

Le principe de la classification simultanée ressemble à celui qui guide le choix d'un modèle en classification à base de mélanges. Lorsqu'on classifie un seul échantillon, on peut utiliser pour modéliser cet unique échantillon, l'un des multiples mélanges existant. On peut aussi imaginer et employer un mélange d'un type nouveau dont les composantes relatent une spécificité de la donnée. La qualité du classifieur obtenu sera d'autant meilleure que le modèle utilisé est réaliste c'est à dire que son écart avec le vrai modèle est faible.

La classification simultanée se place dans un contexte différent puisqu'elle cherche à classifier non pas un seul mais plusieurs échantillons. Cependant (i) son objectif est similaire : améliorer la qualité d'un classifieur et (ii) elle repose sur le même principe d'un modèle réaliste.

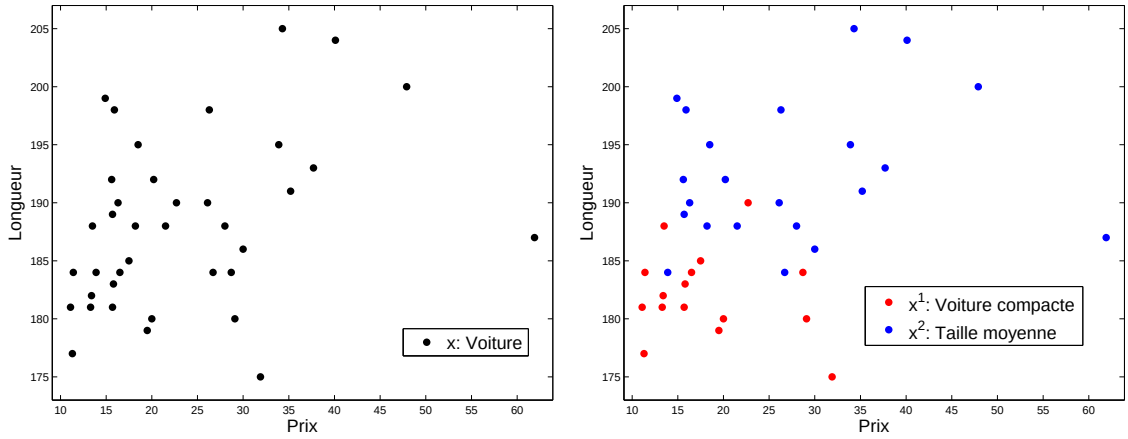
Des situations nombreuses mais ignorées de classification simultanée

Les situations de classification simultanée sont fréquentes, mais elles ne sont quasiment jamais identifiées comme telles.

Mise en évidence de plusieurs échantillons à classifier

Il n'est pas rare que des échantillons pourtant distincts, soient classifiés comme s'ils

provenaient de la même population. La figure 3.1(a) représente un échantillon de voitures décrites par leur prix et leur longueur, dont on aimerait - en les classant par un mélange à deux composantes - retrouver l'origine (USA ou non USA). Or ces voitures sont, par ailleurs, de deux types : compactes ou de taille moyenne, et l'on connaît le type de chacune d'elles. La figure 3.1(b) montre que, si l'on prend en compte le type des voitures, ce n'est pas un mais deux échantillons que l'on a à classer, et que ces deux échantillons semblent issus de populations distinctes. (Il se peut que les échantillons proviennent de la même population et que l'hétérogénéité observée soit un effet d'échantillonnage. Mais si tel est le cas, ce n'est pas pour avoir ignoré leur type.)



(a) *Un échantillon de voirures à classifier...* (b) *... ou deux échantillons ?*

FIGURE 3.1: *Une situation de classification simultanée qui s'ignore*

Lien entre populations

Il n'est pas rare non plus que des échantillons - identifiés comme provenant de populations différentes - possèdent un lien ou recèlent une information commune, mais soient classifiés malgré cela de façon indépendante, sans que l'on tente d'intégrer au modèle l'information qui les relie.

Si l'on admet que les deux échantillons $\mathbf{x}^1$ et $\mathbf{x}^2$ de FIG. 3.1(b) proviennent de deux populations distinctes, leur classification requiert (habituellement) deux procédures indépendantes. Or ces deux échantillons ne sont pas sans lien. D'abord ils sont constitués dans les deux cas de voitures : l'unité statistique est la même. D'autre part les voitures des deux échantillons sont toutes décrites par le prix et la longueur : les descripteurs sont de même nature d'un échantillon à l'autre.

Il n'y aurait pas lieu de mettre en cause une procédure de classification indépendante des échantillons $\mathbf{x}^1$ et $\mathbf{x}^2$ si les individus de $\mathbf{x}^1$ étaient des canards décrits par leur masse et leur taille, et ceux de $\mathbf{x}^2$, des pays décrits par leur Produit Intérieur Brut et leur densité. Mais classifier $\mathbf{x}^1$ et $\mathbf{x}^2$ de façon indépendante, alors que leur unité statistique est de même nature et que leurs descripteurs ont le même sens, c'est perdre une information qui pourrait être utile à la classification. Dans ce contexte la classification simultanée propose au contraire, de formaliser (en un sens à déterminer) l'information commune

aux échantillons $\mathbf{x}^1$ et $\mathbf{x}^2$ par un lien entre les populations dont ils sont issus, puis de les classifier sous l'hypothèse de ce lien.

Cet exemple met en évidence une situation de classification simultanée dans un contexte commercial ou industriel. Citons deux situations, en biologie, propices à la classification simultanée.

Exemples dans un contexte biologique

Les ornithologues s'intéressent de façon récurrente à la détermination du sexe des oiseaux (on parle de sexage), pour étudier par exemple la proportion de mâles et de femelles dans une espèce. Dans ce contexte, des méthodes statistiques comme l'analyse discriminante ([51]) et la classification automatique ([55]) font directement concurrence aux méthodes moléculaires fiables mais coûteuses. La classification (qu'elle soit supervisée ou non) consiste alors, dans un échantillon d'oiseaux décrit par des variables biométriques (continues), à déterminer deux groupes en inférant le paramètre d'un mélange (gaussien) (voir Partie I). La procédure repose sur l'hypothèse que tous les oiseaux de l'échantillon proviennent de la même population. Or, si les oiseaux à sexer proviennent d'une même espèce mais de sous-espèces différentes, il peut être utile (i) de distinguer plusieurs échantillons d'oiseaux (un échantillon par sous-espèce) et (ii) de formaliser leur appartenance à une espèce commune par un lien entre les populations.

Notre second exemple repose sur le même principe mais la structure recherchée est différente. On ignore souvent sur quels critères se constituent les couples dans une espèce animale. F. D'Amico (2010) ([30]) montre sur un échantillon de cincles - une espèce monogame d'oiseaux de l'ordre des passereaux - une préférence homotypique basée sur la couleur du plastron (tâche colorée qui orne la poitrine du Cincle). Or l'échantillon d'étude comporte des oiseaux pyrénéens et irlandais. Pour retrouver les couples d'oiseaux, il peut être utile à notre avis (i) de considérer les cincles des deux régions comme deux échantillons distincts et (ii) de lier les populations de façon à formaliser qu'il s'agit dans tous les cas de cincles.

L'exemple des oiseaux et celui des voitures montrent que l'on peut souvent mettre en évidence dans un échantillon à classifier, l'existence de plusieurs échantillons distincts. La classification simultanée repose alors sur un compromis entre deux hypothèses antagonistes. Dans la première hypothèse les échantillons à classifier proviennent de la même population. Dans la seconde hypothèse, les populations dont sont issus les échantillons n'ont aucun type de relation. La classification simultanée, en supposant un lien entre les populations, propose une hypothèse alternative à ces deux situations opposées, et prétend être dans beaucoup de cas, plus réaliste que chacune d'elles.

L'existence d'échantillons distincts dans un jeu de données est parfois naturelle et n'a pas besoin d'être mise en évidence. C'est le cas lorsque des données ont été recueillies à des instants distincts pas exemple. A ce titre les séries chronologiques constituent un contexte approprié à la classification simultanée.

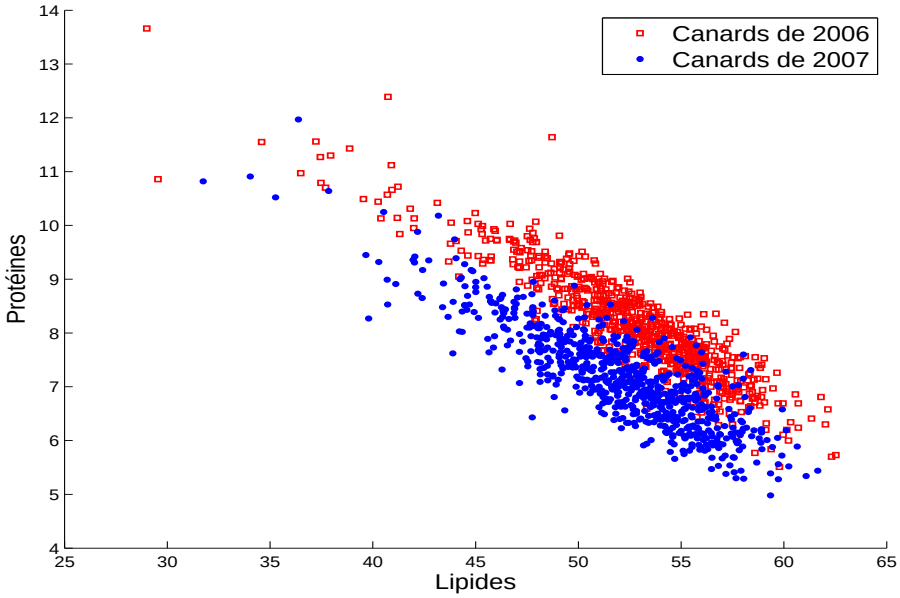


FIGURE 3.2: *Deux échantillons de foies de canards*

Classification simultanée et séries chronologiques

La figure 3.2 représente deux échantillons de canards ayant servi en 2006 et 2007 à la production de foies gras. Ces données proviennent de l'I.N.R.A. de Toulouse [1]. Elles ont été recueillies dans le cadre d'une étude sur la détermination des facteurs responsables du *taux de fonte*. Certains foies de canards perdent une partie importante de leur masse lors de la cuisson et sont inaptes à produire un foie gras. Or on ne connaît pas les facteurs génétiques responsables du taux de fonte. On ne sait pas non plus prédire avant cuisson si un foie présentera un taux de fonte supérieur au seuil requis. La proportion de foies inaptes, bien que faible, engendre un coût non négligeable à un niveau industriel et motive cette étude.

Nous n'apportons pas ici, de solution à la détermination des foies aptes ou inaptes. Nous remarquons juste que la classification simultanée des canards de 2006 et de 2007 est une méthode indiquée si l'on cherche une structure de l'un ou l'autre des échantillons.

L'asymétrie de la queue de sa distribution (voir FIG. 3.2) donne à penser que chaque échantillon est hétérogène. Ainsi, quelle que soit l'année les lipides et les protéines mesurés sur les foies, révèleraient deux groupes de canards (peut-être davantage).

Les populations de 2006 et 2007 sont manifestement différentes. On supposerait donc pour classifier les deux échantillons qu'ils proviennent de mélanges distincts. Pour autant les populations des deux années ne sont pas sans lien : les nuages sont proches par

1. programme ANR Genecan, co financé par AGENAVI et les régions Midi-Pyrénées et Aquitaine

la forme et ils présentent en particulier une zone de faible densité dans la même région du plan correspondant à un taux de lipides faible et un taux de protéines élevé.

Les échantillons eux non plus ne sont pas sans lien : (i) ils sont constitués de canards dans les deux cas, (ii) ils sont décrits pas des variables de même signification et (iii) leur taille est comparable.

Ces ressemblances entre les populations d'une part, entre les échantillons d'autre part, nous incitent si l'on cherche à déterminer des groupes parmi les canards des deux années, à le faire par une méthode de classification simultanée.

Méthodologie et justification du principe

La classification simultanée basée sur des mélanges consiste à : (i) mettre en lumière dans des données à classifier (lorsque ce n'est pas fait de façon naturelle), plusieurs échantillons provenant éventuellement de populations différentes, (ii) essayer de formaliser par un lien réaliste entre populations toute information commune aux différents échantillons (envisager éventuellement plusieurs formes de lien), (iii) inférer le paramètre du modèle sous la contrainte de ce lien, (iv) choisir un modèle parmi les modèles de lien disponibles, et (v) classer les données des différents échantillons.

Le lien établi entre populations prendra la forme d'une contrainte imposée au paramètre des mélanges.
Classifier les échantillons comme s'ils provenaient de la même population revient à supposer que la contrainte est maximale : les mélanges ont le même paramètre.
Classifier les échantillons de façon indépendante revient à supposer que la contrainte est lâche : le paramètre des mélanges n'est soumis qu'aux contraintes propres à chaque population.
La classification simultanée fait un pari. Elle parie qu'il existe un état du lien, intermédiaire à ces deux états limites et plus réaliste que chacun d'eux.

Le lien entre populations est censé diminuer le biais du modèle global. Dans la mesure ou il est réaliste, il doit donc améliorer (en moyenne) d'une part la classification indépendante des échantillons (sans l'hypothèse de ce lien) mais également leur classification sous l'hypothèse d'une origine commune.

3.2 Etat de l'art

La classification simultanée s'inscrit dans la tradition de nombreuses méthodes statistiques dédiées initialement à un échantillon, puis étendues à plusieurs échantillons.

Analyse en composantes principales simultanée

On peut citer en premier lieu l'Analyse en Composantes Principales (ACP) simultanée de B.N. Flury ([41]). L'ACP revient à projeter les points d'un échantillon, sur les axes spectraux de sa matrice des covariances. Quand on dispose de plusieurs échantillons, Flury propose de procéder à leur ACP (simultanée) en supposant que leur matrice des covariances possède la même orientation. Ainsi l'ACP simultanée de plusieurs échantillons repose sur l'hypothèse que les axes principaux ont deux à deux la même direction, hypothèse dont Flury affirme qu'elle est plausible en biologie, dans de nombreux cas.

Analyse procrustéenne généralisée

On peut citer ensuite l'analyse procrustéenne généralisée de Gower ([50]). L'analyse procrustéenne est une méthode pemettant d'estimer l'écart entre deux tableaux $\boldsymbol{A}$ et $\boldsymbol{B}$ de même dimension $d \times n$. Elle se propose de déterminer la rotation $\boldsymbol{S}$ (matrice orthogonale de $\mathbb{R}^{d\times d}$), la dilatation λ (scalaire de $\mathbb{R}_*^+$) et la translation $\boldsymbol{u}$ (vecteur de $\mathbb{R}^d$) qui minimisent le critère :

$$\mathrm{tr}\left[(\boldsymbol{B} - \lambda \boldsymbol{S}\boldsymbol{A} - \boldsymbol{u}\boldsymbol{1}_n')'\boldsymbol{M}(\boldsymbol{B} - \lambda \boldsymbol{S}\boldsymbol{A} - \boldsymbol{u}\boldsymbol{1}_n')\right] \tag{3.1}$$

c'est à dire l'écart - au sens d'une métrique $\boldsymbol{M}$ (matrice de $\mathbb{R}^{d\times d}$ symétrique, définie et positive) - entre le tableau $\boldsymbol{B}$ et le tableau transformé $\lambda\boldsymbol{S}\boldsymbol{A} + \boldsymbol{u}\boldsymbol{1}_n'$. Groenen et Franses (2000) ([53]) emploient cette méthode pour représenter (et comparer) les valeurs de n cotations sur d places boursières à deux instants différents. L'analyse procrustéenne généralisée ([50]) cherche à déterminer, lorsque $\boldsymbol{A}_k$ $(k = 1, \ldots, K)$ sont des tableaux de $\mathbb{R}^{d\times n}$, les rotations $\boldsymbol{S}_k$ $(k = 1, \ldots, K)$ les dilatations λ_k et les translations $\boldsymbol{u}_k$ qui minimisent la distance entre les tableaux transformés $\tilde{\boldsymbol{A}}_k = \lambda_k \boldsymbol{S}_k \boldsymbol{A}_k + \boldsymbol{u}_k \boldsymbol{1}_n'$ $(k = 1, \ldots, K)$ c'est à dire le critère :

$$\sum_{k=1}^{K}\sum_{j=1}^{K} \mathrm{tr}\left[(\tilde{\boldsymbol{A}}_k - \tilde{\boldsymbol{A}}_j)'\boldsymbol{M}(\tilde{\boldsymbol{A}}_k - \tilde{\boldsymbol{A}}_j)\right]. \tag{3.2}$$

Les solutions proposées par Kristof et Wingersky (1971) ([72]), par Gower (1975) ([50]) et Ten Berge (1977) ([101]) reposent toutes sur une optimisation alternée du critère (3.2) et peuvent être vues comme un enchaînement d'analyses procrustéennes simples.

Tenenhaus et Vinzi (2005) ([103]) s'inspirent de cette méthode pour déterminer les matrices de rotations qui lient les facteurs PLS aux variables latentes dans une analyse multi-blocs.

Modèles hiérarchiques pour des partitions emboîtées

Vermunt (2005) ([107]) propose un modèle hiérarchique permettant de déceler d'éventuelles partitions emboîtées dans un jeu de données. Si on demande aux employés de plusieurs entreprises de répondre à un questionnaire et que dans chaque entreprise les employés sont répartis en équipes, les réponses fournies permettent-elles de déterminer une structure parmi les entreprises (H groupes d'entreprises), une autre parmi les

équipes (K groupes d'équipes), et de prévoir à quel regroupement d'entreprises et à quel regroupement d'équipes appartient chaque employé ? Un modèle hiérarchique appliqué à des mélanges de classes latentes permet de répondre à ce problème, mais Vermunt ([107]) propose un modèle plus général, qui s'étend à des données continues, et dont la densité est :

$$f(\bullet\,;\boldsymbol{\theta}) = \sum_{h=1}^{H} \pi^h \sum_{k=1}^{K} \pi_{k|h} f_k(\bullet\,;\boldsymbol{\theta}_k). \tag{3.3}$$

Ce modèle suppose l'existence de deux variables latentes catégorielles multinomiales $\mathbf{Y} \in \{0,1\}^H$ (qui indique le groupe d'entreprises auquel appartient une réponse) dont la h^{e} composante Y^h vaut 1 (et les autres 0) avec probabilité π^h, et $\mathbf{Z} \in \{0,1\}^K$ (qui indique le groupe d'équipes auquel appartient une réponse) dont la k^{e} composante Z_k vaut 1 (et les autres 0) avec probabilité $\pi_{k|h}$ lorsque $Y^h = 1$. $f_k(\bullet;\boldsymbol{\theta}_k)$ est la densité du vecteur aléatoire $\mathbf{X}$ (qui contient la réponse) lorsque $Z_k = 1$ (La loi de $\mathbf{X}$ est supposée indépendante de $\mathbf{Y}$.).

Galimberti et Soffritti (2007) ([47]) emploient ce modèle pour déterminer une structure hiérarchique parmi cent trois provinces italiennes regroupées par régions, et décrites par des variables socio-économiques (continues).

Quand les appartenances de niveau 2 (valeurs de la variables $\mathbf{Y}$) sont connues, la classification hiérarchique est strictement équivalente (dans le cas de populations conditionnelles gaussiennes) à notre méthode de classification simultanée dans le cas le plus simple de lien affine entre populations. Elle consiste à classifier les échantillons en supposant que les mélanges qui les modélisent ne diffèrent que par le poids des composantes.

Modèles adaptatifs de régression

C. Bouveyron et J. Jacques (2010) ([22]) appliquent le principe d'un lien entre populations à des modèles de régression linéaire.

Considérons deux modèles de régression linéaire :

$$Y^h \sim \beta_0^h + (\boldsymbol{\beta}^h)'\mathbf{X}^h\ ;\ h = 1,2, \tag{3.4}$$

où $Y^h \in \mathbb{R}$ est une variable à expliquer, $\mathbf{X}^h \in \mathbb{R}^d$ un vecteur de prédicteurs et $(\beta_0^h, \boldsymbol{\beta}^h) \in \mathbb{R} \times \mathbb{R}^d$, le paramètre de la régression h. Il arrive très souvent que l'on estime les paramètres $(\beta_0^1, \boldsymbol{\beta}^1)$ et $(\beta_0^2, \boldsymbol{\beta}^2)$ de façon indépendante alors que les échantillons d'apprentissage ne sont pas sans lien. Y^1 et Y^2 peuvent désigner le prix d'appartements dans deux régions différentes par exemple et les composantes des vecteurs $\mathbf{X}^1$ et $\mathbf{X}^2$ peuvent être des prédicteurs de même sens (nombre de pièces, loyer etc.) dans les deux régions. C. Bouveyron et J. Jacques (2010) proposent dans ce contexte d'établir un lien entre les paramètres des deux modèles de régression. L'un des modèles qu'ils proposent par exemple, suppose que les régresseurs sont proportionnels : $\beta_0^1\boldsymbol{\beta}^2 = \beta_0^2\boldsymbol{\beta}^1$. Dans ce cas deux valeurs égales des prédicteurs ($\mathbf{x}^1 = \mathbf{x}^2$) conduisent à des prix β_0^2/β_0^1 fois plus élevés dans la région 2 que dans la région 1.

Cette méthode est une alternative réaliste à la régression linéaire classique. Pour mettre en évidence son intérêt, reprenons l'exemple précédent où le prix d'appartements

est expliqué dans deux régions différentes par des prédicteurs de même nature (nombre de pièces, loyer etc.).

Il n'y a aucune raison d'abord, de penser que le prix d'un appartement soit déterminé par ses prédicteurs de la même façon dans les deux régions. Un modèle de régression commun aux deux régions est donc sans doute trop fruste.

On peut envisager deux modèles de régression différents, un dans chaque région. Mais alors les données de la région h ne servent à estimer (dans une conception classique de la régression) que le paramètre de la régression h. Lorsque les données d'apprentissage dans une région sont peu nombreuses, les régresseurs estimés dans cette région sont donc fortement liés aux fluctuations d'échantillonnage.

Imposer une contrainte aux régresseurs des deux modèles diminue la variabilité de leurs estimateurs. Les contraintes envisagées par C. Bouveyron et J. Jacques (2010) s'appuient sur une information véritable des données d'apprentissage : (i) les prédicteurs (nombre de pièces, loyer etc.) sont de même nature dans les deux régions et (ii) les variables expliquées (prix de l'appartement) aussi. Pour cette raison les modèles qu'ils proposent permettent de diminuer conjointement le biais du modèle et la variabilité des régresseurs.

Analyse discriminante généralisée

Dans ce qui suit nous plaçons la classification simultanée dans un contexte non supervisé. Mais la méthode s'inspire de modèles initialement pensés dans un contexte d'analyse discriminante.

En analyse discriminante et en classification semi-supervisée on estime une règle de classement sur des individus dont on connaît l'origine (échantillon d'apprentissage) ; cette règle sert ensuite à déterminer l'origine d'individus sans label (échantillon de test). On suppose implicitement dans cette procédure, que l'échantillon de test et l'échantillon d'apprentissage proviennent de la même population.

Biernacki et al. (2002) ([15]) envisagent que dans certains cas l'échantillon de test $\mathbf{x}^1$ et l'échantillon d'apprentissage $\mathbf{x}^2$ puissent provenir de populations différentes. Ils proposent alors d'établir un lien paramétrique entre les deux populations. L'estimation de ce lien permet de transformer la règle de classement des données d'apprentissage, r^1, en une règle de classement différente pour les données de test, r^2.
Biernacki et al. montrent l'intérêt de cette méthode, appelée analyse discriminante généralisée (ADG), dans un contexte biologique : ils déterminent le sexe d'oiseaux *diomedea* (une sous-espèce de puffins cendrés) en les liant stochastiquement à d'autres puffins, *borealis*, dont ils connaissent le sexe.

Les oiseaux *borealis* et *diomedea* dans [15] sont décrits par des variables continues ; ils

sont supposés provenir de mélanges gaussiens. J. Jacques (2005) ([63]) propose d'étendre l'ADG à des variables binaires. Dans une première étude il discrétise les variables biométriques qui décrivent les oiseaux de [15]. Les échantillons de test (*diomedea*) et d'apprentissage (*borealis*) sont modélisés non plus par des mélanges gaussiens mais par deux mélanges de modèles de classes latentes. Il compare alors trois méthodes : l'analyse discriminante généralisée (qui détermine le sexe des *diomedea* à partir de celui des *borealis* sous l'hypothèse d'un lien paramétrique entre les deux populations), l'analyse discriminante traditionnelle (qui suppose que *borealis* et *diomedea* proviennent de la même population) et la classification automatique des *diomedea* (qui considère que connaître le sexe des *borealis* n'apporte aucune information sur le sexe des *diomedea*). Il montre que l'analyse discriminante généralisée est, de ces trois méthodes, celle qui donne le meilleur taux d'erreur de classement des *diomedea*.

J. Jacques (2005) ([63]) montre dans une seconde étude la pertinence de l'analyse discriminante généralisée, sur des données discrètes (et non discrétisées) issues de la médecine. Il dispose de deux échantillons de patients hospitalisés à des époques différentes pour un cancer du testicule et décrits par des variables binaires (tumeurs séminomateuse ou non, position par rapport à l'âge médian etc.). En établissant un lien stochastique entre les patients des deux échantillons, J. Jacques prédit le risque de deuxième cancer des uns connaissant les cas de rechute chez les autres.

On trouvera dans [64] un autre exemple d'analyse discriminante généralisée sur données réelles binaires. Il s'agit à nouveau de sexer un échantillon d'oiseaux, de l'Atlantique, connaissant le sexe d'oiseaux du Pacifique. Les oiseaux des deux échantillons sont décrits dans [64] par des variables discrètes comme l'auto-coloration ou non de la plume sous-caudale, la présence ou l'absence d'une bande colorée.

Modèles liés de régression logistique

Notons enfin l'extension dans [10] de l'analyse discriminante généralisée aux modèles de Logit.
Les échantillons $\mathbf{x}^1$ et $\mathbf{x}^2$ sont dans un espace de dimension d et l'on suppose qu'ils recèlent une partition à deux classes notées 0 et 1. On peut supposer que la frontière discriminante liée aux règles de classement r^1 et r^2 est linéaire. Elle l'est par exemple lorsque les échantillons $\mathbf{x}^1$ et $\mathbf{x}^2$ sont supposés provenir de mélanges gaussiens homoscédastiques, mais dans de nombreux autres cas également (voir Anderson (1982) [4]). Il existe donc $\beta_0^i \in \mathbb{R}$ et $\boldsymbol{\beta}^i \in \mathbb{R}^d$ $(i = 1, 2)$ tels que pour tout $\boldsymbol{x} \in \mathbb{R}^d$: $r^i(\boldsymbol{x}) = 1$ si $\beta_0^i + \boldsymbol{x}'\boldsymbol{\beta}^i \geq 0$ et 0 sinon. F. Beninel et C. Biernacki (2007) ([10]) proposent d'imposer directement une contrainte sur le paramètre des Logit $(\beta_0^i, \boldsymbol{\beta}^i)$ $(i = 1, 2)$ c'est à dire d'établir un lien entre les frontières discriminantes et non plus entre les populations.

3.3 Plan

Les prémisses de la classification simultaneé consistent (i) à ramener le partitionnement d'un échantillon à la classification de plusieurs échantillons, puis (ii) à établir un lien entre la population d'origine des différents échantillons. La méthode peut donc se décliner sous de multiples formes. D'une part le nombre des mélanges permettant de modéliser les échantillons est important (Le chapitre 1 ainsi que [85] en donnent un aperçu.). D'autre part le lien entre populations a pour seul impératif d'être justifié, ce qui ouvre le champ à de nombreuses possibilités.

Nous nous bornons dans ce travail à montrer la pertinence de la classification simultanée de données continues, dans quelques situations assez générales.

Classification simultanée et mélanges gaussiens

Nous supposons à la section 4.1 que les données de chaque échantillon proviennent d'un mélange gaussien. Dans ce contexte nous établirons un lien stochastique entre les données catégorielles, qui peut s'interpréter géométriquement : les populations conditionnelles se tranforment deux à deux l'une en l'autre, de façon affine. Ce lien géométrique entre populations conditionelles sera justifié. Il permettra de définir des modèles parcimonieux significatifs dont nous montrerons l'utilité pour sexer des échantillons d'oiseaux.

Classification simultanée robuste

La loi normale est sensible aux données atypiques, aberrantes ou bruitées. Lorsque les échantillons à classifier comportent de telles données, il peut être judicieux de supposer à sa place que les populations conditionnelles suivent une loi de Student multivariée. Cette loi suscite un intérêt récent d'une part parce qu'elle est moins sensible aux fluctuations d'échantillonnage que la loi normale et d'autre part parce qu'elle apparaît comme une généralisation de la loi normale (voir [44]). Nous envisagerons donc à la section 4.3 un avatar de la classification simultanée basé sur des mélanges de Student, dont la vocation est d'être robuste à la présence de bruit dans la donnée. On supposera à nouveau un lien stochastique affine entre populations conditionnelles et l'on montrera l'intérêt d'un tel classifieur robuste sur des données financières.

Classification simultanée en grande dimension

Nous évoquerons à la section 4.4, la classification simultanée de données de grande dimension.

Il n'est pas rare - nous l'avons vu précédemment - que l'on puisse distinguer dans un échantillon $\mathbf{x}$ à classifier, deux échantillons (parfois plus) $\mathbf{x}^1$ et $\mathbf{x}^2$. Si $\mathbf{x}$ est situé dans un espace de grande dimension, on peut être tenté afin de réduire la taille du paramètre

à estimer, de considérer que $\mathbf{x}^1$ et $\mathbf{x}^2$ proviennent du même mélange. Mais cette hypothèse augmente le biais du modèle (y compris dans un espace de grande dimension) si les échantillons $\mathbf{x}^1$ et $\mathbf{x}^2$ proviennent clairement de populations distinctes. Si l'on suppose au contraire que les deux échantillons proviennent bien de mélanges différents, la variabilité des estimateurs - déjà importante à cause de la dimension de l'espace des données - augmente encore à cause de la multiplicité des populations. Le contexte de la grande dimension exacerbe donc l'importance du compromis à trouver entre le biais du modèle et la variabilité des estimateurs. A ce titre, des modèles de lien entre populations, semblent trouver un intérêt particulier dans le contexte de la grande dimension.

Un des modèles classiques des données de grande dimension est celui des Factor Analyzers (voir [85] et Section 1.1.4). Il s'agit de distributions normales particulières qui permettent de formaliser le manque d'information qu'apporte la donnée dans certaines directions d'un espace de grande dimension. Nous évoquerons donc à la section 4.4, la possibilité de modéliser et de classifier plusieurs échantillons en grande dimension, en établissant un lien stochastique affine conditionnel entre mélanges de Factor Analyzers.

Classification simultanée basée sur un recouvrement égal des classes

Dans les formes précédentes de la classification simultanée le lien entre populations est conditionnel et il s'interprète comme une transformation stochastique des données catégorielles.
Nous envisagerons au chapitre 5 un lien non plus conditionnel mais global, basé sur un chevauchement identique des classes dans les différents mélanges.

Lorsqu'on observe une difficulté similaire à classifier des échantillons, on peut supposer que l'imbrication des données conditionnelles est la même dans chaque échantillon et traduire cette information en supposant que le recouvrement des classes est le même d'un mélange à l'autre.

Ainsi nous supposerons à la section 5.2 que les mélanges possèdent le même taux d'erreur de classement, et à la section 5.3 que l'entropie gobale de leurs composantes est homogène.

Alors que la classification simultanée repose au chapitre 4 sur des mélanges spécifiques (mélanges gaussiens, mélanges de Student, mélanges de Factor Analyzers) nous proposons à la section 5.3 une forme de la classification simultanée destinée à n'importe quel type de mélange. Nous définissons en effet un algorithme générique $\tilde{\text{EM}}$ dérivé d'EM, qui permet d'estimer le paramètre de mélanges en supposant que l'entropie globale de leurs composantes est la même, et sans faire aucune hypothèse sur la spécificité des populations conditionnelles.

Chapitre 4

Lien affine stochastique entre populations

La section qui suit fait l'objet d'un article en cours de publication ([78]). Cet article est rédigé en anglais et nous proposons d'en reprendre exactement la structure.

4.1 Mélanges gaussiens (cas général)

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 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.

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

4.1.1 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 [85] for a review). In particular, in the multivariate continuous situation, Gaussian mixture model-based clustering has found successful applications in diverse fields : Genetics [97], medicine [85], magnetic resonance imaging [7], astronomy [28]. 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. [104] 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 [98] 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 [41], 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 [50] generalizes to K samples ($K \geq 3$) the classical Procrustes analysis which estimates a geometrical link, established between two samples. Hierarchical mixture models [107] 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 4.1.2, 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 4.1.3. Section 4.1.4 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 4.1.5. Finally in Section 4.1.6 we plan extensions of this work.

4.1.2 From independent to simultaneous Gaussian clustering

We aim to separate H samples into K groups. Describing standard Gaussian model-based clustering (Subsection 4.1.2.1) in this apparently more complex context (H samples instead of one), will be later convenient for introducing simultaneous Gaussian model-based clustering (Subsection 4.1.2.2). 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 $\mathbf{x}^h$ ($h \in \{1, \dots, H\}$) is composed of n^h individuals $\mathbf{x}_i^h$ ($i = 1, \dots, n^h$) of $\mathbb{R}^d$, and arises from a population P^h. In addition, all populations are described by the same d continuous variables.

4.1.2.1 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 ($k = 1, \dots, K$), with probability density function :

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

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 $\Phi_d(\boldsymbol{x}; \boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h)$ 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 $\boldsymbol{\psi}_k^h = (\pi_k^h, \boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h)$.
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 $(\pi_1^h, \dots, \pi_K^h)$.
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 k-th component Z_k^h equals 1 (and the others 0) with probability π_k^h, and $\left(\boldsymbol{X}^h | Z_k^h = 1\right) \sim \Phi_d(\ . \ ; \boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h)$. 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 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 [85] 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 $\hat{\boldsymbol{\psi}}$:

$$t_{i,k}^h(\hat{\boldsymbol{\psi}}) = E(Z_k^h | \boldsymbol{X}^h = \mathbf{x}_i^h; \hat{\boldsymbol{\psi}}). \tag{4.1}$$

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.

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

$$(\boldsymbol{X}^{h'} | Z_k^{h'} = 1) \sim \xi_k^{h,h'} \left(\boldsymbol{X}^h | Z_k^h = 1 \right). \tag{4.2}$$

This model implies 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 ($j \in \{1, \dots, d\}$) 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 :

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

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 $\mathscr{H}_2$–, then the only possible transformation is an affine map. Indeed, De Meyer et al. [35] 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 $(h, h') \in \{1, \dots, H\}^2$ and all $k \in \{1, \dots, K\}$, 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 :

$$(\boldsymbol{X}^{h'} | Z_k^{h'} = 1) \sim \boldsymbol{D}_k^{h,h'} (\boldsymbol{X}^h | Z_k^h = 1) + \boldsymbol{b}_k^{h,h'}. \tag{4.3}$$

Relation (4.2) constitutes the keystone of the simultaneous Gaussian model-based clustering framework, and (4.3) is its affine form involved from the two previous hypotheses $\mathscr{H}_1$ and $\mathscr{H}_2$.

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 $\mathscr{H}_3$–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 4.1.5), this assumption may be weakened as we remark it at the end of Subsection 4.1.4.4.

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 :

$$\boldsymbol{\Sigma}_k^{h'} = \boldsymbol{D}_k^{h,h'}\ \boldsymbol{\Sigma}_k^h\ \boldsymbol{D}_k^{h,h'} \quad \text{and} \quad \boldsymbol{\mu}_k^{h'} = \boldsymbol{D}_k^{h,h'}\boldsymbol{\mu}_k^h + \boldsymbol{b}_k^{h,h'}. \tag{4.4}$$

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

4.1.2.3 A useful and statistically meaningful interpretation of the linear stochastic link

Each covariance matrix can be decomposed into :

$$\boldsymbol{\Sigma}_k^h = \boldsymbol{T}_k^h \boldsymbol{R}_k^h \boldsymbol{T}_k^h, \tag{4.5}$$

where $\boldsymbol{T}_k^h$ is the diagonal matrix of conditional standard deviations in C_k^h component–for all $(i,j) \in \{1,\dots,d\}^2$: $\boldsymbol{T}_k^h(i,j) = \sqrt{\boldsymbol{\Sigma}_k^h(i,j)}$ 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 (4.5) is unique, Relation (4.4) involves for every $(h,h') \in \{1,\dots,H\}^2$ and every $k \in \{1,\dots,K\}$ 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 (4.3) is equivalent therefore to postulating that conditional correlations are equal through the populations.

This interpretation of the affine link between the conditional populations (4.3) allows the model to keep all its sense when simultaneous clustering is envisaged in a relaxed context–as in Subsection 4.1.5.4–where the samples to be classified are described by different descriptor sets.

4.1.3 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 (4.4) between conditional populations.

4.1.3.1 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 [85], 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 [28] 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.

4.1.3.2 Interpopulation models

Thus we can also imagine some meaningful constraints on the parametric link (4.4). 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 (4.4), 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 (4.4) becomes :

$$\boldsymbol{\Sigma}_k^{h'} = \boldsymbol{D}_k^{h,h'}\ \boldsymbol{\Sigma}_k^h\ \boldsymbol{D}_k^{h,h'} \quad \text{and} \quad \boldsymbol{\mu}_k^{h'} = \boldsymbol{D}_k^{h,h'}\boldsymbol{\mu}_k^h. \tag{4.6}$$

As the first equality involves the following relation between the conditional standard

deviation matrices :

$$\boldsymbol{T}_k^{h'} = \boldsymbol{D}_k^{h,h'}\, \boldsymbol{T}_k^h, \tag{4.7}$$

we deduce then from the second one :

$$\left(\boldsymbol{T}_k^{h'}\right)^{-1} \boldsymbol{\mu}_k^{h'} = \left(\boldsymbol{T}_k^h\right)^{-1} \boldsymbol{\mu}_k^h. \tag{4.8}$$

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.

4.1.3.3 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 (4.4)–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 4.1 displays all allowed combinations of intra and interpopulation models.

4.1.3.4 Requirements about identifiability

For a given permutation σ in $\mathscr{S}_H$ (symmetric group on $\{1, \ldots, H\}$), and another one τ in $\mathscr{S}_K$, $\boldsymbol{\psi}_\tau^\sigma$ 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 :

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

TABLE 4.1: *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		$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
$\boldsymbol{I}$, $\alpha^{h,h'}\boldsymbol{I}$, $\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	• (.)	• (.)	• (•)	• (•)
	$\boldsymbol{b}^{h,h'}$	• (.)	• (.)	• (•)	• (•)
	$\boldsymbol{b}_k^{h,h'}$	∘ (.)	• (.)	• (•)	• (•)
π (π^h)					
$\alpha_k^{h,h'}\boldsymbol{I}$, $\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	. (.)	• (.)	. (.)	• (•)
	$\boldsymbol{b}_k^{h,h'}$	. (.)	• (.)	. (.)	• (•)

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 (4.4) 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 4.1.

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.

4.1.3.5 Model selection

In a parametric model-based clustering context the BIC criterion (see [98] and see also [73] 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 [93], [45].

TABLE 4.2: *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 $\gamma = \dfrac{d^2+d}{2}$ 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.*

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

The BIC of a model is defined here by :

$$BIC = -\ell(\hat{\boldsymbol{\psi}}; \mathbf{x}) + \frac{\nu}{2}\log(n), \tag{4.9}$$

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 4.2 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 4.1.2.1) and simultaneous clustering (Subsection 4.1.2.2).

4.1.4 Parameter estimation

After a useful reparameterization (Subsection 4.1.4.1), a GEM procedure for estimating the model parameters by maximum likelihood is described in Subsections 4.1.4.2 to 4.1.4.4. An alternative and simplified estimation process is proposed then, in Subsection 4.1.4.5, for some specific models.

4.1.4.1 A useful reparameterization

The parametric link between the Gaussian parameters (4.4) 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 $h \in \{2, \dots, H\}$, $\boldsymbol{\theta}^h = \left[\left(\pi_k^h, \boldsymbol{D}_k^h, \boldsymbol{b}_k^h\right); k = 1, \dots, K\right]$, where $\boldsymbol{D}_k^h = \boldsymbol{D}_k^{1,h}$ and $\boldsymbol{b}_k^h = \boldsymbol{b}_k^{1,h}$. Let us note Θ the space described by the vector $\boldsymbol{\theta} = (\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^H)$ when $\boldsymbol{\psi}$ 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 $\boldsymbol{\psi}$ 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 P^1 population. Indeed the bijective correspondance between the parameter spaces Θ and Ψ ensures that the model inference is invariant by relabelling the populations.

4.1.4.2 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 :

$$l_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{h=1}^{H} \sum_{i=1}^{n^h} \sum_{k=1}^{K} z_{i,k}^h \log\left(\pi_k^h \Phi_d\left(\mathbf{x}_i^h; \boldsymbol{D}_k^h \boldsymbol{\mu}_k^1 + \boldsymbol{b}_k^h, \boldsymbol{D}_k^h \boldsymbol{\Sigma}_k^1 \boldsymbol{D}_k^h\right)\right), \tag{4.10}$$

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. [37] 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 $E\left[l_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) | \boldsymbol{X} = \mathbf{x}\right]$ 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 (4.1) 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 (4.10), 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.

4.1.4.3 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 :

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

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

– *Vectors* $\boldsymbol{b}_k^h$

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

$$\boldsymbol{b}_k^{h+} = \left[\sum_{k=1}^{K} \hat{n}_k^h \left(\boldsymbol{D}_k^h \boldsymbol{\Sigma}_k^{1+} \boldsymbol{D}_k^h \right)^{-1} \right]^{-1} \left[\sum_{k=1}^{K} \hat{n}_k^h \left(\boldsymbol{D}_k^h \boldsymbol{\Sigma}_k^{1+} \boldsymbol{D}_k^h \right)^{-1} \left(\bar{\mathbf{x}}_k^h - \boldsymbol{D}_k^h \boldsymbol{\mu}_k^{1+} \right) \right], \tag{4.11}$$

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 :

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

vhere

– $u_k^h = \sum_{i=1}^{n^h} t_{i,k}^h \left(\mathbf{x}_i^h - \boldsymbol{b}_k^{h+} \right)' \left(\boldsymbol{\Sigma}_k^{1+} \right)^{-1} \boldsymbol{\mu}_k^{1+}$ 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 $E\left[l_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) | \boldsymbol{X} = \mathbf{x}\right]$, keeps on being 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.

4.1.4.5 An alternative sequential estimate

According to Subsections 4.1.4.3 and 4.1.4.4, $\boldsymbol{\psi}$ 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 $\boldsymbol{b}^{h,h'} = \boldsymbol{0}$) the conditional link (4.3) stretches over unconditional populations :

$$\boldsymbol{X}^{h'} \sim \boldsymbol{D}^{h,h'}\boldsymbol{X}^{h} + \boldsymbol{b}^{h,h'}. \tag{4.12}$$

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 4.3.

TABLE 4.3: *Link parameter least-square estimates in the sequential estimation method.* $\bar{\mathbf{x}}^h = (1/n^h)\sum_{i=1}^{n^h}\mathbf{x}_i^h$ *and* $\hat{\boldsymbol{S}}^h = (1/n^h)\sum_{i=1}^{n^h}(\mathbf{x}_i^h - \bar{\mathbf{x}}^h)(\mathbf{x}_i^h - \bar{\mathbf{x}}^h)'$ *denote respectively the empirical center and the empirical covariance matrix of the whole population* P^h.

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

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 independently from P^1 population, a simple and traditional EM algorithm devoted to Gaussian mixture estimation, can be involved. Softwares as MIXMOD [17] 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 4.3 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 $(\pi, \boldsymbol{D}^{h,h'}, \boldsymbol{b}^{h,h'})$, $(\pi, \boldsymbol{D}^{h,h'}, \boldsymbol{0})$ 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 4.3. 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 4.3 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 4.1.4.3 and 4.1.4.4).

4.1.5 A biological example

4.1.5.1 The data

In [104] 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 4.1: *Three samples of Cory's Shearwaters described by variables of identical meaning.*

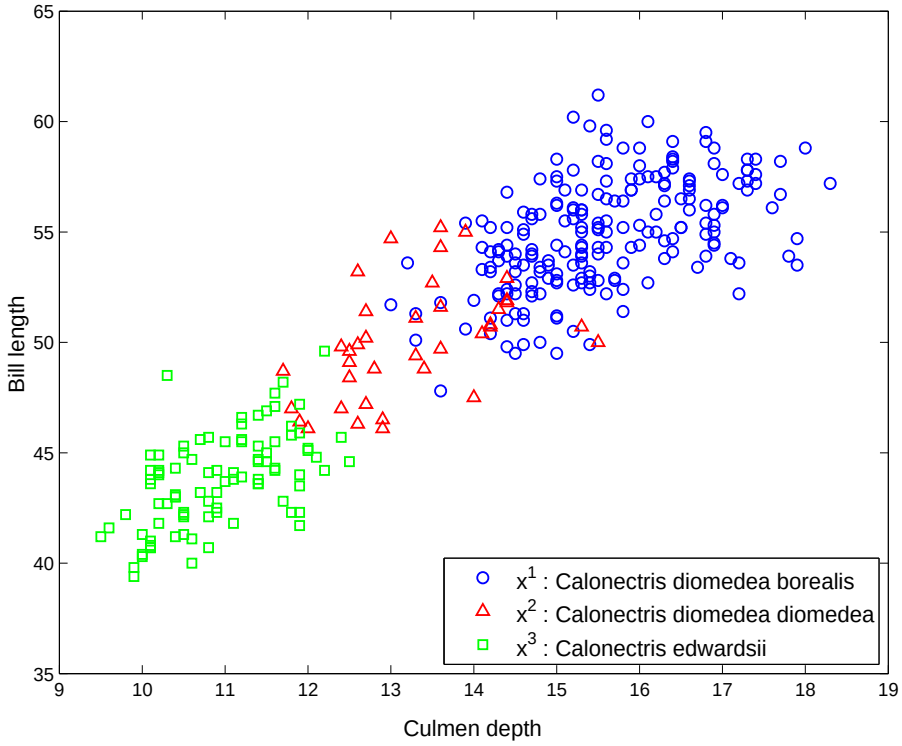


Figure 4.1 displays the birds in the plane of the culmen depth and the bill length.

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.

4.1.5.2 Partitioning when the cluster number is known

The following experiments have been exhibited in [18]. We applied on the three seabird samples each of the 66 allowed models of simultaneaous clustering displayed in Table 4.1. Since the birds must be clustered 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 4.1.4.2)) 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 (4.3) does not hold in the three seabird populations and compute then the BIC value of the model so inferred. In Table 4.4, the BIC values obtained by the independent clustering method, have been computed according to (4.9). 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 4.4 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 [15] 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 [15]).

Remark : Table 4.5 displays BIC values and all associated errors rates obtained by sequential estimation (Subsection 4.1.4.5). BIC values are greater than the corresponding BIC of Table 4.4–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.

TABLE 4.4: *BIC value and (error rate) in simultaneous (full ML estimates) and independent clustering (2 groups) of Shearwaters.*

			π		π_k	
			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
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)
		$\boldsymbol{b}_k^{h,h'}$	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)
		$\boldsymbol{b}_k^{h,h'}$	4079.6 (11.61)	4105.2 (11.90)	4079.9 (40.77)	4095.8 (45.83)
	$\alpha_k^{h,h'}\ \boldsymbol{I}$	$\mathbf{0}$	.	4282.9 (32.14)	.	4279.4 (38.69)
		$\boldsymbol{b}_k^{h,h'}$	.	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)
		$\boldsymbol{b}_k^{h,h'}$	4094.9 (33.33)	4122.2 (11.31)	4101.9 (41.96)	4122.7 (15.77)
	$\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	.	4097.5 (11.90)	.	4099.2 (14.88)
		$\boldsymbol{b}_k^{h,h'}$	.	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)
		$\boldsymbol{b}_k^{h,h'}$	.	.	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)
		$\boldsymbol{b}_k^{h,h'}$	.	.	4081.5 (41.96)	4102.9 (45.83)
	$\alpha_k^{h,h'}\ \boldsymbol{I}$	$\mathbf{0}$	.	.	.	4129.5 (47.32)
		$\boldsymbol{b}_k^{h,h'}$	.	.	.	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)
		$\boldsymbol{b}_k^{h,h'}$	.	.	4107.6 (42.86)	4128.5 (15.18)
	$\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	.	.	.	4101.8 (45.24)
		$\boldsymbol{b}_k^{h,h'}$	.	.	.	4153.6 (16.37)
Independent			**4139.8** (**12.50**)	4218.2 (38.39)	4143.0 (29.17)	4219.7 (40.18)

That example shows that the alternative sequential method can provide for less some

TABLE 4.5: *Sequential estimation : BIC value and (error rate) in simultaneous and independent clustering (2 groups) of Shearwaters.*

			π		π_k	
			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
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)

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.

4.1.5.3 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 and the related results were presented in [76]. 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. 4.6, 4.4, 4.7 and 4.8. Obviously no empirical error rate is displayed when $K \neq 2$.

TABLE 4.6: *BIC value in simultaneous (full ML estimates) and independent clustering (1 group) of Shearwaters.*

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

TABLE 4.7: *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

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

TABLE 4.8: *BIC value in simultaneous (full ML estimates) and independent clustering (4 groups) of Shearwaters.*

			π		π_k	
			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
	$\boldsymbol{I}$	$\mathbf{0}$	4357.8	4429.2	4341.7	4444.6
		$\boldsymbol{b}^{h,h'}$	4075.9	4157.5	4079.9	4162.0
		$\boldsymbol{b}_k^{h,h'}$	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
		$\boldsymbol{b}_k^{h,h'}$	4135.8	4222.3	4137.9	4219.0
π	$\alpha_k^{h,h'}\,\boldsymbol{I}$	$\mathbf{0}$	.	4360.4	.	4362.6
		$\boldsymbol{b}_k^{h,h'}$	.	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
		$\boldsymbol{b}_k^{h,h'}$	4153.7	4243.5	4155.0	4229.0
	$\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	.	4213.9	.	4207.9
		$\boldsymbol{b}_k^{h,h'}$	.	4320.8	.	4304.9
	$\boldsymbol{I}$	$\mathbf{0}$	.	.	4078.7	4169.9
		$\boldsymbol{b}^{h,h'}$	.	.	4084.2	4165.9
		$\boldsymbol{b}_k^{h,h'}$	.	.	4151.5	4220.9
	$\alpha^{h,h'}\,\boldsymbol{I}$	$\mathbf{0}$	.	.	4078.7	4175.7
		$\boldsymbol{b}^{h,h'}$	.	.	4087.3	4163.7
		$\boldsymbol{b}_k^{h,h'}$	.	.	4151.9	4224.5
π^h	$\alpha_k^{h,h'}\,\boldsymbol{I}$	$\mathbf{0}$	.	.	.	4193.2
		$\boldsymbol{b}_k^{h,h'}$	.	.	.	4235.8
	$\boldsymbol{D}^{h,h'}$	$\mathbf{0}$	.	.	4073.1	4155.2
		$\boldsymbol{b}^{h,h'}$	.	.	4107.4	4175.0
		$\boldsymbol{b}_k^{h,h'}$	.	.	4168.7	4243.8
	$\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	.	.	.	4228.5
		$\boldsymbol{b}_k^{h,h'}$	.	.	.	4318.1
Independent			**4159.6**	4363.4	4171.8	4359.3

comparing the best BIC values obtained in both methods, Table 4.9 confirms when $K = 1, 3,$ or 4, that advantage of the simultaneous clustering method on the inde-

TABLE 4.9: *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

pendent 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 4.9, 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 ($K = 2$).

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

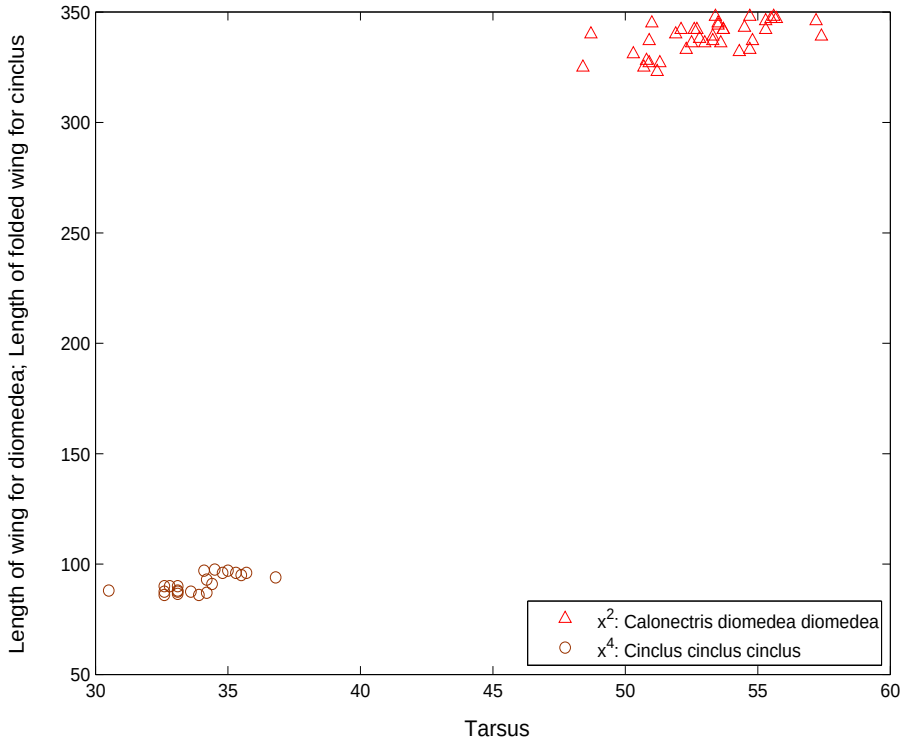
Simultaneous clustering relies, among other things, on the assumption that samples to be classified are described by variables of identical meaning. However in many concrete situations descriptors do not have exactly the same sense in some sample or other. The parsimonious models of simultaneous clustering 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) [32] 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 4.2) according to their sex and then the cluster number is set to 2. Table 4.10 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.

FIGURE 4.2: *Two bird samples described by variables close in meaning.*



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%).

4.1.6 Concluding remarks

This work is a scope enlargement of clustering based on Gaussian mixtures. It displays models (exhibited at first in [79], in a more restricted version) 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 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

TABLE 4.10: *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**)

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 4.1.5.2 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 [107]–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 [85], mixtures of Student distributions are applied when the data include outliers [85]. Some combined use of both factor analyzers and t-distributions seems to give interesting results in microarray gene-expression data clustering [87]. 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.

4.2 Application aux séries chronologiques

Il est fréquent que deux échantillons considérés à des instants différents, soient composés d'unités statistiques de même nature et soient décrits par des variables de même signification. Dans ce contexte les méthodes de classification traditionnelles, même les méthodes basées sur des mélanges, considèrent généralement que les échantillons proviennent de la même population, ou bien qu'ils proviennnent de populations différentes sans lien entre elles. Nous allons montrer sur deux exemples, l'un géologique (Section 4.2.1), l'autre biologique (Section 4.2.2), l'intérêt des modèles de classification simultanée dans ce contexte de séries chronologiques.

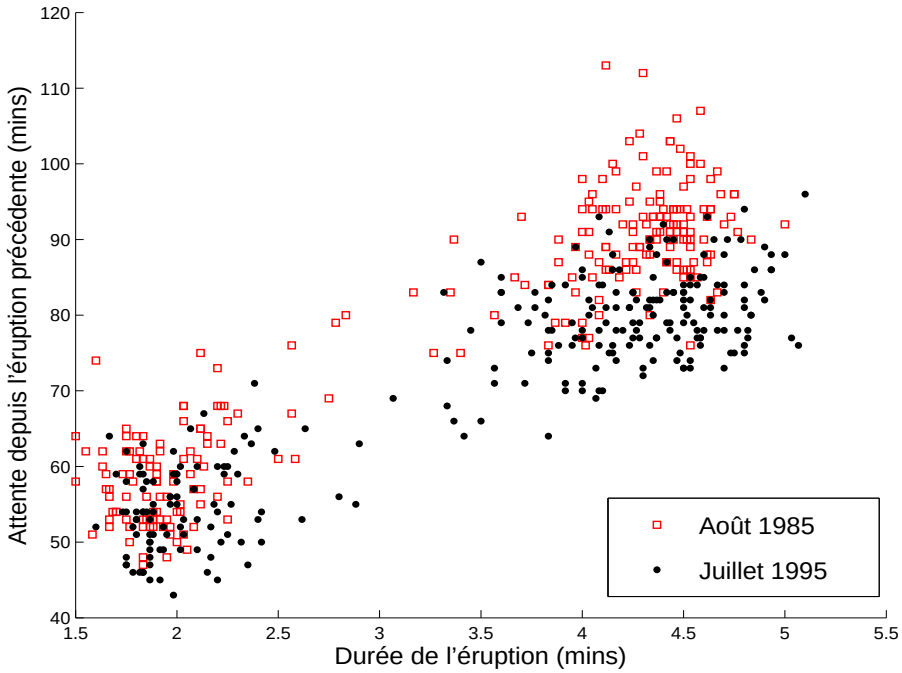


FIGURE 4.3: *Eruptions du geyser Old Faithful (Yellowstone National Park) à dix ans d'intervalle*

4.2.1 Evolution de la structure des éruptions d'OldFaithful

Les données de la section 2.2.2 qui ont permis de déterminer deux groupes parmi les éruptions d'Old Faithful, datent du mois d'Août 1985. Nous rappelons qu'il s'agissait de $n^1 = 272$ éruptions décrites chacune par sa durée (en mins) et par l'intervalle de temps (en mins) qui la sépare de l'éruption précédente. On trouvera sur le site `http://wwww.whfreeman.com/statistics/ips/EESEE1/OFDATA.TXT` un échantillon similaire de $n^2 = 252$ éruptions du geyser Old Faithful, décrites par les mêmes variables Durée et Attente. Ces nouvelles données ont été enregistrées par Rick A. Hutchinson, géologue au Parc du Yellowstone, dix ans plus tard, au cours du mois de Juillet 1995. La figure 4.3 représente la durée des éruptions et l'attente entre deux éruptions au cours des deux périodes.

Si les éruptions du geyser ont évolué dans leur distribution d'après la figure 4.3, elles semblent cependant avoir conservé, d'une période à l'autre, leur structure en deux groupes. Nous proposons de montrer l'intérêt de la classification simultanée pour retrouver dans chaque échantillon ces deux groupes d'éruptions interprétés à la section 2.2.2 comme les éruptions courtes ou longues.

Classification simultanée vs. indépendante des éruptions d'Old Faithful

Le tableau TAB. 4.11 donne les valeurs d'ICL obtenues par quelques uns des modèles de classification simultanée et par les modèles de classification indépendante.

			π		π_k	
			$\boldsymbol{\Sigma}$	$\boldsymbol{\Sigma}_k$	$\boldsymbol{\Sigma}$	$\boldsymbol{\Sigma}_k$
π	$\boldsymbol{I}$	$\mathbf{0}$	2346.5	2339.6	2335.3	2327.1
	$\boldsymbol{D}^{h,h'}$	$\boldsymbol{b}^{h,h'}$	2239.4	2236.6	2228.2	**2224.7**
		$\boldsymbol{b}_k^{h,h'}$	2238.1	2237.3	2226.9	2225.7
	$\boldsymbol{D}_k^{h,h'}$	$\boldsymbol{b}_k^{h,h'}$	.	2241.5	.	2230.2
π^h	$\boldsymbol{D}^{h,h'}$	$\boldsymbol{b}^{h,h'}$	.	.	2230.7	2227.2
		$\boldsymbol{b}_k^{h,h'}$	.	.	2229.4	2228.0
	$\boldsymbol{D}_k^{h,h'}$	$\boldsymbol{b}_k^{h,h'}$	.	.	.	2232.5
Independent			2240.9	2246.9	**2232.1**	2237.6

TABLE 4.11: *ICL des modèles gaussiens simultanés et indépendants ($K = 2$ groupes), estimés sur les données d'Old Faithful.*

Le modèle interpopulation de contrainte maximale, $(\pi, \boldsymbol{I}, \mathbf{0};\ .\ ,\ .\ )$, est systématiquement déclassé par ICL. Supposer que les deux échantillons d'éruptions proviennent de la même population conduit aux valeurs d'ICL les pires, et cela quelles que soient les contraintes envisagées sur les composantes (homoscédasticité ou hétéroscédasticité, homogénéité ou hétérogénéité de leur poids). Ceci confirme ce que laissait supposer FIG. 4.3 : les deux échantillons d'éruptions proviennent de populations distinctes et la distribution des éruptions d'Old Faithful a bien évolué de 1985 à 1995.
Mais l'hypothèse opposée selon laquelle il n'existe aucun lien entre les deux populations d'éruptions (qui correspond aux modèles de classification indépendante), ne recueille pas non plus la faveur d'ICL.
La meilleure valeur d'ICL est obtenue par le modèle $\left(\pi, \boldsymbol{D}^{h,h'}, \boldsymbol{b}^{h,h'}; \pi_k, \boldsymbol{\Sigma}_k\right)$ qui suppose un état intermédiaire du lien entre les populations d'éruptions des deux périodes.

Remarque. La structure en deux groupes des éruptions d'Old Faithful est une hypothèse largement répandue. Cependant, nous avons vu à la section 2.2.2 qu'elle ne fait pas l'unanimité. Le tableau 4.12 montre que, quel que soit le nombre de groupes considéré entre 1 et 5, l'hypothèse selon laquelle (i) les deux échantillons d'éruptions proviennnent de populations différentes et (ii) les deux populations d'éruptions ne sont pas sans lien, cette hypothèse est systématiquement retenue par ICL. La ligne 'simultané, pop. égales' indique pour chaque valeur de K considérée, la meilleure valeur d'ICL obtenue en considérant que les deux échantillons d'éruptions proviennent de la même population (modèles correspondant à la ligne 3 de TAB. 4.11). La ligne 'simultané, pop. distinctes' indique la meilleure valeur d'ICL obtenue en considérant que les deux populations sont distinctes mais liées de façon stochastique (modèles correspondant aux lignes 4 à 9 de TAB. 4.11). La ligne 'pop. indépendantes' indique la meilleure valeur d'ICL obtenue en considérant que les populations d'éruptions sont sans lien (modèles correspondant à la ligne 10 de TAB. 4.11).

K	1	2	3	4	5
simultané, pop. égales	2645.9	2327.1	2335.8	2363.5	2411.9
simultané, pop. distinctes	**2543.2**	**2224.7**	**2224.6**	**2236.9**	**2253.0**
pop. indépendantes	2544.7	2232.1	2239.5	2255.4	2277.4

TABLE 4.12: *Les meilleures valeurs d'ICL obtenues par les modèles de* TAB. *4.11 sur les données d'Old Faithful, pour un nombre variable de groupes*

4.2.2 Evolution de canards à foie gras

Les modèles du tableau 4.11 donnent également des résultats encourageants en faveur de la classification simultanée, lorsqu'ils sont inférés sur les deux échantillons de canards à foie gras (de 2006 et 2007) présentés à la section 3. Rappelons brièvement la problématique du foie gras de canard et la raison pour laquelle on s'intéresse à la proportion de lipides et de protéines des foies.

La problématique des canards à foie gras

Des canards gavés ne sont pas tous aptes à produire un foie gras. En effet pour qu'un foie puisse être commercialisé, il faut qu'une fois cuit, il ne soit pas trop maigre. Or il existe deux types de causes à la maigreur excessive d'un foie cuit, qui tiennent à la proportion de lipides et de protéines avant cuisson. Si la proportion de lipides est trop faible avant cuisson, le foie cuit sera évidemment trop maigre. Mais lorsque la proportion de lipides est au contraire trop importante, le taux de fonte est élevé, et le foie cuit devient également trop maigre. Ainsi dans un contexte industriel, les foies de canards se répartissent en deux groupes : les foies aptes ou inaptes à la commercialisation. Mais on distingue à nouveau deux groupes parmi les foies inaptes, suivant qu'ils sont trop maigres ou trop gras avant cuisson.

Sans prétendre apporter une solution à la détermination des foies aptes ou inaptes, nous juxtaposons à la problématique industrielle les faits suivants. (i) D'après FIG. 3.2 les populations de canards sont distinctes, mais elles ne semblent pas sans lien. (ii) La modélisation des échantillons et le choix d'un modèle confirment l'hypothèse d'un lien entre les populations de canards. Aux modèles de mélanges traditionnels, le critère BIC préfère en effet les modèles de classification simultanée (voir TAB. 4.13). (iii) Le critère BIC hésite à déterminer deux ou trois groupes dans chaque échantillon de canards (TAB. 4.13). (iv) Pour chaque échantillon de canards, la partition en deux groupes et la partition en trois groupes (issues de la classification simultanée) sont emboîtées (FIG. 4.4(d) et FIG. 4.4(b) ; TAB. 4.15).

Classification simultanée vs. indépendante des canards de 2006 et 2007

Chacun des modèles de TAB. 4.11 est inféré sur les deux échantillons de canards ($n^1 = 741$ canards en 2006 et $n^2 = 736$ canards en 2007) pour un nombre de groupes

variant de 1 à 5. Les valeurs de BIC obtenues sont réparties, comme dans TAB. 4.12 pour ICL, en trois catégories, selon la nature du lien établi entre les populations de canards. Le tableau 4.13 présente la meilleure valeur de BIC obtenue dans chaque catégorie.

K	1	2	3	4	5
simultané, pop. égales	5807.0	5738.8	5666.4	5680.0	5676.8
simultané, pop. distinctes	5148.5	**5058.4**	**5057.9**	**5062.1**	**5073.4**
pop. indépendantes	**5141.7**	5068.9	5082.6	5098.3	5102.3

TABLE 4.13: *Les meilleures valeurs de BIC obtenues par les modèles de* TAB. *4.11 sur les données de canards, pour un nombre variable de groupes*

On observe (une nouvelle fois) que BIC rejette (sauf pour $K = 1$ groupe) les deux hypothèses antagonistes : (i) les deux populations sont identiques, (ii) les deux populations sont distinctes et il n'existe aucun lien entre elles. En effet, c'est l'hypothèse intermédiaire d'un lien entre les populations de canards, qui emporte la faveur de BIC.

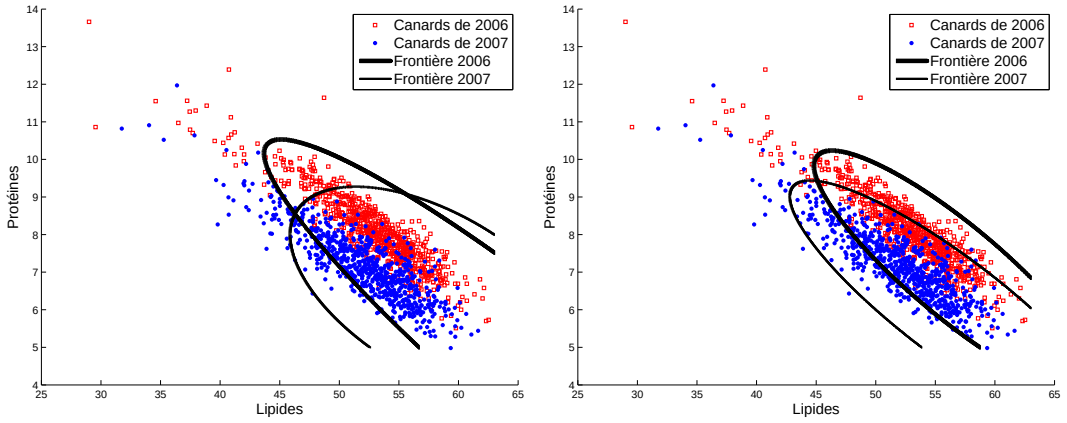
On observe d'autre part que BIC semble hésiter entre deux et trois groupes (TAB. 4.13). Nous allons voir que les deux modèles de classification simultanée retenus pour $K = 2$ et $K = 3$ groupes ($BIC = 5058.4$ et $BIC = 5057.9$) conduisent à une partition plus cohérente que celle déduite des modèles de classification indépendante ($BIC = 5068.9$ et $BIC = 5082.6$).

La figure 4.4 représente la frontière de classement des échantillons de canards, inférée pour $K = 2$ ou $K = 3$ groupes, de façon indépendante ou simultanée.

D'après FIG. 4.4(b) et FIG. 4.4(d), pour chaque échantillon, les partitions obtenues en classification simultanée semblent emboîtées. A l'évidence, si l'on compare FIG. 4.4(a) et FIG. 4.4(c), les partitions que l'on obtient en classification indépendante ne le sont pas.

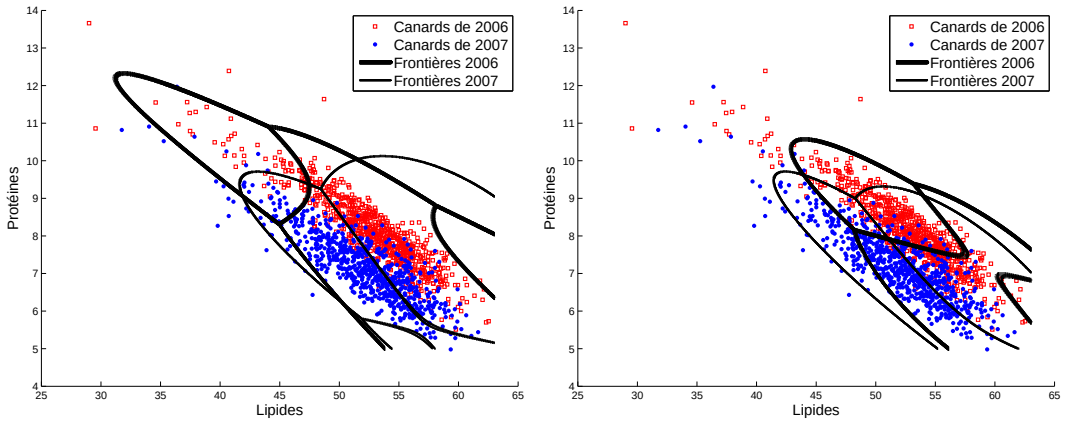
Le quasi-emboîtement des partitions estimées de façon simultanée, est confirmé par leur tableau de confusion. A chaque échantillon (canards de 2006 ou 2007), et à chaque nombre de groupes considéré ($K = 2$ ou $K = 3$), correspond une partition. On note A_i (resp. $\mathrm{A'}_i$) le groupe i ($i = 1, 2, 3$) dans la partition à trois groupes des canards de 2006 (resp. de 2007) et B_j (resp. $\mathrm{B'}_j$) le groupe j ($j = 1, 2$) dans la partition à deux groupes des canards de la même année. Les tableaux TAB. 4.14 et TAB. 4.15 indiquent les effectifs des canards de 2006 (resp. de 2007) dans chacun des groupes du type $\mathrm{A}_i \cap \mathrm{B}_j$ (resp. $\mathrm{A'}_i \cap \mathrm{B'}_j$) lorsque les partitions sont estimées de façon indépendante (TAB. 4.14) ou simultanée (TAB. 4.15). Ainsi, les groupes A_i (resp. B_j) de TAB. 4.15 correpondent à la partition des canards de 2006 représentée par FIG. 4.4(d) (resp. par FIG. 4.4(b)). Le tableau 4.14 nous apprend par exemple, que les canards de 2006 sont pour quarante et un d'entre eux, à la fois dans le groupe 2 de la partition à trois groupes, et dans le groupe 1 de la partition à deux groupes, lorsque les partitions sont inférées de façon indépendante.

On remarque que la distribution des effectifs des deux échantillons est proche lorsque les partitions sont estimées de façon simultanée (voir TAB. 4.15). Ce n'est pas le cas lorsque les partitions sont estimées de façon indépendante (voir TAB. 4.14).



(a) *Classif. indépendante et* $K = 2$ *groupes* $(BIC = 5068.9)$

(b) *Classif. simultanée et* $K = 2$ *groupes* $(BIC = 5058.4)$



(c) *Classif. indépendante et* $K = 3$ *groupes* $(BIC = 5082.6)$

(d) *Classif. simultanée et* $K = 3$ *groupes* $(BIC = \mathbf{5057.9})$

FIGURE 4.4: *Comparaison des frontières de classement des deux échantillons de canards, obtenues en classification simultanée et indépendante pour deux groupes ou trois groupes*

		$K = 3$		
	groupe	$A_1 < A'_1 >$	$A_2 < A'_2 >$	$A_3 < A'_3 >$
$K = 2$	$B_1 < B'_1 >$	$672 < 44 >$	$41 < 19 >$	$0 < 0 >$
	$B_2 < B'_2 >$	$0 < 447 >$	$24 < 1 >$	$4 < 225 >$

TABLE 4.14: *Classification indépendante (Comparaison dans chaque échantillon de canards, de la partition à deux groupes et à trois groupes)*

D'autre part TAB. 4.15 confirme l'emboîtement des partitions à deux et à trois groupes lorsqu'elles sont estimées de façon simultanée (Nous avions jusqu'à maintenant constaté l'emboîtement graphiquement, en comparant FIG. 4.4(b) et FIG. 4.4(d)). En effet le groupe B_1 (resp. B'_1) semble contenir A_1 et A_2 (resp. A'_1 et A'_2) et le groupe B_2 (resp. B'_2) correspond à peu de choses près au groupe A_3 (resp. A'_3). D'après TAB. 4.14 les partitions estimées de façon indépendante ne possèdent pas cette proprité d'emboî-

		$K=3$		
	groupe	$A_1 < A'_1 >$	$A_2 < A'_2 >$	$A_3 < A'_3 >$
$K=2$	$B_1 < B'_1 >$	$230 < 218 >$	$475 < 486 >$	$1 < 4 >$
	$B_2 < B'_2 >$	$0 < 4 >$	$8 < 5 >$	$27 < 19 >$

TABLE 4.15: *Classification simultanée (Comparaison dans chaque échantillon de canards, de la partition à deux groupes et à trois groupes)*

tement.

4.3 Mélanges de Student (Classification simultanée robuste)

Classifier grâce à un modèle de mélange gaussien peut être un mauvais choix, lorsque les données comportent du bruit ou des outliers par exemple. En effet les données atypiques influencent beaucoup l'estimation des paramètres gaussiens et perturbent fortement la règle de classement. Supposer que les populations conditionnelles ont été générées par des lois de Student - dont la queue est plus longue que celle de lois normales - constitue alors une alternative. Les valeurs obtenues des probabilités conditionnelles sont plus éloignées de 0 et de 1 que dans le cas gaussien (voir FIG. 4.5), et le poids des données atypiques dans l'estimation du paramètre est amoindri.

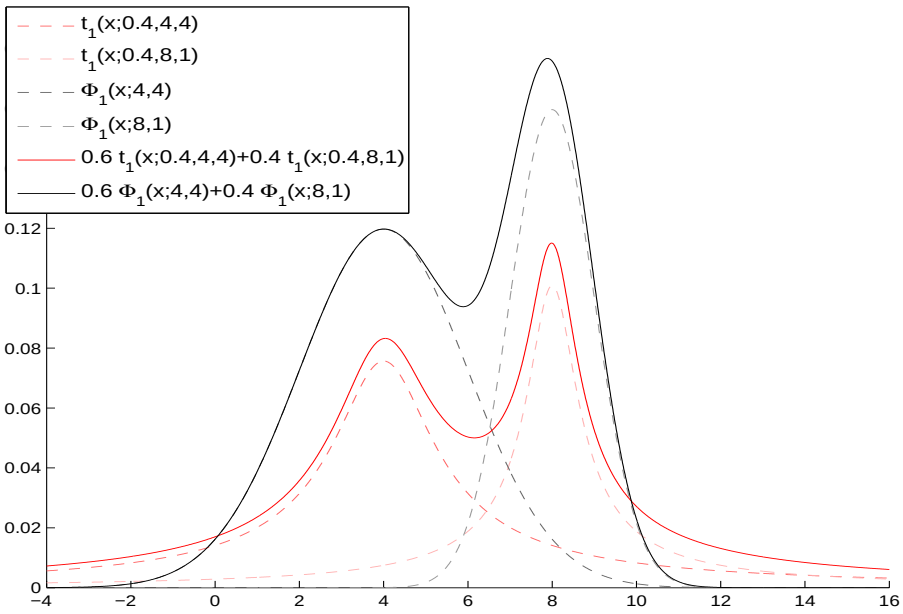


FIGURE 4.5: *Comparaison d'un mélange de lois normales et d'un mélange de lois de Student.*

Hoogerheide et al. ([61]) passent en revue les raisons de l'intérêt récent que suscitent les mélanges de Student. Bishop et Svensén ([20]) montrent que l'usage de mélanges de

Student est pertinent pour détecter un bruit introduit dans les données réelles de [91]. McLachlan et Peel ([85]) montrent que ce type de mélange convient mieux qu'un mélange gaussien pour déterminer le sexe des crabes de [23].

4.3.1 Généralités sur la loi de Student multivariée

Une extension de la loi normale

Un vecteur aléatoire $\boldsymbol{X}$ de $\mathbb{R}^d$ suit une loi de Student de paramètre $(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ s'il admet pour fonction de densité :

$$t_d(\boldsymbol{x}; \nu, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\Gamma(\frac{\nu+d}{2})(\pi\nu)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}}{\Gamma(\nu/2)[1 + (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/\nu]^{(\nu+d)/2}}, \tag{4.13}$$

où $\nu \in \mathbb{R}_*^+$ désigne le degré de liberté, $\boldsymbol{\mu} \in \mathbb{R}^d$ un paramètre de localisation, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ une matrice de produit scalaire ($\boldsymbol{\Sigma}$ est symétrique, définie, positive), et Γ la fonction Gamma d'Euler.
$\boldsymbol{X}$ n'a de moyenne (resp. de matrice des covariances) que si $\nu > 1$ (resp. $\nu > 2$) et dans ce cas elle coïncide avec $\boldsymbol{\mu}$ (resp. avec $\frac{\nu}{\nu-2}\boldsymbol{\Sigma}$). Le degré de liberté ν contrôle la longueur de queue de la distribution : plus ν est petit plus la queue de la distribution de $\boldsymbol{X}$ est longue, et lorsque ν tend vers $+\infty$ la loi de $\boldsymbol{X}$ tend en probabilité vers la loi normale de paramètre $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (voir FIG. 4.6)

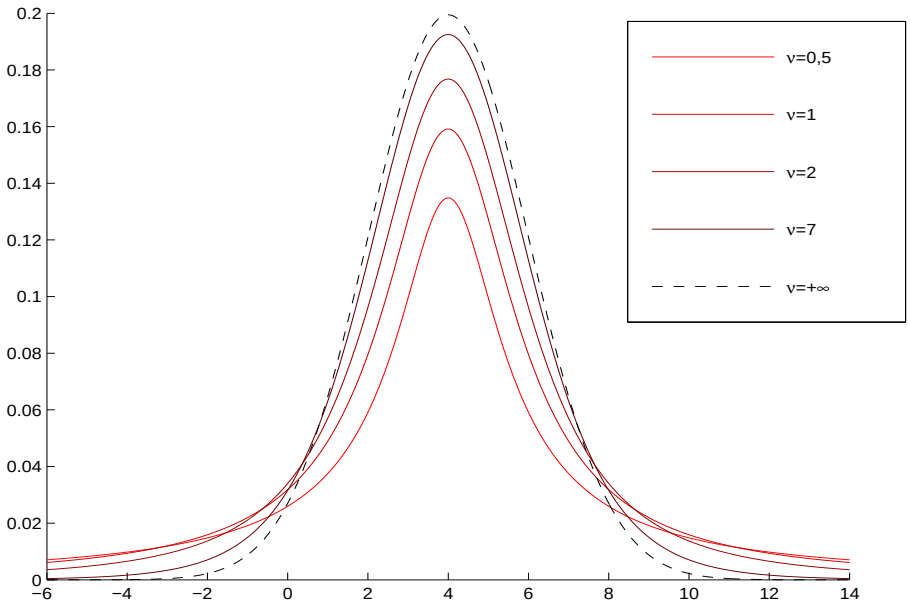


FIGURE 4.6: *Densités de probabilité de* $t_1(\boldsymbol{x}; \nu, 4, 4)$ *pour différentes valeurs de* ν*. On observe que* $t_1(\boldsymbol{x}; +\infty, 4, 4) = \Phi_1(\boldsymbol{x}; 4, 4)$.

Estimation par maximum de vraisemblance

La donnée d'un échantillon $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de $\mathbb{R}^d$ ne suffit pas à elle seule, pour estimer le paramètre $(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ d'une loi de Student par maximum de vraisemblance. Les estimateurs ne sont pas explicites, pas même lorsqu'on suppose ν connu. Il est possible cependant d'estimer $(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ grâce à un algorithme EM, après avoir mis en évidence la structure de donnée manquante sous-jacente à la loi de Student.

Théorème 1. *Un vecteur aléatoire $\boldsymbol{X}$ de $\mathbb{R}^d$ suit une loi de Student de paramètre $(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ si et seulement s'il existe une variable aléatoire U à valeurs dans $\mathbb{R}_*^+$ distribuée selon la loi $\gamma_{\nu/2,\nu/2}$ et telle que $(\boldsymbol{X}|U=u)$ suit la loi normale $\mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}/u)$.*

Preuve. 1. ($\Leftarrow$) La densité du couple $(\boldsymbol{X}, U)$ est :

$$g(\boldsymbol{x}, u) = C. \exp\left[-\frac{\nu + (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}.u + \frac{d+\nu-2}{2}.\ln u\right] \mathbb{1}_{\mathbb{R}_*^+}(u),$$

avec $(\boldsymbol{x}, u) \in \mathbb{R}^d \times \mathbb{R}_*^+$ et $C = \dfrac{|\boldsymbol{\Sigma}|^{-1/2}(\nu/2)^{\nu/2}}{(2\pi)^{d/2}\Gamma(\nu/2)}$.

On en déduit la densité de $\boldsymbol{X}$:

$$\int_{\mathbb{R}} g(\boldsymbol{x}, u)\mathrm{d}u = t_d(\boldsymbol{x}; \nu, \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

en changeant de variable d'intégration : $s = \dfrac{\nu + (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}.u.$

2. ($\Rightarrow$) U est une variable aléatoire réelle qui, conditionnellement à la valeur $\boldsymbol{x}$ de $\boldsymbol{X}$, suit la loi gamma de premier paramètre $(\nu+d)/2$ et de second paramètre $(\nu + (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))/2$.

   La densité du couple $(\boldsymbol{X}, U)$ est donc $g(\boldsymbol{x}, u)$.

   La densité marginale relative à la variable U est :

$$\int_{\mathbb{R}^d} g(\boldsymbol{x}, u)\mathrm{d}\boldsymbol{x}. \tag{4.14}$$

   Cette densité est celle de la loi $\gamma_{\nu/2,\nu/2}$.

□

Ainsi la génèse d'une donnée $\mathbf{x} \in \mathbb{R}^d$ selon une loi de Student de paramètre $(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ peut se décomposer en deux étapes :

- on détermine d'abord une donnée $u \in \mathbb{R}_*^+$ selon la loi $\gamma_{\nu/2,\nu/2}$ (destinée à modifier le volume de $\mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$) ;
- on génère ensuite $\mathbf{x} \in \mathbb{R}^d$ selon la loi normale $\mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}/u)$.

Supposons qu'un échantillon $\mathbf{x}_i$ $(i = 1, \ldots, n)$ provient d'une loi de Student de paramètre $(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. On peut considérer (i) que chaque $\mathbf{x}_i$ a été généré par une loi normale centrée en $\boldsymbol{\mu}$, dont la matrice des covariances ne diffère de $\boldsymbol{\Sigma}$ que par le volume : on note $\boldsymbol{\Sigma}/u_i$ $(u_i \in \mathbb{R}_*^+)$ cette matrice, et (ii) que les données u_i $(i = 1, \ldots, n)$ sont générées par la loi $\gamma_{\nu/2,\nu/2}$.
Ainsi les points $\mathbf{x}_i$ constituent la donnée observée, les réels u_i la donnée manquante, et les couples $(\mathbf{x}_i, u_i)$ la donnée complète permettant d'estimer $(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ par maximum de vraisemblance dans un algorithme EM.

4.3.2 Classification d'échantillons d'origines multiples

On dispose de H échantillons $\mathbf{x}^h$ $(h = 1, \ldots, H)$ de $\mathbb{R}^d$ à partitionner en K groupes. Un échantillon $\mathbf{x}^h$ est constitué de n^h individus $\mathbf{x}_i^h$ $(i = 1, \ldots, n^h)$ et on suppose qu'il provient d'une population P^h. Le contexte est le même que dans le cas gaussien : les populations sont décrites par les mêmes d variables continues et on cherche le même nombre de groupes dans chacun des échantillons.

4.3.2.1 Mélanges de Student et classification indépendante

On considère ici que les échantillons présentent des outliers, du bruit, des valeurs extrêmes ou tout ce qui pourrait donner à penser qu'un mélange de Student conviendrait mieux à la classification de chacun des échantillons, qu'un mélange gaussien.

On suppose que chaque échantillon $\mathbf{x}^h$ provient d'un mélange de Student à K composantes dont la densité est :

$$f(\boldsymbol{x}; \boldsymbol{\psi}^h) = \sum_{k=1}^{K} \pi_k^h t_d(\boldsymbol{x}; \nu_k^h, \boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h) \ ; \ \boldsymbol{x} \in \mathbb{R}^d. \tag{4.15}$$

Les coefficients π_k^h $(k = 1, \ldots, K)$ sont les proportions du mélange (pour tout k : $\pi_k^h > 0$, et $\sum_{k=1}^K \pi_k^h = 1$) ; $\nu_k^h (\in \mathbb{R}_*^+)$, $\boldsymbol{\mu}_k^h$ et $\boldsymbol{\Sigma}_k^h$ désignent respectivement le degré de liberté, le paramètre de localisation et la matrice des produits scalaires de la composante k. Le paramètre du mélange P^h est $\boldsymbol{\psi}^h = \left(\boldsymbol{\pi}^h, \nu_1^h, \ldots, \nu_K^h, \boldsymbol{\omega}_1^h, \ldots, \boldsymbol{\omega}_K^h\right)$ avec $\boldsymbol{\pi}^h = \left(\pi_1^h, \ldots, \pi_K^h\right)$ et $\boldsymbol{\omega}_k^h = \left(\boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h\right)$.

Comme dans le cas gaussien, la composante d'origine d'un point $\mathbf{x}_i^h$ est indiquée par un vecteur binaire $\boldsymbol{z}_i^h \in \{0, 1\}^K$ dont la k^{e} composante $z_{i,k}^h$ vaut 1 (et les autres 0) si et seulement si $\mathbf{x}_i^h$ provient de la composante k.

Estimer le paramètre $\boldsymbol{\psi} = \left(\boldsymbol{\psi}^h\right)_{h=1,\dots,H}$ en maximisant sa log-vraisemblance :

$$\ell\left(\boldsymbol{\psi}; \mathbf{x}\right) = \sum_{h=1}^{H} \sum_{i=1}^{n^h} \ln f(\mathbf{x}_i^h; \boldsymbol{\psi}^h) \tag{4.16}$$

calculée sur la donnée observée $\mathbf{x} = \bigcup_{h=1}^{H} \mathbf{x}^h$, revient à maximiser indépendamment la log-vraisemblance $\ell^h\left(\boldsymbol{\psi}^h; \mathbf{x}^h\right)$ de chaque paramètre $\boldsymbol{\psi}^h$ calculée sur l'échantillon $\mathbf{x}^h$. Mais contrairement au cas gaussien, on ne peut pas estimer le paramètre $\boldsymbol{\psi}^h$ du mélange h grâce à EM si l'on considère que la donnée manquante relative à l'échantillon h ne consiste qu'en la partition inconnue $\boldsymbol{z}^h = \{\boldsymbol{z}_i^h; i = 1, \dots, n^h\}$.

D'après la caractérisation précédente d'une loi de Student multivariée (Théorème 1) si le point $\mathbf{x}_i^h$ a été généré par la composante k du mélange P^h, on peut le voir comme provenant d'une loi normale centrée en $\boldsymbol{\mu}_k^h$ dont la matrice des covariances - que l'on note pour cette raison $\boldsymbol{\Sigma}_k^h/u_i^h$ ($u_i^h \in \mathbb{R}_*^+$) - ne diffère de $\boldsymbol{\Sigma}_k^h$ que par le volume. Ainsi pour l'observation i dans l'échantillon h, la donnée manquante u_i^h complète celle de $\boldsymbol{z}_i^h$.

La donnée complète relative à l'échantillon h, notée $\mathbf{x}_{\mathrm{c}}^h$, est constituée des triplets $\left(\mathbf{x}_i^h, \boldsymbol{z}_i^h, u_i^h\right)$ ($i = 1, \dots, n^h$), les points $\mathbf{x}_i^h$ ($i = 1, \dots, n^h$) à classifier formant la donnée observée relative à cet échantillon, et les couples $(\boldsymbol{z}_i^h, u_i^h)$ ($i = 1, \dots, n^h$), la donnée manquante que l'on note $\mathbf{x}_{\mathrm{m}}^h$. On suppose que les triplets $\left(\mathbf{x}_i^h, \boldsymbol{z}_i^h, u_i^h\right)$ sont des réalisations de vecteurs aléatoires indépendants $\left(\boldsymbol{X}_i^h, \boldsymbol{Z}_i^h, U_i^h\right)$ identiquement distribués à $\left(\boldsymbol{X}^h, \boldsymbol{Z}^h, U^h\right) \in \mathbb{R}^d \times \{0,1\}^K \times \mathbb{R}_*^+$ avec :

$$\boldsymbol{Z}^h \sim \mathscr{M}_K(1; \pi_1^h, \dots, \pi_K^h), \tag{4.17}$$

$$\left(U^h | Z_k^h = 1\right) \sim \gamma_{\nu_k^h/2, \nu_k^h/2} \tag{4.18}$$

et

$$\left(\boldsymbol{X}^h | U^h = u, Z_k^h = 1\right) \sim \mathscr{N}_d(\boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h/u). \tag{4.19}$$

La log-vraisemblance du paramètre $\boldsymbol{\psi}^h$ calculée sur la donnée complète $\mathbf{x}_{\mathrm{c}}^h = \left\{(\mathbf{x}_i^h, \boldsymbol{z}_i^h, u_i^h); i = 1, \dots, n^h\right\}$ relative à l'échantillon h est alors :

$$l\left(\boldsymbol{\psi}^h; \mathbf{x}_{\mathrm{c}}^h\right) = l_1\left(\boldsymbol{\pi}^h; \boldsymbol{z}^h\right) + \sum_{k=1}^{K} l_2\left(\nu_k^h; \mathbf{x}_{\mathrm{m}}^h\right) + \sum_{k=1}^{K} l_3\left(\boldsymbol{\omega}_k^h; \mathbf{x}_{\mathrm{c}}^h\right) \tag{4.20}$$

avec :

$$l_1\left(\boldsymbol{\pi}^h; \boldsymbol{z}^h\right) = \sum_{k=1}^{K} \sum_{i=1}^{n^h} z_{i,k}^h \ln \pi_k^h \ , \tag{4.21}$$

$$l_2\left(\nu_k^h;\mathbf{x}_{\mathrm{m}}^h\right)=\sum_{i=1}^{n^h} z_{i,k}^h\{-\ln\Gamma\left(\frac{\nu_k^h}{2}\right)+\frac{\nu_k^h}{2}\ln\left(\frac{\nu_k^h}{2}\right)+\left(\frac{d+\nu_k^h}{2}-1\right)\ln u_i^h-\frac{\nu_k^h}{2}u_i^h\},\quad(4.22)$$

et :

$$l_3\left(\boldsymbol{\omega}_k^h;\mathbf{x}_{\mathrm{c}}^h\right)=\sum_{i=1}^{n^h} z_{i,k}^h\{-\frac{d}{2}\ln(2\pi)-\frac{1}{2}\ln|\boldsymbol{\Sigma}_k^h|-\frac{1}{2}u_i^h\left(\mathbf{x}_i^h-\boldsymbol{\mu}_k^h\right)'\left(\boldsymbol{\Sigma}_k^h\right)^{-1}\left(\mathbf{x}_i^h-\boldsymbol{\mu}_k^h\right)\}.\quad(4.23)$$

La décomposition (4.20) montre que si les données manquantes sont connues (ou estimées) dans l'échantillon h, toute la donnée relative à cet échantillon n'est pas utile pour estimer chacune des composantes de $\boldsymbol{\psi}^h$. La partition inconnue $\boldsymbol{z}^h=\{\boldsymbol{z}_i^h;i=1,\ldots,n^h\}$ suffit par exemple à estimer les proportions π_k^h $(k=1,\ldots,K)$. Les degrés de liberté ν_k^h $(k=1,\ldots,K)$ ne dépendent que des données manquantes $\mathbf{x}_{\mathrm{m}}^h=\{(\boldsymbol{z}_i^h,u_i^h);i=1,\ldots,n^h\}$. Les paramètres $\boldsymbol{\mu}_k^h$ et $\boldsymbol{\Sigma}_k^h$ $(k=1,\ldots,K)$ nécessitent par contre toute la donnée relative à l'échantillon h pour être estimés.

Chaque algorithme EM permettant d'optimiser la vraisemblance d'un paramètre $\boldsymbol{\psi}^h$ consiste alternativement à estimer sa vraisemblance complétée $l\left(\boldsymbol{\psi}^h;\mathbf{x}^h\right)$ conditionnellement à la donnée observée $\mathbf{x}^h$ (étape E), puis à déterminer la valeur de $\boldsymbol{\psi}^h$ qui rend maximale cette estimation (étape M).

Etape E : En notant $\tilde{\boldsymbol{\psi}}^h$ la valeur courante du paramètre $\boldsymbol{\psi}^h$ on estime $l\left(\boldsymbol{\psi}^h;\mathbf{x}^h\right)$ par sa moyenne sur les valeurs de la donnée manquante $\mathbf{x}_{\mathrm{m}}^h=\{(\boldsymbol{z}_i^h,u_i^h);i=1,\ldots,n^h\}$ considérée comme aléatoire et paramétrée par $\tilde{\boldsymbol{\psi}}^h$, conditionnellement à la donnée observée $\mathbf{x}^h=\{\mathbf{x}_i^h;1,\ldots,n^h\}$:

$$Q(\boldsymbol{\psi}^h;\tilde{\boldsymbol{\psi}}^h)=E_{\left\{\boldsymbol{Z}_i^h,U_i^h\right\}}\left(l\left(\boldsymbol{\psi}^h;\{\boldsymbol{X}_i^h,\boldsymbol{Z}_i^h,U_i^h\}\right)|\boldsymbol{X}_i^h=\mathbf{x}_i^h;\tilde{\boldsymbol{\psi}}^h\right)\quad(4.24)$$

La formule de l'espérance totale permet de calculer cette espérance par rapport à la loi des variables U_i^h conditionnellement aux vecteurs $\boldsymbol{Z}_i^h$ d'abord, puis par rapport à la loi des vecteurs $\boldsymbol{Z}_i^h$:

$$Q^h(\boldsymbol{\psi}^h;\tilde{\boldsymbol{\psi}}^h)=E_{\left\{\boldsymbol{Z}_i^h\right\}}\left[E_{\left\{U_i^h\right\}}\left(l\left(\boldsymbol{\psi}^h;\{\boldsymbol{X}_i^h,\boldsymbol{Z}_i^h,U_i^h\}\right)|\boldsymbol{X}_i^h=\mathbf{x}_i^h;\tilde{\boldsymbol{\psi}}^h\right);\tilde{\boldsymbol{\psi}}^h\right]\quad(4.25)$$

Daprès le théorème 1, $\left(U_i^h|\boldsymbol{X}_i^h=\mathbf{x}_i^h,Z_{i,k}^h=1\right)$ suit une loi γ dont le premier paramètre est $(\tilde{\nu}_k^h+d)/2$ et le second $\left(\tilde{\nu}_k^h+\left(\mathbf{x}_i^h-\tilde{\boldsymbol{\mu}}_k^h\right)'\left(\tilde{\boldsymbol{\Sigma}}_k^h\right)^{-1}\left(\mathbf{x}_i^h-\tilde{\boldsymbol{\mu}}_k^h\right)\right)/2$. On en déduit que cette variable aléatoire à valeur dans $\mathbb{R}_*^+$ vaut en moyenne :

$$\tilde{u}_{i,k}^h=\frac{\tilde{\nu}_k^h+d}{\tilde{\nu}_k^h+\left(x_i^h-\tilde{\boldsymbol{\mu}}_k^h\right)'\left(\tilde{\boldsymbol{\Sigma}}_k^h\right)^{-1}\left(x_i^h-\tilde{\boldsymbol{\mu}}_k^h\right)},\quad(4.26)$$

et que son logarithme est une variable aléatoire d'espérance :

$$\ln \tilde{u}_{i,k}^h + \frac{\Gamma'((\tilde{\nu}_k^h + d)/2)}{\Gamma((\tilde{\nu}_k^h + d)/2)} - \ln((\tilde{\nu}_k^h + d)/2). \tag{4.27}$$

Par ailleurs la variable $(Z_{i,k}^h | \boldsymbol{X}_i^h = \mathbf{x}_i^h)$ vaut en moyenne :

$$t_{i,k}^h = \frac{\tilde{\pi}_k^h t_d(\mathbf{x}_i^h; \tilde{\nu}_k^h, \tilde{\boldsymbol{\mu}}_k^h, \tilde{\boldsymbol{\Sigma}}_k^h)}{\sum_{j=1}^K \tilde{\pi}_j^h t_d(\mathbf{x}_i^h; \tilde{\nu}_j^h, \tilde{\boldsymbol{\mu}}_j^h, \tilde{\boldsymbol{\Sigma}}_j^h)}. \tag{4.28}$$

On en déduit :

$$Q(\boldsymbol{\psi}^h; \tilde{\boldsymbol{\psi}}^h) = Q_1(\boldsymbol{\pi}^h; \tilde{\boldsymbol{\psi}}^h) + \sum_{k=1}^K Q_2(\nu_k^h; \tilde{\boldsymbol{\psi}}^h) + \sum_{k=1}^K Q_3(\boldsymbol{\omega}_k^h; \tilde{\boldsymbol{\psi}}^h) \ , \tag{4.29}$$

avec :

$$Q_1(\pi_k^h; \tilde{\boldsymbol{\psi}}^h) = \sum_{k=1}^K \sum_{i=1}^{n^h} t_{i,k}^h \ln \pi_k^h \ , \tag{4.30}$$

$$\begin{aligned} Q_2(\nu_k^h; \tilde{\boldsymbol{\psi}}^h) = \sum_{i=1}^{n^h} t_{i,k}^h \{ & - \ln \Gamma\left(\frac{\nu_k^h}{2}\right) + \frac{\nu_k^h}{2} \ln\left(\frac{\nu_k^h}{2}\right) - \frac{\nu_k^h}{2} \tilde{u}_i^h \\ & + \left(\frac{d + \nu_k^h}{2} - 1\right) \times \left(\ln \tilde{u}_{i,k}^h + \frac{\Gamma'((\tilde{\nu}_k^h + d)/2)}{\Gamma((\tilde{\nu}_k^h + d)/2)} - \ln((\tilde{\nu}_k^h + d)/2)\right)\} \end{aligned} \tag{4.31}$$

et :

$$\begin{aligned} Q_3(\boldsymbol{\omega}_k^h; \tilde{\boldsymbol{\psi}}^h) = \sum_{i=1}^{n^h} t_{i,k}^h \{ & - \frac{d}{2} \ln(2\pi) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_k^h| \\ & - \frac{1}{2} \tilde{u}_i^h \left(\mathbf{x}_i^h - \boldsymbol{\mu}_k^h\right)' \left(\boldsymbol{\Sigma}_k^h\right)^{-1} \left(\mathbf{x}_i^h - \boldsymbol{\mu}_k^h\right)\}. \end{aligned} \tag{4.32}$$

Etape M : La décomposition (4.29) montre que $Q(\boldsymbol{\psi}^h; \tilde{\boldsymbol{\psi}}^h)$ est additivement séparable. On peut l'optimiser séparément par rapport aux composantes $\boldsymbol{\pi}^h$, ν_k^h et $\boldsymbol{\omega}_k^h$ du paramètre $\boldsymbol{\psi}^h$.

Les proportions des composantes dans les mélanges sont estimées grâce aux probabilités conditionnelles uniquement, selon :

$$\pi_k^h = \left(\sum_{i=1}^{n^h} t_{i,k}^h\right) / n^h. \tag{4.33}$$

Ces estimateurs sont identiques à ceux des proportions dans n'importe quel modèle de mélange.

Les paramètres de localisation et les matrices de produits scalaires sont estimés par :

$$\boldsymbol{\mu}_k^h = \left(\sum_{i=1}^{n^h} t_{i,k}^h \tilde{u}_{i,k}^h \mathbf{x}_i^h \right) / \left(\sum_{i=1}^{n^h} t_{i,k}^h \tilde{u}_{i,k}^h \right), \tag{4.34}$$

et :

$$\boldsymbol{\Sigma}_k^h = \left(\sum_{i=1}^{n^h} t_{i,k}^h \tilde{u}_{i,k}^h (\mathbf{x}_i^h - \boldsymbol{\mu}_k^h)(\mathbf{x}_i^h - \boldsymbol{\mu}_k^h)' \right) / \left(\sum_{i=1}^{n^h} t_{i,k}^h \right). \tag{4.35}$$

Les estimateurs (4.34) et (4.35) ressemblent beaucoup aux estimateurs du maximum de vraisemblance du paramètre d'une loi normale. Cela tient à ce que chaque terme de (4.32) est la vraisemblance du paramètre d'une loi normale à la pondération près des points $\mathbf{x}_i^h$. La formule (4.34) montre en particulier que l'estimateur de $\boldsymbol{\mu}_k^h$ est la moyenne empirique de l'échantillon $\mathbf{x}^h$ dont chaque point $\mathbf{x}_i^h$ est pondéré de $t_{i,k}^h \tilde{u}_i^h$.

On peut fixer le degré de liberté des composantes afin d'accélerer EM (Il ne s'agit pas d'améliorer les qualités intrinsèques de l'algorithme mais de simplifier l'algorithme en soustrayant de l'estimation une composante du paramètre). Dans ce cas l'étape M se limite aux estimations précédentes qui sont explicites. Mais les degrés de liberté de chaque mélange constituent un paramètre de réglage de la robustesse du modèle puisqu'ils déterminent la longueur de queue de distribution des populations conditionnelles. Les inférer au même titre que les autres paramètres permet au modèle de faire dépendre sa robustesse de la donnée elle-même. Si l'on fait ce choix les coefficients ν_k^h doivent être estimés à chaque étape M comme solutions des équations de vraisemblance :

$$\partial Q_2(\nu_k^h; \tilde{\boldsymbol{\psi}}) / \partial \nu_k^h = 0 \; ; k = 1, \ldots, K. \tag{4.36}$$

On peut enfin envisager d'estimer les degrés de liberté en considérant que ce sont des entiers non nuls et en optimisant (4.31) de façon discrète. On évite ainsi de recourir à une méthode numérique pour résoudre (4.36) mais on augmente l'algorithme d'une difficulté combinatoire.

4.3.2.2 Mélanges de Student et classification simultanée

Comme dans le cas gaussien, puisque les échantillons $\mathbf{x}^h$ $(h = 1, \ldots, H)$ sont décrits par les mêmes d variables continues et que l'on recherche le même nombre de groupes dans chacun d'entre eux, on établit un lien stochastique entre populations conditionnelles. Pour tout $k \in \{1, \ldots, K\}$ et tout couple $(h, h') \in \{1, \ldots, H\}^2$ on suppose qu'il existe une application $\xi_k^{h,h'} : \mathbb{R}^d \to \mathbb{R}^d$ telle que :

$$(\boldsymbol{X}^{h'}|Z_k^{h'}=1) \sim \xi_k^{h,h'}\left(\boldsymbol{X}^{h}|Z_k^{h}=1\right). \tag{4.37}$$

Comme les échantillons sont décrits par des variables de même signification on suppose ici encore que la composante j du vecteur aléatoire $(\boldsymbol{X}^{h'}|Z_k^{h'}=1)$ ne dépend stochastiquement que de la composante j du vecteur $(\boldsymbol{X}^{h}|Z_k^{h}=1)$ c'est à dire que chaque application $\xi_k^{h,h'}$ vérifie la propriété :

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

Ainsi chaque composante $\left(\xi_k^{h,h'}\right)^{(j)}$ $(j \in \{1,\dots,d\})$ de l'application $\xi_k^{h,h'}$ peut s'identifier à une application de $\mathbb{R}$ dans $\mathbb{R}$ qui transforme stochastiquement la variable aléatoire $(\boldsymbol{X}^{h}|Z_k^{h}=1)^{(j)}$ - distribuée selon une loi de Student univariée à ν_k^h degrés de liberté - en la variable aléatoire $(\boldsymbol{X}^{h'}|Z_k^{h'}=1)^{(j)}$ qui suit elle aussi une loi de Student univariée, à $\nu_k^{h'}$ degré de liberté. Si on suppose de plus que chaque application $\left(\xi_k^{h,h'}\right)^{j}$ est continûment dérivable (on note $\mathscr{H}_2$ cette hypothèse lorsqu'elle est vérifiée pour tout $j \in \{1,\dots,d\}$), alors $\left(\xi_k^{h,h'}\right)^{j}$ est forcément affine. En effet il existe deux applications (et seulement deux) de $\mathbb{R}$ dans $\mathbb{R}$ continûment dérivables qui transforment stochastiquement une variable aléatoire de Student en une autre, et ces deux applications sont affines (voir Annexe B Conséquence du Théorème 1).

On en déduit que pour tout $k \in \{1,\dots,K\}$ et tout $(h,h') \in \{1,\dots,H\}^2$ il existe une matrice $\boldsymbol{D}_k^{h,h'} \in \mathbb{R}^{d\times d}$ diagonale, inversible et un vecteur $\boldsymbol{b}_k^{h,h'} \in \mathbb{R}^d$ tels que :

$$(\boldsymbol{X}^{h'}|Z_k^{h'}=1) \sim \boldsymbol{D}_k^{h,h'}(\boldsymbol{X}^{h}|Z_k^{h}=1) + \boldsymbol{b}_k^{h,h'}. \tag{4.38}$$

Propriété 1. *Si $\boldsymbol{X}$ est un vecteur aléatoire de $\mathbb{R}^d$ distribué selon une loi de Student de paramètre $(\nu,\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\boldsymbol{D} \in \mathbb{R}^{d\times d}$ une matrice inversible et $\boldsymbol{b}$ un vecteur de $\mathbb{R}^d$ alors le vecteur aléatoire $\boldsymbol{D}\boldsymbol{X}+\boldsymbol{b}$ suit la loi de Student de paramètre $(\nu,\boldsymbol{D}\boldsymbol{\mu}+\boldsymbol{b},\boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D})$.*

Preuve. On note

$$\xi : \left| \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ \boldsymbol{x} & \longmapsto & \boldsymbol{D}\boldsymbol{x}+\boldsymbol{b} \end{array} \right.$$

La densité de $\boldsymbol{X}$ est $t_d(\boldsymbol{x};\nu,\boldsymbol{\mu},\boldsymbol{\Sigma})$. On en déduit que celle de $\boldsymbol{Y}=\xi(\boldsymbol{X})$ est :

$$t_d(\xi^{-1}(\boldsymbol{y});\nu,\boldsymbol{\mu},\boldsymbol{\Sigma}) / \left|\boldsymbol{J}_{\xi^{-1}(\boldsymbol{y})}\xi\right| = t_d(\boldsymbol{y};\nu,\boldsymbol{D}\boldsymbol{\mu}+\boldsymbol{b},\boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D}). \tag{4.39}$$

($\boldsymbol{J}_{\xi^{-1}(\boldsymbol{y})}\xi$ désigne la matrice jacobienne de l'application ξ, calculée au point $\xi^{-1}(\boldsymbol{y})$.) □

Puisqu'une transformation affine (inversible) ne modifie pas le degré de liberté d'un vecteur de Student, les populations conditionnelles ont le même dégré de liberté. Pour tout $k \in \{1, \dots, K\}$:

$$\nu_k^1 = \nu_k^2 = \cdots = \nu_k^H = \nu_k. \tag{4.40}$$

Le modèle (4.38) implique également que les paramètres de localisation de la composante k dans les mélanges h et h' sont liés par :

$$\boldsymbol{\mu}_k^{h'} = \boldsymbol{D}_k^{h,h'} \boldsymbol{\mu}_k^h + \boldsymbol{b}_k^{h,h'}, \tag{4.41}$$

et les matrices de produits scalaires par :

$$\boldsymbol{\Sigma}_k^{h'} = \boldsymbol{D}_k^{h,h'} \boldsymbol{\Sigma}_k^h \boldsymbol{D}_k^{h,h'}. \tag{4.42}$$

On ne doit pas confondre les égalités (4.41)et (4.42) avec les propriétés générales de l'espérance (linéarité) et de la variance d'un vecteur aléatoire. Rappelons en effet que $E(\boldsymbol{X}^h | Z_k^h = 1) = \boldsymbol{\mu}_k^h$ si et seulement si $\nu_k^h > 1$, et $V(\boldsymbol{X}^h | Z_k^h = 1) = \boldsymbol{\Sigma}_k^h$ si et seulement si $\nu_k^h > 2$. Or (4.41) et (4.42) sont vérifiées sous le modèle (4.38) même lorsque les degrés de liberté sont strictement inférieurs à 1.

Les relations (4.40) (4.41)et (4.42) caractérisent le modèle (4.38) c'est à dire le modèle général de lien affine entre mélanges de Student.
Des mélanges de Student K-modaux (K composantes par mélange) P^h $(h = 1, \dots, H)$ - Rappelons que la composante k du mélange h est paramétrée par sa proportion dans le mélange π_k^h, son degré de liberté ν_k^h, son paramètre de localisation $\boldsymbol{\mu}_k^h$ et sa matrice de produits scalaires $\boldsymbol{\Sigma}_k^h$. - constituent un modèle de lien affine pour la classification simultanée si :

- les degrés de liberté des composantes de même label verifient (4.40),
- étant données deux composantes de même label k (quelconque) dans deux populations h et h' (quelconques) il existe une matrice $\boldsymbol{D}_k^{h,h'}$ de $\mathbb{R}^{d \times d}$ diagonale et inversible, et un vecteur $\boldsymbol{b}_k^{h,h'}$ de $\mathbb{R}^d$ tels que (4.41) et (4.42) .

Sans autre hypothèse sur les vecteurs $\boldsymbol{b}_k^{h,h'}$ la condition (4.41) n'est pas contraignante et les relations (4.40) et (4.42) suffisent à caractériser le modèle général de lien affine entre mélanges de Student.

Le modèle général (4.38) est identifiable - l'identifiabilité du paramètre $\boldsymbol{\psi}$ est nécessaire à la consistence de l'estimateur du maximum de vraisemblance - à une permutation près du label des populations et à une permutation près (la même dans chaque population) du label des composantes.

Le lien paramétrique entre les composantes de Student de même label, qui caractérise le modèle, lui en revanche, n'est pas identifiable. En effet, si les égalités (4.41) et (4.42) sont vérifiées pour un couple $[\boldsymbol{D}_k^{h,h'}, \boldsymbol{b}_k^{h,h'}]$, elles le sont aussi pour le couple $[-\boldsymbol{D}_k^{h,h'}, 2\boldsymbol{D}_k^{h,h'}\boldsymbol{\mu}_k^h + \boldsymbol{b}_k^{h,h'}]$.

Supposons désormais que les matrices $\boldsymbol{D}_k^{h,h'}$ en plus d'être diagonales et inversibles, sont positives. Cela revient à postuler que la corrélation de deux covariables conditionnelles quelconques (lorsqu'elle existe, c'est à dire lorsque le degré de liberté des vecteurs conditionnels considérés est supérieur à 2) ne change pas de signe d'une population à l'autre. Dans ce cas - mais pas uniquement dans celui là - le lien paramètrique entre composantes de Student est identifiable : les couples $[\boldsymbol{D}_k^{h,h'}, \boldsymbol{b}_k^{h,h'}]$ vérifiant (4.41) et (4.42) sont uniques.

On peut définir des modèles parcimonieux en imposant, comme dans le cas gaussien, des contraintes portant directement sur les paramètres des mélanges (modèles intra-populations) et des contraintes portant sur le paramètre de lien (modèles inter-populations).

Modèles intra-populations

On peut supposer que dans chaque mélange les proportions des composantes sont égales (π) ou libres (π_k) et, indépendamment, que les degrés de liberté sont égaux (ν) ou libres (ν_k). On peut imposer aux matrices de produits scalaires $\boldsymbol{\Sigma}_k^h$ $(k = 1, \ldots, K)$ d'être égales ($\boldsymbol{\Sigma}_k^h = \boldsymbol{\Sigma}^h$) ou faire l'hypothèse qu'elles sont libres.

Remarque. Supposer que les matrices de produits scalaires d'un mélange de Student sont égales ne signifie pas (même si les degrés de liberté des composantes du mélange sont strictement supérieurs à 2) que les composantes de ce mélange sont homoscédastiques, mais simplement que leurs matrices de covariances ont même forme et même orientation (Elles peuvent en effet avoir des volumes différents.). Pour que le mélange soit homoscédastique, il faut en plus que les degrés de liberté des composantes soient égaux.

Modèles inter-populations

On peut envisager les mêmes contraintes sur les matrices de transformation $\boldsymbol{D}_k^{h,h'}$ et sur les vecteurs de translation $\boldsymbol{b}_k^{h,h'}$ que dans le cas gaussien.

Dans le cas général les matrices $\boldsymbol{D}_k^{h,h'}$ sont diagonales inversibles positives. Elles peuvent aussi être indépendantes de la composante ($\boldsymbol{D}_k^{h,h'} = \boldsymbol{D}^{h,h'}$), agir uniformément sur les variables ($\boldsymbol{D}_k^{h,h'} = \alpha_k^{h,h'}\boldsymbol{I}; \alpha_k^{h,h'} \in \mathbb{R}_*^+$), agir uniformément sur les variables et indépendamment de la composante ($\boldsymbol{D}_k^{h,h'} = \alpha^{h,h'}\,\boldsymbol{I}; \alpha^{h,h'} \in \mathbb{R}_*^+$) ou être toutes égales à l'identité ($\boldsymbol{D}_k^{h,h'} = \boldsymbol{I}$) si l'on considère que les populations conditionnelles ne diffèrent

que par leur localisation. Les vecteurs de translations $\boldsymbol{b}_k^{h,h'}$ sont libres dans le cas général. On peut supposer qu'ils sont homogènes en k ($\boldsymbol{b}_k^{h,h'} = \boldsymbol{b}^{h,h'}$), ou nuls ($\boldsymbol{b}_k^{h,h'} = \mathbf{0}$). Enfin on peut supposer que les vecteurs de proportions mélange $(\pi_1^h, \ldots, \pi_K^h)$ $(h = 1, \ldots, H)$ sont libres (π^h) ou égaux (π).

On note $\left(\pi^h, \boldsymbol{D}_k^{h,h'}, \boldsymbol{b}_k^{h,h'}; \nu_k, \pi_k, \boldsymbol{\Sigma}_k^h\right)$ le modèle général de lien affine entre mélanges de Student.

Bien que $\boldsymbol{\Sigma}_k^h$ ne fasse pas référence dans cette notation à un paramètre gaussien, on peut établir une correspondance entre chaque modèle de lien affine entre mélanges gaussiens et deux modèles de lien affine entre mélanges de Student. Ainsi le modèle général $\left(\pi^h, \boldsymbol{D}_k^{h,h'}, \boldsymbol{b}_k^{h,h'}; \pi_k, \boldsymbol{\Sigma}_k^h\right)$ défini dans le cas gaussien donne lieu à deux modèles dans le cas de mélanges de Student $\left(\pi^h, \boldsymbol{D}_k^{h,h'}, \boldsymbol{b}_k^{h,h'}; \nu, \pi_k, \boldsymbol{\Sigma}_k^h\right)$ et $\left(\pi^h, \boldsymbol{D}_k^{h,h'}, \boldsymbol{b}_k^{h,h'}; \nu_k, \pi_k, \boldsymbol{\Sigma}_k^h\right)$ qui ne diffèrent que par l'hypothèse faite sur les degrés de liberté.

Les incompatibilités entre certaines hypothèses intra et inter-populations remarquées dans le cas gaussien demeurent dans le cas de Student. On ne pouvait pas dans le cas gaussien supposer les composantes de chaque mélange homoscédastiques si les matrices $\boldsymbol{D}_k^{h,h'}$ étaient libres. Dans le cas de mélanges de Student, si les matrices de transformation sont homogènes en k, les matrices de produits scalaires ne peuvent pas être libres.

Un modèle qui était identifiable dans le cas gaussien correspond toujours à un modèle identifiable dans le cas de mélanges de Student, quelle que soit l'hypothèse que l'on formule sur les degrés de liberté des composantes.
En revanche certains modèles non identifiables dans le cas gaussien correspondent à des modèles identifiables dans le cas de mélanges de Student. Supposer dans le cas gaussien que les proportions mélanges π_k^h valent toutes $1/K$, que les matrices $\boldsymbol{D}_k^{h,h'}$ sont homogènes en k et que les vecteurs $\boldsymbol{b}_k^{h,h'}$ sont libres, conduit à un modèle non identifiable. Dans le cas de mélanges de Student les mêmes contraintes n'empêchent pas le modèle d'être identifiable si les degrés de liberté sont libres d'une composante à l'autre. Le tableau 4.16 indique parmi les combinaisons possibles de contraintes intra et inter-populations, celles qui sont autorisées ainsi que les modèles identifiables.

Dans ce contexte de lien affine entre mélanges de Student, les composantes $\boldsymbol{\psi}^h$ $(h = 1, \ldots, H)$ du paramètre $\boldsymbol{\psi}$ ne peuvent plus être estimées indépendamment les unes des autres. Les relations (4.40) (4.41) (4.42) qui les lient empêchent en effet l'optimisation indépendante de leur vraisemblance. Un algorithme EM qui optimiserait la vraisemblance du paramètre $\boldsymbol{\psi}$ n'est pas non plus envisageable, la log-vraisemblance de $\boldsymbol{\psi}$ calculée sur la donnée complète $\mathbf{x}_c$ n'ayant plus d'optimum explicite. On peut cependant recourir à un algorithme GEM dont les propriétés de convergence sont identiques à celles d'EM ([37]).

Les relations (4.40) (4.41) (4.42) permettent (puisque les matrices $\boldsymbol{D}_k^{h,h'}$ et les vecteurs $\boldsymbol{b}_k^{h,h'}$ sont uniques) de reparamétrer le modèle décrit par $\boldsymbol{\psi}$ en définissant $\boldsymbol{\theta}^1 = \boldsymbol{\psi}^1$,

			Modèles intra-populations							
			ν				ν_k			
			π		π_k		π		π_k	
Modèles inter-populations			$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$	$\boldsymbol{\Sigma}^h$	$\boldsymbol{\Sigma}_k^h$
π (π^h)		$\mathbf{0}$	• (.)	• (.)	•(•)	•(•)	• (.)	• (.)	•(•)	•(•)
	$\boldsymbol{I}$,$\alpha^{h,h'}\boldsymbol{I}$,$\boldsymbol{D}^{h,h'}$	$\boldsymbol{b}^{h,h'}$	• (.)	• (.)	•(•)	•(•)	• (.)	• (.)	•(•)	•(•)
		$\boldsymbol{b}_k^{h,h'}$	∘ (.)	• (.)	•(•)	•(•)	• (.)	• (.)	•(•)	•(•)
	$\alpha_k^{h,h'}\boldsymbol{I}$,$\boldsymbol{D}_k^{h,h'}$	$\mathbf{0}$	. (.)	• (.)	. (.)	• (•)	. (.)	• (.)	. (.)	• (•)
		$\boldsymbol{b}_k^{h,h'}$	. (.)	• (.)	. (.)	• (•)	. (.)	• (.)	. (.)	• (•)

TABLE 4.16: *Combinaisons autorisées de contraintes intra/interpopulations et modèles identifiables. On note '.' une combinaison non autorisée, '∘' une combinaison autorisée correspondant à un modèle non identifiable, et '•', une combinaison autorisée correspondant à un modèle identifiable.*

et pour tout $h \in \{2, \ldots, H\}$, $\boldsymbol{\theta}^h = \left[\left(\pi_k^h, \boldsymbol{D}_k^h, \boldsymbol{b}_k^h\right); k = 1, \ldots, K\right]$, où $\boldsymbol{D}_k^h = \boldsymbol{D}_k^{1,h}$ et $\boldsymbol{b}_k^h = \boldsymbol{b}_k^{1,h}$. On note $\boldsymbol{\theta} = (\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^H)$, et Θ l'espace décrit par $\boldsymbol{\theta}$ lorsque le paramètre $\boldsymbol{\psi}$ décrit l'espace Ψ. Il existe alors une bijection canonique entre les espaces Ψ et Θ et il est équivalent d'estimer $\boldsymbol{\psi}$ ou $\boldsymbol{\theta}$ en maximisant leur vraisemblance sur respectivement Ψ ou Θ.

L'algorithme GEM par lequel on estime le nouveau paramètre $\boldsymbol{\theta}$ consiste, à partir d'une valeur initiale du paramètre, à alterner les deux étapes suivantes :

Etape E : A partir de la valeur en cours $\tilde{\boldsymbol{\theta}}$ du paramètre $\boldsymbol{\theta}$, les probabilités conditionnelles sont calculées selon (4.28) et les coefficients u_i^h sont estimés selon (4.26).

Etape GM : Les proportions des mélanges sont estimées selon $\pi_k^h = \sum_{i=1}^{n^h} t_{i,k}^h / n^h$ quand on les suppose libres, par $\pi_k^h = \sum_{k=1}^{K} \sum_{i=1}^{n^h} t_{i,k}^h / n$ si elles sont homogènes en h, et par $\pi_k^h = 1/K$ quand on les suppose égales.

Les degrés de liberté sont estimés en résolvant les équations de vraisemblance :

$$\frac{\partial}{\partial \nu_k}\left[\sum_{h=1}^{H}\sum_{i=1}^{n^h} t_{i,k}^h\{-\ln\Gamma\left(\frac{\nu_k}{2}\right)+\frac{\nu_k}{2}\ln\left(\frac{\nu_k}{2}\right)-\frac{\nu_k}{2}\tilde{u}_i^h+\left(\frac{d+\nu_k}{2}-1\right)\times\left(\ln\tilde{u}_{i,k}^h+\frac{\Gamma'((\tilde{\nu}_k+d)/2)}{\Gamma((\tilde{\nu}_k+d)/2)}-\ln((\tilde{\nu}_k+d)/2)\right)\}\right]=0, \tag{4.43}$$

lorsqu'ils sont libres, et l'équation :

$$\frac{\partial}{\partial \nu_k}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{n^h} t_{i,k}^h\{-\ln\Gamma\left(\frac{\nu_k}{2}\right)+\frac{\nu_k}{2}\ln\left(\frac{\nu_k}{2}\right)-\frac{\nu_k}{2}\tilde{u}_i^h+\left(\frac{d+\nu_k}{2}-1\right)\times\left(\ln\tilde{u}_{i,k}^h+\frac{\Gamma'((\tilde{\nu}_k+d)/2)}{\Gamma((\tilde{\nu}_k+d)/2)}-\ln((\tilde{\nu}_k+d)/2)\right)\}\right]=0, \tag{4.44}$$

lorsqu'ils sont homogènes.

Les autres composantes du paramètre sont estimées itérativement de façon à augmenter leur log-vraisemblance attendue :

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{i=1}^{n^h} t_{i,k}^h\{-\frac{d}{2}\ln(2\pi)-\frac{1}{2}\ln|\boldsymbol{\Sigma}_k^1|-\ln|\boldsymbol{D}_k^h|-\frac{1}{2}\tilde{u}_i^h\left(\left(\boldsymbol{D}_k^h\right)^{-1}(\mathbf{x}_i^h-\boldsymbol{b}_k^h)-\boldsymbol{\mu}_k^1\right)'\left(\boldsymbol{\Sigma}_k^1\right)^{-1}\left(\left(\boldsymbol{D}_k^h\right)^{-1}(\mathbf{x}_i^h-\boldsymbol{b}_k^h)-\boldsymbol{\mu}_k^1\right)\}, \tag{4.45}$$

en alternant (dans cet ordre) les étapes suivantes :

Estimation du paramètre de localisation des classes

Le paramètre de localisation des classes de la population de référence est estimé par :

$$\boldsymbol{\mu}_k^1=\frac{\sum_{h=1}^{H}\sum_{i=1}^{n^h} t_{i,k}^h\tilde{u}_{i,k}^h\left(\boldsymbol{D}_k^h\right)^{-1}(\mathbf{x}_i^h-\boldsymbol{b}_k^h)}{\sum_{h=1}^{H}\sum_{i=1}^{n^h} t_{i,k}^h\tilde{u}_{i,k}^h}. \tag{4.46}$$

Estimation des matrices de produits scalaires

Les matrices de produits scalaires de la population de référence sont estimées par :

$$\boldsymbol{\Sigma}_k^1=\frac{\sum_{h=1}^{H}\sum_{i=1}^{n^h} t_{i,k}^h\tilde{u}_{i,k}^h\left[\left(\boldsymbol{D}_k^h\right)^{-1}(\mathbf{x}_i^h-\boldsymbol{b}_k^h)-\boldsymbol{\mu}_k^1\right]\left[\left(\boldsymbol{D}_k^h\right)^{-1}(\mathbf{x}_i^h-\boldsymbol{b}_k^h)-\boldsymbol{\mu}_k^1\right]'}{\sum_{h=1}^{H}\sum_{i=1}^{n^h} t_{i,k}^h}, \tag{4.47}$$

quand elles sont libres, et par :

$$\boldsymbol{\Sigma}_k^1=\frac{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{n^h} t_{i,k}^h\tilde{u}_{i,k}^h\left[\left(\boldsymbol{D}_k^h\right)^{-1}(\mathbf{x}_i^h-\boldsymbol{b}_k^h)-\boldsymbol{\mu}_k^1\right]\left[\left(\boldsymbol{D}_k^h\right)^{-1}(\mathbf{x}_i^h-\boldsymbol{b}_k^h)-\boldsymbol{\mu}_k^1\right]'}{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{n^h} t_{i,k}^h}, \tag{4.48}$$

quand on les suppose homogènes.

Estimation des vecteurs de translation

Les vecteurs de translation sont estimés par :

$$\boldsymbol{b}_k^h = \left(\sum_{i=1}^{n^h} t_{i,k}^h \tilde{u}_{i,k}^h \mathbf{x}_i^h \right) / \left(\sum_{i=1}^{n^h} t_{i,k}^h \tilde{u}_{i,k}^h \right) - \boldsymbol{D}_k^h \boldsymbol{\mu}_k^1, \tag{4.49}$$

quand ils sont libres, et par :

$$\boldsymbol{b}_k^h = \left[\sum_{i=1}^{n^h} \sum_{k=1}^{K} t_{i,k}^h \tilde{u}_{i,k}^h \left(\boldsymbol{D}_k^h \boldsymbol{\Sigma}_k^1 \boldsymbol{D}_k^h \right)^{-1} \right]^{-1} \left[\sum_{i=1}^{n^h} \sum_{k=1}^{K} t_{i,k}^h \tilde{u}_{i,k}^h \left(\boldsymbol{D}_k^h \boldsymbol{\Sigma}_k^1 \boldsymbol{D}_k^h \right)^{-1} \left(\mathbf{x}_i^h - \boldsymbol{D}_k^h \boldsymbol{\mu}_k^1 \right) \right] \tag{4.50}$$

quand ils sont homogènes.

Estimation des matrices de transformation

Quel que soit le modèle considéré, (4.45) est concave par rapport à $\left(\boldsymbol{D}_k^h\right)^{-1}$ et l'on peut toujours l'optimiser (lorsque les autres composantes du paramètre sont fixées) par un algorithme ad-hoc. Si son maximum n'est pas explicite dans les modèles $\boldsymbol{D}_k^{h,h'}$ et $\boldsymbol{D}^{h,h'}$, il l'est, en revanche, dans les modèles $\alpha_k^{h,h'}\boldsymbol{I}$ et $\alpha^{h,h'}\boldsymbol{I}$; Il correspond alors respectivement à $\boldsymbol{D}_k^h = \alpha_k^h I$ et $\boldsymbol{D}_k^h = \alpha^h I$ avec :

$$\alpha_k^h = \frac{4(\mathbf{1}_d' \boldsymbol{Q}_k^h \mathbf{1}_d)}{-(\boldsymbol{L}_k^h \mathbf{1}_d) + \sqrt{\left(\boldsymbol{L}_k^h \mathbf{1}_d\right)^2 - 8dC_k^h (\mathbf{1}_d' \boldsymbol{Q}_k^h \mathbf{1}_d)}}, \tag{4.51}$$

et

$$\alpha^h = \frac{4(\mathbf{1}_d' \boldsymbol{Q}^h \mathbf{1}_d)}{-(\boldsymbol{L}^h \mathbf{1}_d) + \sqrt{\left(\boldsymbol{L}^h \mathbf{1}_d\right)^2 - 8dC^h (\mathbf{1}_d' \boldsymbol{Q}^h \mathbf{1}_d)}}, \tag{4.52}$$

où :
$$\boldsymbol{Q}_k^h = \sum_{i=1}^{n^h} t_{i,k}^h \tilde{u}_{i,k}^h \mathrm{diag}(\mathbf{x}_i^h - \boldsymbol{b}_k^h) \left(\boldsymbol{\Sigma}_k^1\right)^{-1} \mathrm{diag}(\mathbf{x}_i^h - \boldsymbol{b}_k^h),$$

$$\boldsymbol{L}_k^h = -2(\boldsymbol{\mu}_k^1)' (\boldsymbol{\Sigma}_k^1)^{-1} \sum_{i=1}^{n^h} t_{i,k}^h \tilde{u}_{i,k}^h \mathrm{diag}(\mathbf{x}_i^h - \boldsymbol{b}_k^h),\ C_k^h = -2\hat{n}_k^h,\ C^h = \sum_{k=1}^{K} C_k^h,\ \boldsymbol{L}^h = \sum_{k=1}^{K} \boldsymbol{L}_k^h,$$

$Q^h = \sum_{k=1}^{K} \boldsymbol{Q}_k^h$ et $\mathbf{1}_d$ désigne le vecteur de $\mathbb{R}^d$ dont toutes les composantes valent 1.

4.3.3 Application numérique à des données financières

On prédit généralement la faillite d'une entreprise en estimant sa capacité à rembourser ses dettes. Des modèles ad-hoc permettent de mesurer par un calcul de score, la distance de l'entreprise à une situation critique de non remboursement. Or ces modèles considèrent la banqueroute d'un point de vue statique. Prenant ce point de vue à contre-pied, J. Du Jardin et E. Séverin mettent en évidence dans [39] le caractère dynamique de la faillite. Ils notent qu'elle résulte d'une évolution et s'attachent à étudier la

trajectoire des entreprises dans le temps, afin tout à la fois de caractériser et de prévenir les situations de banqueroute.

Dans cette section nous reprenons l'étude de cas menée par C. Biernacki et A. Lourme ([80]), sur les données de J. Du Jardin et E. Séverin ([39]). Nous proposons d'appliquer la classification simultanée robuste à l'analyse typologique de plusieurs échantillons d'entreprises considérées à des instants distincts, et nous interprétons les modèles inférés comme une formalisation de la dynamique qui conduit certaines des entreprises à la faillite.

On considère deux échantillons d'entreprises $\mathbf{x}^1$ ($n^1 = 428$ entreprises observées en 2002, dont 49.5% ont fait faillite) et $\mathbf{x}^2$ ($n^2 = 461$ entreprises de 2003, dont 47.7% ont fait faillite). Les entreprises des deux années sont décrites par les mêmes rapports de variables économétriques : Excédent Brut Exploitation/Actif Total, Valeur Ajoutée/Chiffre d'Affaire, Stock/Dette Court Terme, Fournisseur/Chiffre d'Affaire. FIG. 4.7 représente les entreprises de 2002 et 2003 dans le plan canonique [Excédent Brut Exploitation/Actif Total, Stock/Dette Court Terme] et montre que les deux échantillons proviennent de populations distinctes.

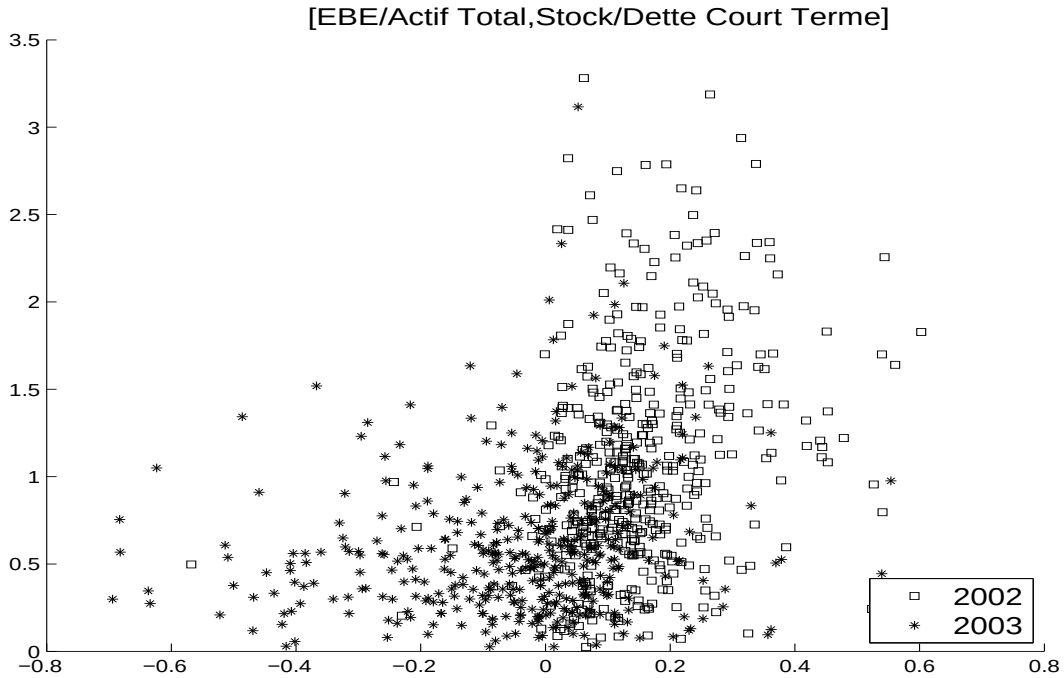


FIGURE 4.7: *Deux échantillons d'entreprises de* 2002 *et* 2003.

On notera que ces données se placent bien dans un contexte de classification simultanée. D'une part les descripteurs des entreprises sont les mêmes d'une année à l'autre. D'autre part on recherche dans les deux échantillons d'entreprises une partition dont on pourra donner la même interprétation financière.

TAB. 4.17 indique les différentes valeurs du critère ICL (1.51) obtenues en classifiant

les échantillons indépendamment et simultanément, pour un nombre de groupes variant de 1 à 5. Les modèles de classification simultanée font référence ici, aux modèles autorisés de TAB. 4.16 qui supposent les matrices $\boldsymbol{D}_k^{h,h'}$ et les vecteurs $\boldsymbol{b}_k^{h,h'}$ libres ou homogènes en k. Ils sont inférés grâce à l'algorithme GEM décrit à la section précédente (On trouvera dans [80] le détail de la procédure, le nombre d'itérations de l'algorithme, le nombre de points de départs, etc.).
Le critère ICL, décrit à la section 1.3 comme plus robuste que BIC, semble mieux convenir à la détermination de groupes interprétables parmi ces données financières. Nous pensons en effet qu'il peut éviter (comme dans le cas gaussien) que l'inadéquation (éventuelle) du modèle conditionnel ne conduise à l'inférence de fausses composantes.

K	1	2	3	4	5
Classif. simultanée	-1169.7	-1191.3	$\mathbf{-1202.0}$	-1183.4	-1131.3
Classif. indépendante	-1154.6	$\mathbf{-1163.6}$	-1072.1	-1127.7	-1098.3

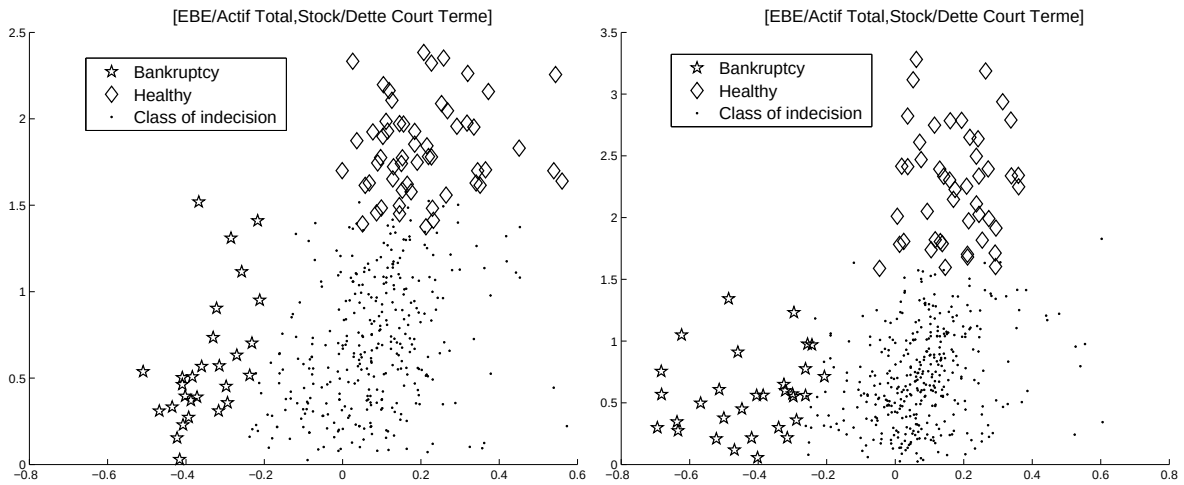
TABLE 4.17: *Les meilleures valeurs d'ICL obtenues en classification indépendante et simultanée, pour un nombre variable de groupes*

La préférence d'ICL se porte sur un modèle de classification simultanée correspondant à $K = 3$ groupes (voir TAB. 4.17). Si l'on compare la partition induite par ce modèle à la vraie partition des entreprises (tableau de confusion TAB. 4.18), on observe que les groupes 1 et 2 inférés, sont fortement corrélés aux entreprises respectivement en faillite et en bonne santé. Ainsi le modèle à trois groupes retenu, éclaire la typologie des entreprises d'un jour nouveau. L'état de faillite ou de bonne santé est facile à déterminer (voir FIG. 4.8) pour quelques entreprises (les groupes 1 et 2 représentent respectivement 7% et 13% des données) et plus difficile à déceler pour la majeure partie d'entre elles (le groupe 3 comporte 80% des données).

	groupe 1	groupe 2	groupe 3
non-faillite	3	94	360
faillite	56	10	366

TABLE 4.18: *Comparaison de la vraie partition faillite/non-faillite et de la partition relative au meilleur modèle de la classification simultanée ($ICL = -1202.0$)*

Les groupes inférés (faillite, santé, indécision) parmi les entreprises de 2002 et 2003 peuvent évidemment être décrits par les paramètres conditionnels, selon une méthodologie habituelle en matière de mélanges. Mais focalisons-nous ici sur l'interprétation de l'évolution des entreprises que permet le modèle de classification simultanée retenu. Le meilleur modèle selon ICL est $(\pi, \boldsymbol{D}^{h,h'}, \boldsymbol{b}^{h,h'}; \nu, \pi_k, \boldsymbol{\Sigma}^h)$. Cela signifie que la proportion des entreprises dans les trois groupes (faillite, santé, indécision) est invariante de 2002 à 2003. D'autre part d'après ce modèle, le centre et la corrélation des covariables évoluent au cours du temps de façon homogène d'une classe à l'autre : les matrices $\boldsymbol{D}_k^{1,2}$ et les vecteurs $\boldsymbol{b}_k^{1,2}$ ne dépendent pas de k. Plus précisément, les paramètres de transition estimés sont $\hat{\boldsymbol{D}}^{1,2} = \mathrm{diag}(1.12, 0.95, 1.20, 0.93)$ et $\hat{\boldsymbol{b}}^{1,2} = 10^{-3}.(-18.2, 2, -102, -1)'$. Ainsi,



(a) *année* 2002 (b) *année* 2003

FIGURE 4.8: *Partition (faillite, bonne santé, indécision) des entreprises de* 2002 *et* 2003*, estimée en classification simultanée robuste (basée sur un mélange de Student).*

ce sont essentiellement les rapports EBE/Actif Total et Stock/Dette Court Terme qui caractérisent le déplacement des trois catégories d'entreprises (faillite, santé, indécision) d'une année à l'autre.
Ce résultat fait sens d'un point de vue financier. Les principaux prédicteurs de la faillite d'une entreprise sont en effet la performance et les liquidités, dont les rapports EBE/Actif Total et Stock/Dette Court Terme constituent des indicateurs. Il est généralement admis que les modifications de la structure financière d'une entreprise sont une conséquence de l'évolution de ces deux variables.

Ainsi, non seulement ICL choisit un modèle de classification simultanée mais ce modèle permet à la fois (i) d'obtenir une partition des entreprises dont on peut interpréter les classes, (ii) de caractériser chaque classe d'entreprise par la valeur du paramètre conditionnel inféré, (iii) de caractériser l'évolution des entreprises en faillite ou en bonne santé par la valeur du paramètre de transition.

La classification indépendante des entreprises est moins pertinente. Le modèle auquel elle conduit n'est pas le meilleur pour ICL. Ce modèle détermine $K = 2$ groupes parmi les entreprises des deux années mais, comme le montre TAB. 4.19, les groupes inférés ne sont pas corrélés à la santé réelle des entreprises. Enfin le modèle inféré en classification indépendante ne permet pas d'interpréter l'évolution des catégories d'entreprises au cours du temps.

	groupe 1	groupe 2
non-faillite	228	229
faillite	289	143

Table 4.19: *Comparaison de la vraie partition faillite/non-faillite et de la partition relative au meilleur modèle de la classification indépendante ($ICL = -1163.6$)*

4.4 Mélanges de Factor Analyzers (Une perspective pour la classification simultanée en grande dimension)

Les résultats encourageants de la classification simultanée nous incitent à étendre la méthode aux données de grande dimension. La problématique et les notations sont les mêmes qu'à la section 4.1. On dispose de H échantillons $\mathbf{x}^h = (\mathbf{x}_i^h; i = 1, \ldots, n^h)$ $(h = 1, \ldots, H)$ de $\mathbb{R}^d$ décrits par d variables continues, à partitionner en K groupes chacun ; on suppose que les données de l'échantillon h proviennent de façon indépendante d'un mélange gaussien de densité :

$$f(\boldsymbol{x}; \boldsymbol{\psi}^h) = \sum_{k=1}^{K} \pi_k^h \Phi_d(\boldsymbol{x}; \boldsymbol{\mu}_k^h, \boldsymbol{\Sigma}_k^h), \ \boldsymbol{x} \in \mathbb{R}^d. \tag{4.53}$$

Dans cette section on considère que la dimension d est grande et que l'information apportée par la donnée fait défaut dans certaines directions de l'espace.

On peut réduire la variabilité du modèle de façon élémentaire, en supposant que les différents échantillons proviennent de la même population et que le paramètre des mélanges est identique : $\boldsymbol{\psi}^1 =, \ldots, = \boldsymbol{\psi}^H$. Mais en grande dimension, on manque d'information ; il ne faut donc pas se priver de celle qui est disponible. D'une part les échantillons sont distincts ; on doit donc au moins envisager que les populations d'origine le sont. D'autres part les échantillons sont décrits par les mêmes variables et l'on recherche le même nombre de groupes dans chacun d'eux ; cela nous encourage comme dans les sections précédentes à établir un lien entre les groupes deux à deux. Nous chercherons donc à la fois à établir une relation entre populations conditionnelles et à formaliser le défaut d'information dans certaines régions de l'espace.

Nous avons décrit à la section 1.1.4, deux types de mélanges gaussiens employés en grande dimension. Les premiers sont les modèles de C. Bouveyron ([21]). Ils reposent sur la décomposition canonique des matrices de covariances :

$$\boldsymbol{\Sigma}_k^h = \boldsymbol{S}_k^h \boldsymbol{\Lambda}_k^h (\boldsymbol{S}_k^h)', \tag{4.54}$$

où $\boldsymbol{\Lambda}_k^h$ est la matrice diagonale des valeurs propres de $\boldsymbol{\Sigma}_k^h$ (rangées par ordre décroissant) et $\boldsymbol{S}_k^h$ une matrice orthogonale de $\mathbb{R}^{d \times d}$. La parcimonie des modèles de C. Bouveyron sous aucune de ses formes n'est compatible avec la transformation stochastique mutuelle (4.3) des populations conditionnelles. (4.3) ne permet pas de supposer homogènes

les dernières valeurs propres de $\boldsymbol{\Sigma}_k^h$ (parcimonie par homogénéisation du paramètre dans une classe : $\boldsymbol{\Lambda}_k^h = \mathrm{diag}(\alpha_{1,k}^h + \beta_k^h, \ldots, \alpha_{\delta,k}^h + \beta_k^h, \beta_k^h, \ldots, \beta_k^h)$). La transformation (4.3) ne permet pas non plus de supposer homogène l'orientation des classes (parcimonie par homogénéisation du paramètre entre classes : $\boldsymbol{S}_1^h = \cdots = \boldsymbol{S}_K^h$).

Tout en traduisant l'idée que les données conditionnelles vivent dans des sous-espaces de dimension réduite, les mélanges de Factor Analyzers permettent, eux, d'établir un lien entre ces sous-espaces, interprétable géométriquement.

Les mélanges de Factor Analyzers dans le contexte simultané

Les mélanges de Factor Analyzers supposent que chaque matrice de covariances dans la population h peut s'écrire :

$$\boldsymbol{\Sigma}_k^h = \boldsymbol{B}_k^h (\boldsymbol{B}_k^h)' + \boldsymbol{\Delta}_k^h, \tag{4.55}$$

où $\boldsymbol{B}_k^h$ désigne une matrice de dimension $d \times b$ ($b < d$) et $\boldsymbol{\Delta}_k^h$ une matrice diagonale, définie et positive. Pour ce modèle, la donnée de la classe k vit essentiellement dans le sous-espace $\mathfrak{B}_k^h$ engendré par $\boldsymbol{B}_k^h(\bullet, j)$ ($j = 1, \ldots, b$), les vecteurs colonnes de $\boldsymbol{B}_k^h$. En effet la loi du vecteur conditionnel relatif à cette classe, peut se décomposer en :

$$(\mathbf{X}^h | Z_k^h = 1) \sim \boldsymbol{\mu}_k^h + \boldsymbol{B}_k^h \boldsymbol{U}_k^h + \boldsymbol{e}_k^h. \tag{4.56}$$

Les vecteurs latents (gaussiens, centrés, réduits et indépendants) $\boldsymbol{U}_k^h$ ($k = 1, \ldots, K$) de $\mathbb{R}^b$, sont les Factor Analyzers des populations conditionnelles. Leur estimation (conditionnelle) permet une représentation simplifiée des données dans les sous-espaces latents $\mathfrak{B}_k^h$. Leurs réalisations constituent, avec celles de la variable catégorielle $\mathbf{Z}^h$, la donnée manquante permettant d'estimer le paramètre $\boldsymbol{\psi} = (\boldsymbol{\psi^h}; h = 1, \ldots, H)$ dans un algorithme AECM (voir [85], chap. 8). Les vecteurs (gaussiens, centrés et indépendants) $\boldsymbol{e}_k^h$ ($k = 1, \ldots, K$) de $\mathbb{R}^d$, sont indépendants des Factor Analyzers et leur matrice des covariances $\boldsymbol{\Delta}_k^h$ représente la dispersion de l'alea résiduel dans chaque classe.

Nous proposons d'établir un lien paramétrique entre les sous-espaces latents de données conditionnelles $\mathfrak{B}_k^h$ ($h = 1, \ldots, H$), qui peut s'interpréter de façon géométrique.

Une hypothèse de lien entre les sous-espaces latents

Nous supposerons que pour tout k et tout couple (h, h'), il existe une matrice $\boldsymbol{D}_k^{h,h'} \in \mathbb{R}^{d \times d}$ diagonale, définie et positive telle que :

$$\boldsymbol{B}_k^{h'} = \boldsymbol{D}_k^{h,h'} \boldsymbol{B}_k^h. \tag{4.57}$$

Une telle matrice $\boldsymbol{D}_k^{h,h'}$ peut s'identifier à une composée d'affinités orthogonales, qui transforme le sous-espace latent $\mathfrak{B}_k^h$ en $\mathfrak{B}_k^{h'}$. En effet chaque vecteur colonne $\boldsymbol{B}_k^h(\bullet, j)$ ($j \in \{1, \ldots, b\}$) se transforme selon (4.57) en $\boldsymbol{B}_k^{h'}(\bullet, j)$.
Ainsi selon ce modèle, les sous-espaces latents des données conditionnelles sont distincts,

mais il existe un lien de type géométrique entre eux.

Notons que lorsque $b = 1$, l'égalité (4.57) n'est pas contraignante : il existe toujours une composée d'affinités orthogonales qui transforme mutuellement deux sous-espaces vectoriels de dimension 1. Dans ce cas, (4.53) est un modèle de mélanges de Factor Analyzers indépendants.

Dans les autres cas, la transformation géométrique mutuelle des sous-espaces $\mathfrak{B}_k^h$ peut également être vue ainsi : dans les sous-espaces latents, les données conditionnelles sont corrélées de façon homogène d'une population à l'autre.

La parcimonie de ce type de mélanges pourra porter sur des paramètres propres à chaque population : la dimension b des sous-espaces latents (de 1 à $d-1$) et les paramètres $\boldsymbol{\Delta}_k^h$ de dispersion des erreurs (isotropiques ($\boldsymbol{\Delta}_k^h = \delta_k^h \boldsymbol{I}_d$; $\delta_k^h \in \mathbb{R}_*^+$), homogènes en k et libres en h, homogènes en k et en h, etc.). On pourra combiner ces deux types de parcimonie avec d'autres contraintes (commes celles envisagées à la section 4.1.3) portant sur le paramètre $\boldsymbol{D}_k^{h,h'}$ de transformation mutuelle des sous-espaces latents.

Chapitre 5

Contrainte de recouvrement égal des classes

5.1 Introduction

Lorsqu'on classifie un seul échantillon par inférence d'un mélange, imposer une contrainte au modèle a pour effet de réduire la dimension du paramètre et de diminuer la variabilité des estimateurs. Si la contrainte imposée n'est pas réaliste, elle augmente le biais du modèle estimé et, par conséquent, l'erreur moyenne de classement. Si au contraire la contrainte imposée est réaliste (si elle traduit une propriété vraie des données conditionnelles), elle améliore la qualité du classifieur.

Nous avons appliqué ce principe au chapitre (4), à la classification de plusieurs échantillons. Lorsque les échantillons sont décrits par des variables de même signification et que l'on y cherche le même nombre de groupes (éventuellement des groupes que l'on interprète de la même façon) nous avons vu qu'il pouvait être judicieux d'établir un lien conditionnel entre populations mélanges. Or cette forme de la classification simultanée perd l'essentiel de sa justification lorsque les descripteurs n'ont pas le même sens d'un échantillon à l'autre. Aussi proposons-nous dans ce chapitre, une forme nouvelle de la classification simultanée basée sur la formalisation d'un lien non plus conditionnel mais global, entre les populations.

On dispose de H échantillons $\mathbf{x}^h = (\mathbf{x}_i^h; i = 1, \ldots, n^h)$ $(h = 1, \ldots, H)$ de $\mathbb{R}^d$, à partitionner en K groupes chacun, et l'on suppose que les données de chaque échantillon $\mathbf{x}^h$ proviennent d'un mélange :

$$f^h(\boldsymbol{x}; \boldsymbol{\psi}^h) = \sum_{k=1}^{K} \pi_k^h f_k^h(\boldsymbol{x}; \boldsymbol{\alpha}_k^h); \boldsymbol{x} \in \mathbb{R}^d. \tag{5.1}$$

On rencontre parfois la même difficulté à classifier les différents échantillons parce que les données conditionnelles y sont imbriquées de façon similaire. On peut alors penser que les variables - même si elles n'ont pas la même signification d'un échantillon à l'autre - possèdent le même pouvoir discriminant, qu'elles contribuent de façon semblable à l'hétérogénité des données dans chaque échantillon. Or l'hétérogénéité des données

conditionnelles résulte directement du chevauchement des classes dans le (vrai) modèle. Supposer que les composantes des mélanges f^h se chevauchent de façon similaire, formalise une imbrication identique des groupes recherchés et devrait donc, dans la mesure ou cette contrainte est réaliste, améliorer la qualité du classifieur de chaque échantillon.

La classification simultanée dans ce chapitre, repose sur l'hypothèse d'un chevauchement homogène des classes dans les populations. Nous traduirons cette hypothèse en supposant à la section 5.2 que l'erreur de Bayes des différents mélanges est homogène, puis en supposant à la section 5.3, que leur entropie globale est homogène.

5.2 Egalisation du taux d'erreur de classement

L'erreur de classement du mélange h (son erreur de Bayes) est définie par :

$$\epsilon^h = \sum_{k=1}^{K} \pi_k^h \int_{\tilde{\Omega}_k^h} f_k^h(\boldsymbol{x}; \boldsymbol{\alpha}_k^h) \mathrm{d}\boldsymbol{x}. \tag{5.2}$$

Ω_k^h désigne la région de $\mathbb{R}^d$ dans laquelle prédomine la composante k du mélange :

$$\Omega_k^h = \{\boldsymbol{x} \in \mathbb{R}^d; \forall j \neq k, \pi_k^h f_k^h(\boldsymbol{x};, \boldsymbol{\alpha}_k^h) > \pi_j^h f_j^h(\boldsymbol{x}; \boldsymbol{\alpha}_j^h)\}, \tag{5.3}$$

et $\tilde{\Omega}_k^h$ est son complémentaire dans $\mathbb{R}^d$. Dans le cas de mélanges gaussiens homoscédastiques d'ordre deux par exemple, l'erreur de classement du mélange h est :

$$\epsilon^h = \Phi_1 \left(-\left[(\boldsymbol{\mu}_1^h - \boldsymbol{\mu}_2^h)' (\boldsymbol{\Sigma}^h)^{-1} (\boldsymbol{\mu}_1^h - \boldsymbol{\mu}_2^h) \right]^{1/2} / 2 \right), \tag{5.4}$$

où $\boldsymbol{\mu}_k^h$ $(k = 1, 2)$ désignent les centres des composantes, $\boldsymbol{\Sigma}^h$ leur matrice des covariances commune, et Φ_1 la fonction de répartition de la loi normale centrée réduite en dimension 1.
Dans les autres cas (mélanges hétéroscédastiques d'ordre deux ou mélanges gaussiens d'ordre supérieur), l'erreur de classement n'est pas explicite, pas même à la fonction de répartition près, de la loi normale centrée réduite.

Pour chaque mélange, l'erreur de classement peut être vue comme une mesure du recouvrement de ses composantes : un mélange dont les composantes sont bien séparées (resp. peu séparées) se caractérise par une erreur de classement proche de 0 (resp. proche de 1). Imposer une erreur de Bayes homogène entre les mélanges :

$$\epsilon^1 =, \ldots, = \epsilon^H, \tag{5.5}$$

apparaît donc comme une manière de traduire un chevauchement identique de leurs composantes. La figure 5.1 présente un exemple de mélanges gaussiens homoscédastiques dont l'erreur de classement est homogène.
Ainsi, les égalités (5.5) constituent une contrainte imposée au paramètre du modèle, qui traduit en un sens particulier une difficulté similaire à classifier des échantillons.

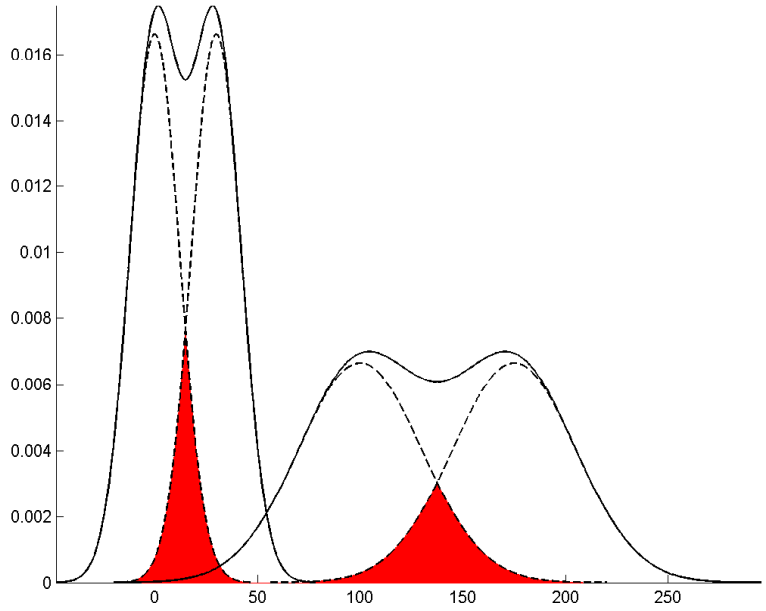


FIGURE 5.1: *Deux mélanges gaussiens homoscédastiques (d'ordre 2) ayant la même erreur de classement (en rouge)*

Au chapitre 4 la contrainte imposée aux mélanges était conditionnelle : elle consistait en un lien paramétrique entre leurs composantes. Ici la contrainte imposée par (5.5) est globale : elle porte sur les populations mélanges et non plus sur les populations conditionnelles. Mais la fonction de ce nouveau type de contrainte est identique. Sa vocation est en effet d'améliorer la qualité du modèle, à la mesure de son réalisme.

5.2.1 Le cas gaussien homoscédastique multivarié

L'erreur de classement d'un mélange (gaussien) n'est pas explicite lorsque ce mélange est hétéroscédastique ou d'ordre supérieur à deux. Aussi nous plaçons-nous ici, dans le contexte particulier de la détermination d'une structure à deux groupes ($K = 2$) par des mélanges homoscédastiques.
On suppose que chaque population P^h est un mélange de deux composantes normales de même poids, centrées respectivement en $\boldsymbol{\mu}_k^h$ $(k = 1, 2)$ et dont la matrice des covariances commune est $\boldsymbol{\Sigma}^h$. D'autre part on soumet les mélanges à la contrainte (5.5) selon laquelle leur erreur de Bayes est homogène.

Le paramètre du modèle peut alors être estimé par l'algorithme EM dédié aux mélanges, présenté à la section 1.2.2.
Son étape E consiste à déterminer la partition probabiliste $\boldsymbol{t} = (\boldsymbol{t}^h)_{h=1,\dots,H}$ de la donnée, à partir de la partition floue $\boldsymbol{t}^h = \left(t_{i,k}^h\right)_{\substack{i=1,\dots,n^h \\ k=1,\dots,K}}$ de chaque échantillon, induite par la valeur courante du paramètre.
Optimiser la log-vraisemblance complétée attendue du paramètre à l'étape M, revient

alors à minimiser le critère :

$$\sum_{h=1}^{H} n^h \left\{ \ln \left| \boldsymbol{\Sigma}^h \right| + \operatorname{tr} \left[\boldsymbol{W}^h \left(\boldsymbol{\Sigma}^h \right)^{-1} \right] \right\}, \tag{5.6}$$

où $\boldsymbol{W}^h = (1/n^h) \sum_{k=1}^{2} \sum_{i=1}^{n^h} t_{i,k}^h (\mathbf{x}_i^h - \boldsymbol{\mu}_k^h)(\mathbf{x}_i^h - \boldsymbol{\mu}_k^h)'$.
A l'optimum de (5.6) sous la contrainte (5.5), les centres des classes sont estimés par :

$$\hat{\boldsymbol{\mu}}_k^h = \left(\lambda^h \bar{\mathbf{x}}^h + \gamma^h \bar{\mathbf{x}}_k^h \right) / \left(\lambda^h + \gamma^h \right) \tag{5.7}$$

et les matrices de covariances par :

$$\hat{\boldsymbol{\Sigma}}^h = \hat{\boldsymbol{W}}^h + \lambda^h / n^h \left(\hat{\boldsymbol{\mu}}_1^h - \hat{\boldsymbol{\mu}}_2^h \right) \left(\hat{\boldsymbol{\mu}}_1^h - \hat{\boldsymbol{\mu}}_2^h \right)', \tag{5.8}$$

où $\bar{\mathbf{x}}^h$ désigne le barycentre empirique de la population h ($\bar{\mathbf{x}}^h = (n_1^h \bar{\mathbf{x}}_1^h + n_2^h \bar{\mathbf{x}}_2^h)/n^h$) et γ^h la demi-moyenne harmonique de ses effectifs conditionnels ($\gamma^h = n_1^h n_2^h / n^h$). Les coefficients λ^h ($h = 1, \ldots, H$) sont des multiplicateurs de Lagrange qui lient à l'optimum le gradient de la fonction de coût (5.6) à celui de la contrainte (5.5). Leur somme vaut zéro et ils peuvent être approchés numériquement grâce à (5.5). On remarquera que les formules (5.7) et (5.8) coïncident lorsque les multiplicateurs λ^h sont nuls c'est à dire lorsqu'on relaxe la contrainte (5.5), avec les estimateurs classiques du paramètre d'un mélange gaussien homoscédastique dans l'étape M de l'algorithme EM. Lorsque les multiplicateurs de Lagrange sont non nuls, c'est à dire lorsque la contrainte (5.5) est active, l'estimateur $\hat{\boldsymbol{\mu}}_k^h$ est un barycentre de la moyenne empirique $\bar{\mathbf{x}}^h$ et de la moyenne conditionnelle $\bar{\mathbf{x}}_k^h$. Par ailleurs, on peut montrer de façon équivalente à (5.8) que $\hat{\boldsymbol{\Sigma}}^h = \boldsymbol{P}^h + \lambda^h \gamma^h / (\lambda^h + \gamma^h) \boldsymbol{Q}^h$, où $\boldsymbol{P}^h = \sum_{k=1}^{2} \sum_{i=1}^{n^h} t_{i,k}^h (\mathbf{x}_i^h - \bar{\mathbf{x}}_k^h)(\mathbf{x}_i^h - \bar{\mathbf{x}}_k^h)' / n^h$ apparaît comme une matrice des covariances intra-classes et $\boldsymbol{Q}^h = \left(\bar{\mathbf{x}}_1^h - \bar{\mathbf{x}}_2^h \right) \left(\bar{\mathbf{x}}_1^h - \bar{\mathbf{x}}_2^h \right)' / n^h$ comme une matrice des covariances inter-classes.
Supposer que l'erreur de Bayes des mélanges est homogène, revient donc dans l'étape M de l'algorithme EM, à estimer (i) le centre de chaque classe non plus comme la moyenne empirique attendue, mais comme un barycentre des moyennes attendues et (ii) la matrice des covariances commune aux classes non plus par la seule matrice intra-classes $\boldsymbol{P}^h$, mais par $\boldsymbol{P}^h$ complétée de la matrice inter-classes $\boldsymbol{Q}^h$.

5.2.2 Application à des données simulées

Supposons que l'erreur de Bayes des différents mélanges est homogène dans le vrai modèle. Puisque les estimateurs du maximum de vraisemblance sont convergents, il est asymptotiquement équivalent de supposer ou de ne pas supposer que l'erreur de Bayes est homogène dans le modèle inféré. Donc cette contrainte n'a lieu d'être spécifiée que pour des échantillons de taille finie. Elle permet alors en réduisant la taille du paramètre, d'accélérer la convergence des estimateurs. On peut se demander (i) en deçà de quelle taille d'échantillon il est utile de préciser que l'erreur de Bayes est homogène et (ii) si

cette taille d'échantillon dépend du chevauchement des classes dans le vrai modèle.

Nous allons apporter une réponse empirique à ces deux questions en étudiant l'effet de l'hypothèse (5.5) sur la vitesse de convergence des estimateurs, pour un chevauchement variable des composantes dans le vrai modèle.

Le vrai modèle consiste en deux mélanges gaussiens ($H = 2$) homoscédastiques d'ordre deux ($K = 2$), dont les composantes ont le même poids et dont l'erreur de Bayes est homogène (en dimension $d = 2$). On note $\boldsymbol{\psi}^0$ son paramètre.
Le modèle inféré consiste en deux mélanges de même nature que dans le vrai modèle (homoscédastiques d'ordre deux, aux composantes de même poids) dont on suppose soit que l'erreur de Bayes est homogène (on note alors $\boldsymbol{\psi}_{\Delta\epsilon}$ son paramètre), soit au contraire que l'erreur de Bayes est libre (son paramètre est noté $\boldsymbol{\psi}$).
La taille $n_k^h \in \{10, 20, 50, 100, 200\}$ des données conditionnelles étant fixée, deux cents échantillons sont générés, sur lesquels on infère successivement les modèles de paramètre $\boldsymbol{\psi}$ et $\boldsymbol{\psi}_{\Delta\epsilon}$. Pour une erreur de Bayes variable dans le vrai modèle, les tableaux 5.1 et 5.2 donnent les valeurs moyennes de BIC, du taux d'erreur (τ) et les écarts-types relatifs à ces moyennes (entre parenthèses).

n_k^h	BIC			τ (erreur apparente)		
	$\boldsymbol{\psi}^0$	$\hat{\boldsymbol{\psi}}$	$\hat{\boldsymbol{\psi}}_{\Delta\epsilon}$	$\boldsymbol{\psi}^0$	$\hat{\boldsymbol{\psi}}$	$\hat{\boldsymbol{\psi}}_{\Delta\epsilon}$
10	155.1(5.8)	148.5(6.0)	145.2(6.1)	5.94(4.04)	19.90(11.12)	13.31(8.94)
20	300.2(7.5)	294.7(8.4)	290.0(7.8)	6.27(2.66)	18.29(11.03)	9.72(6.54)
50	729.4(12.8)	725.7(13.6)	720.1(13.2)	5.86(1.67)	13.91(9.79)	6.51(2.43)
100	1444.9(16.5)	1442.2(18.9)	1435.4(16.7)	5.92(1.00)	11.99(9.49)	6.10(1.08)
200	2867.6(25.1)	2868.0(29.4)	2857.8(25.3)	5.91(0.88)	10.92(9.78)	6.01(0.90)

TABLE 5.1: *Un cas de chevauchement faible des composantes :* $\epsilon^1 = \epsilon^2 = 5.90\%$

Quelle que soit la taille d'échantillon et quel que soit le degré de chevauchement des composantes dans le vrai modèle, on observe que : $BIC(\hat{\boldsymbol{\psi}}_{\Delta\epsilon}) < BIC(\hat{\boldsymbol{\psi}}) < BIC(\boldsymbol{\psi}^0)$. Ainsi le modèle inféré en supposant homogène l'erreur de Bayes, est-il systématiquement préféré par BIC au modèle inféré sans cette contrainte. BIC détecte donc l'homogénéité de l'erreur de Bayes comme une contrainte du vrai modèle, même pour une taille d'échantillon modeste.

Le comportement du taux d'erreur est plus sensible que celui de BIC au recouvrement des classes. Lorsque les composantes sont fortement séparées (TAB. 5.1), le taux d'erreur détecte très bien que (5.5) est une contrainte du vrai modèle. D'une part le taux relatif au modèle $\hat{\boldsymbol{\psi}}_{\Delta\epsilon}$ est considérablement meilleur que celui du modèle $\hat{\boldsymbol{\psi}}$, même pour une taille réduite de la donnée. D'autre part, pour $n_k^h = 200$ le taux correspondant à $\hat{\boldsymbol{\psi}}_{\Delta\epsilon}$ est quasiment à sa limite asymptotique. (Pour la même taille d'échantillon $\tau(\hat{\boldsymbol{\psi}})$ est encore loin de sa limite asymptotique.)

n_k^h	BIC			τ (erreur apparente)		
	$\boldsymbol{\psi}^0$	$\hat{\boldsymbol{\psi}}$	$\hat{\boldsymbol{\psi}}_{\Delta\epsilon}$	$\boldsymbol{\psi}^0$	$\hat{\boldsymbol{\psi}}$	$\hat{\boldsymbol{\psi}}_{\Delta\epsilon}$
10	150.0(6.1)	142.0(6.8)	139.6(6.8)	26.70(6.45)	34.35(6.96)	33.72(7.00)
20	288.9(8.5)	281.5(9.0)	279.0(8.9)	26.92(5.18)	35.16(6.47)	33.71(6.61)
50	702.8(15.8)	695.3(16.2)	692.4(16.2)	27.01(3.04)	36.67(5.83)	34.26(6.03)
100	1391.0(20.0)	1384.4(19.9)	1381.1(20.0)	27.10(2.14)	36.27(5.57)	34.04(5.49)
200	2760.5(26.0)	2753.9(25.9)	2750.4(25.9)	27.07(1.68)	36.25(5.69)	33.66(5.35)

TABLE 5.2: *Un cas de chevauchement fort des composantes :* $\epsilon^1 = \epsilon^2 = 27.21\%$

Lorsque les composantes des mélanges se chevauchent fortement (TAB. 5.2), le taux d'erreur lié à $\hat{\boldsymbol{\psi}}_{\Delta\epsilon}$ est meilleur que celui de $\hat{\boldsymbol{\psi}}$; mais leur différence est moins importante que dans le cas précédent. D'autre part la convergence de $\tau(\hat{\boldsymbol{\psi}}_{\Delta\epsilon})$ comme celle de $\tau(\hat{\boldsymbol{\psi}})$, semble être soit plus lente soit plus erratique que dans TAB. 5.1. Cette observation peut s'interpréter ainsi. Le nombre des extrema locaux de la vraisemblance, augmente avec le chevauchement des composantes. Il est donc probable que la convergence de $\tau(\hat{\boldsymbol{\psi}})$ (resp. de $\tau(\hat{\boldsymbol{\psi}}_{\Delta\epsilon})$) soit parasitée par l'estimation dans certains cas, d'extrema locaux et non globaux. L'homogénéité de l'erreur de Bayes imposée dans l'estimation de $\hat{\boldsymbol{\psi}}_{\Delta\epsilon}$, diminuerait donc le nombre des extrema locaux de la vraisemblance. Cela explique que la convergence de $\tau(\hat{\boldsymbol{\psi}}_{\Delta\epsilon})$, bien que lente, soit moins erratique que celle de $\tau(\hat{\boldsymbol{\psi}})$.

5.3 Egalisation de l'entropie globale des classes

Supposer que l'erreur de Bayes est homogène conduit en pratique et dans beaucoup de cas (pour des mélanges hétéroscédastiques par exemple ou pour des mélanges d'ordre supérieur à deux), à des difficultés numériques rendant impossible l'inférence du modèle. Nous proposons dans cette section de traduire le chevauchement similaire des classes en supposant comme homogène non plus l'erreur de Bayes, mais l'entropie globale des composantes. Nous définirons un algorithme nommé $\tilde{\text{E}}$M, spécifique à cette contrainte. Il s'agit d'un algorithme général qui dépasse le cadre gaussien et permet d'inférer le paramètre d'un grand nombre de mélanges dont les composantes se recouvrent de façon similaire.

On dispose toujours de H échantillons $\mathbf{x}^h = (\mathbf{x}_i^h; i = 1, \ldots, n^h)$ $(h = 1, \ldots, H)$ de $\mathbb{R}^d$, à partitionner en K groupes chacun, et l'on suppose à nouveau que les données de chaque échantillon $\mathbf{x}^h$ proviennent du mélange :

$$f^h(\boldsymbol{x}; \boldsymbol{\psi}^h) = \sum_{k=1}^{K} \pi_k^h f_k^h(\boldsymbol{x}; \boldsymbol{\alpha}_k^h); \boldsymbol{x} \in \mathbb{R}^d. \tag{5.9}$$

5.3.1 Un algorithme ad-hoc : ẼM

Interprétation de l'inférence simultanée du paramètre des mélanges

D'après l'interprétation de l'algorithme EM basée sur la formule d'Hathaway (voir Section 1.2.3), l'inférence du paramètre $\boldsymbol{\psi}^h = \{(\pi_k^h, \boldsymbol{\alpha}_k^h); k = 1, \dots, K\}$ de chaque mélange par EM, revient à optimiser alternativement un critère portant à la fois sur la vraisemblance de $\boldsymbol{\psi}^h$ et sur l'entropie d'une partition floue $\boldsymbol{w}^h = (w_{i,k}^h)_{\substack{i=1,\dots,n^h \\ k=1,\dots,K}}$ de la donnée $\mathbf{x}^h$. Mais $\boldsymbol{\psi} = \left(\psi^h\right)_{h=1,\dots,H}$, le paramètre du modèle, peut être estimé de façon équivalente par un algorithme EM unique. L' étape E consiste alors, à partir de la valeur courante du paramètre, à estimer la partition floue $\boldsymbol{w} = (\boldsymbol{w}^h)_{h=1,\dots,H}$ de la donnée qui optimise le critère :

$$C(\boldsymbol{\psi}, \boldsymbol{w}) = \underbrace{\sum_{h=1}^{H} \sum_{i=1}^{n^h} \ln \left[\sum_{k=1}^{K} \pi_k^h f_k^h(\mathbf{x}_i^h; \boldsymbol{\alpha}_k^h) \right]}_{L(\boldsymbol{\psi})} - \underbrace{\sum_{h=1}^{H} \sum_{i=1}^{n^h} \sum_{k=1}^{K} \left(w_{i,k}^h \ln w_{i,k}^k - w_{i,k}^h \ln t_{i,k}^k \right)}_{KL(\boldsymbol{w},\boldsymbol{t})}. \tag{5.10}$$

Les coefficients $t_{i,k}^h$ désignent les probabilités conditionnelles induites par la valeur courante du paramètre. Le second terme de (5.10) peut être vu comme l'écart de Kullback entre les partitions probabilistes $\boldsymbol{t}$ et $\boldsymbol{w}$. Son premier terme $L(\boldsymbol{\psi})$ est la log-vraisemblance de $\boldsymbol{\psi}$. Puisqu'il est indépendant de $\boldsymbol{w}$, l'optimum de (5.10) est atteint lorsque les partitions floues $\boldsymbol{w}$ et $\boldsymbol{t}$ coïncident.

Mais le critère (5.10) peut également s'écrire :

$$C(\boldsymbol{\psi}, \boldsymbol{w}) = \underbrace{\sum_{h=1}^{H} \sum_{i=1}^{n^h} \sum_{k=1}^{K} w_{i,k}^h \ln \left[\pi_k^h f_k^h(\mathbf{x}_i^h; \boldsymbol{\alpha}_k^h) \right]}_{LC(\boldsymbol{\psi},\boldsymbol{w})} \underbrace{- \sum_{h=1}^{H} \sum_{i=1}^{n^h} \sum_{k=1}^{K} w_{i,k}^h \ln w_{i,k}^h}_{E(\boldsymbol{w})}. \tag{5.11}$$

Le premier terme de (5.11) coïncide avec la log-vraisemblance de $\boldsymbol{\psi}$ complétée de la partition floue $\boldsymbol{w}$ et le second correspond à l'entropie empirique de la partition $\boldsymbol{w}$ (voir Section 1.2.3). Optimiser (5.11) avec $\boldsymbol{w} = \boldsymbol{t}$, revient donc à déterminer l'optimum de la log-vraisemblance attendue du paramètre $\boldsymbol{\psi}$; c'est l'étape M d'EM.

Egalisation de l'entropie globale des composantes

L'entropie globale des composantes de (5.9) se définit comme la divergence de Kullback totale entre les lois conditionnelles pondérées et la loi mélange :

$$\mathscr{E}^h = -\sum_{k=1}^{K} KL(\pi_k^h f_k^h, f^h). \tag{5.12}$$

Lorsque le chevauchement des composantes est fort, (5.12) est proche de $\ln K$. Des composantes bien séparées se manifestent au contraire par une entropie globale (5.12)

proche de 0.
Aussi peut-on traduire l'imbrication similaire des composantes des différents mélanges, en supposant que leur entropie globale est homogène :

$$\mathscr{E}^1 = \cdots = \mathscr{E}^H. \tag{5.13}$$

Mais l'inférence de $\boldsymbol{\psi}$ sous la contrainte (5.13) conduit à des difficultés numériques plus grandes encore que dans la section précédente. (5.13) est en effet hautement non linéaire et l'optimisation (par rapport à $\boldsymbol{\psi}$) de (5.11) sous la contrainte (5.13) n'est pas réalisable.

Nous proposons donc de mesurer la séparation des classes dans le mélange h non pas par l'entropie théorique de ses composantes (5.12) mais par l'entropie empirique normalisée :

$$e(\boldsymbol{t}^h) = -(1/n^h) \sum_{k=1}^{K} \sum_{i=1}^{n^h} t_{i,k}^h \ln t_{i,k}^h. \tag{5.14}$$

L'ordre de grandeur de la statistique $e(\boldsymbol{t}^h)$ est celui du paramètre $\mathscr{E}^h$: elle est d'autant plus proche de 0 (resp. de $\ln K$) que les composantes du mélange h sont séparées (resp. se chevauchent). On peut d'ailleurs montrer que $e(\boldsymbol{t}^h)$ converge en loi vers $\mathscr{E}^h$.

Nous traduirons le recouvrement identique des classes en supposant que l'entropie empirique globale des composantes est homogène d'un mélange à l'autre, contrainte que l'on note :

$$\gamma(\boldsymbol{t}) \ : \ e(\boldsymbol{t}^1) = \cdots = e(\boldsymbol{t}^H). \tag{5.15}$$

Il est important de remarquer que la contrainte (5.15) imposée à l'étape M d'EM pour optimiser (5.11), porte sur le paramètre $\boldsymbol{\psi}$ mais qu'elle induit une restriction de l'espace de la partition floue. Chaque partition floue $\boldsymbol{t}$ relative au paramètre $\boldsymbol{\psi}$ estimé à l'étape M, vérifie (5.15). Par conséquent, pour chaque partition floue $\boldsymbol{w}$ estimée à l'étape E suivante, l'entropie empirique normalisée des populations est homogène. Les partitions floues $\boldsymbol{w}$ successivement estimées dans l'étape E, vérifient la contrainte $\gamma(\boldsymbol{w})$.

$\tilde{\text{E}}$M : un algorithme spécifique pour les modèles d'entropie homogène

L'optimum de (5.11) (par rapport à $\boldsymbol{\psi}$) est aussi difficile à obtenir sous la contrainte (5.15) que sous la contrainte (5.13). Aussi doit-on renoncer à un algorithme EM qui permettrait d'inférer le paramètre $\boldsymbol{\psi}$ en supposant en un sens strict, que l'entropie des composantes est homogène.
Nous proposons ici un algorithme de substitution $\tilde{\text{E}}$M, dérivé de l'algorithme EM. Ses estimateurs n'optimisent pas (pas tout à fait) la vraisemblance de $\boldsymbol{\psi}$ sous la contrainte (5.15). Mais ils convergent vers les estimateurs du maximum de vraisemblance lorsque (5.15) est une contrainte du vrai modèle.

Lorsqu'on impose au paramètre $\boldsymbol{\psi}$ la contrainte $\gamma(\boldsymbol{t})$ à l'étape M d'EM, la partition floue $\boldsymbol{w}$ que l'on estime à l'étape E suivante, vérifie la contrainte $\gamma(\boldsymbol{w})$. L'algorithme

$\tilde{\text{E}}$M consiste à relaxer $\gamma(\boldsymbol{t})$ à l'étape M, mais à maintenir $\gamma(\boldsymbol{w})$ à l'étape E.

Ainsi à partir d'une valeur initiale du paramètre $\boldsymbol{\psi}$, l'algorithme $\tilde{\text{E}}$M alterne les deux étapes suivantes.

- *Etape $\tilde{E}$.* On détermine la partition floue $\boldsymbol{w}$ qui optimise (5.10) sous la contrainte $\gamma(\boldsymbol{w})$. Cela revient à déterminer parmi les partitions floues $\boldsymbol{w}$ vérifiant $\gamma(\boldsymbol{w})$ (homogénéité de l'entropie globale normalisée), la plus proche au sens de l'écart de Kullback, de la partition $\boldsymbol{t}$ induite par la valeur courante du paramètre.

- *Etape M.* On optimise (5.11) par rapport à $\boldsymbol{\psi}$, en prenant pour $\boldsymbol{w}$ la partition floue déterminée à l'étape $\tilde{\text{E}}$.

On notera que l'étape M d'$\tilde{\text{E}}$M est strictement identique à celle d'un algorithme EM classique. Elle consiste à déterminer le paramètre qui rend optimale la log-vraisemblance complétée attendue. Seule l'étape $\tilde{\text{E}}$ distingue les deux algorithmes en imposant une contrainte à la partition floue recherchée.
Dans le cas de deux classes ($K = 2$) par exemple, la partition $\boldsymbol{w}$ déterminée à l'étape $\tilde{\text{E}}$, vérifie :

$$w_{i,k}^h = 1 \Big/ \left(1 + (t_{i,1}^h / t_{i,2}^h)^{(-1)^k/(1+\lambda^h/n^h)}\right). \tag{5.16}$$

Les coefficients $t_{i,k}^h$ désignent toujours les probabilités conditionnelles induites par le paramètre en cours. Les coefficients λ^h sont des multiplicateurs de Lagrange dont la valeur est approchée numériquement grâce à la contrainte $\gamma(\boldsymbol{w})$.

L'algorithme $\tilde{\text{E}}$M ainsi défini est générique : aucune hypothèse n'a été faite sur la spécificité du mélange (5.9). Il peut être implémenté dans n'importe quel type de mélange dès lors que les estimateurs qui permettent d'optimiser la log-vraisemblance complétée, sont connus.

Lorsque la contrainte (5.13) est vraie ou lorsque (5.15) est vérifiée asypmtotiquement dans le vrai modèle, les algorithmes EM avec ou sans la contrainte (5.15) et l'algorithme $\tilde{\text{E}}$M sont équivalents asymptotiquement.
Si au contraire (5.15) n'est pas vérifiée de façon asymptotique dans le vrai modèle, on peut s'attendre à ce que la vraisemblance du paramètre soit suffisamment petite pour qu'un critère de choix de modèle comme BIC (basé sur la vraisemblance) écarte le modèle inféré par $\tilde{\text{E}}$M, et lui préfère un modèle estimé par un algorithme EM traditionnel.

5.3.2 Illustrations

Application à des données simulées

Sur des données simulées, nous mettons en évidence les propriétés de convergence de l'algorithme $\tilde{\text{E}}$M, énoncées précédemment. Les résultats expérimentaux de ce para-

graphe ont été présentés dans [77], ainsi que ceux du paragraphe suivant.
Les données consistent en deux échantillons de même taille $n^1 = n^2 \in \{40, 100\}$, générés chacun par un mélange gaussien hétéroscédastique dont les composantes ont le même poids. L'entropie globale des composantes est homogène et les composantes de chaque mélange se chevauchent de façon relativement importante (leur erreur de Bayes vaut 24.33 dans les deux cas). On note $\boldsymbol{\psi}^0$ le paramètre du vrai modèle ; $\hat{\boldsymbol{\psi}}$ désigne le paramètre du modèle estimé (deux mélanges gaussiens hétéroscédastiques aux composantes de même poids) par l'algorithme EM, et $\tilde{\boldsymbol{\psi}}$ le paramètre du modèle estimé par $\tilde{\text{E}}$M. Le tableau 5.3 indique la valeur moyenne ainsi que l'écart-type de BIC et de τ (taux d'erreur de classement) relatifs aux paramètres précédents, et calculés par simulation de deux cents échantillons.

On observe d'abord qu'$\tilde{\text{E}}$M accélère davantage la convergence de BIC que celle de τ : le modèle inféré par $\tilde{\text{E}}$M donne une valeur de BIC meilleure que celle obtenue par EM et cela, même pour une petite taille des échantillons ($n^h = 40$). Il faut atteindre une taille plus importante des échantillons ($n^h = 100$) pour que la valeur de τ obtenue par $\tilde{\text{E}}$M soit à son tour meilleure que celle obtenue par EM.
Mais la taille des échantillons à partir de laquelle $\tilde{\text{E}}$M identifie clairement (5.15) comme une contrainte du vrai modèle est relativement modeste puisque dès $n^h = 100$, les valeurs de BIC et τ relatives à $\tilde{\text{E}}$M sont, dans les deux cas, meilleures que celles obtenues par EM.

n^h	$\tau(\boldsymbol{\psi}^0)$	$BIC(\hat{\boldsymbol{\psi}})$	$\tau(\hat{\boldsymbol{\psi}})$	$BIC(\tilde{\boldsymbol{\psi}})$	$\tau(\tilde{\boldsymbol{\psi}})$
40	25.35 (4.61)	342.0 (8.74)	31.32 (6.81)	340.1 (8.74)	31.57 (7.09)
100	24.34 (2.90)	813.7 (15.57)	28.79 (5.2)	811.4 (15.56)	28.22 (4.92)

TABLE 5.3: *Valeur moyenne et (écart-type) de BIC et du taux d'erreur de classement, obtenus sur des données simulées*

Application à des données réelles

Nous considérons ici les deux échantillons de *Calonectris diomedea borealis* ($n^1 = 206$ oiseaux dont 45% de femelles) et *Calonectris diomedea diomedea* ($n^2 = 38$ oiseaux dont 58% de femelles) introduits à la section 4.1. De l'avis des biologistes à l'origine de ces données, il est plausible que les mâles et les femelles décrits par les cinq variables biométriques disponibles (hauteur du bec, longueur du tarse, etc.) soient mélangés de façon similaire dans les deux espèces, ce que corrobore, nous allons le voir, l'inférence d'un modèle par l'algorithme $\tilde{\text{E}}$M.

Comme à la section 4.1, nous envisagerons plusieurs types de contraintes sur les mélanges gaussiens qui modélisent les oiseaux de chaque espèce. Nous supposerons alternativement que leurs composantes sont homoscédastiques ($\boldsymbol{\Sigma}_1$) ou hétéroscédastiques ($\boldsymbol{\Sigma}_k$), et que leur poids est homogène (π_1) ou libre (π_k). Nous combinerons par ailleurs les quatre modèles intra-populations ainsi obtenus avec chacune des hypothèses suivantes : (i) les deux sous-espèces d'oiseaux proviennent de la même population, (ii) les

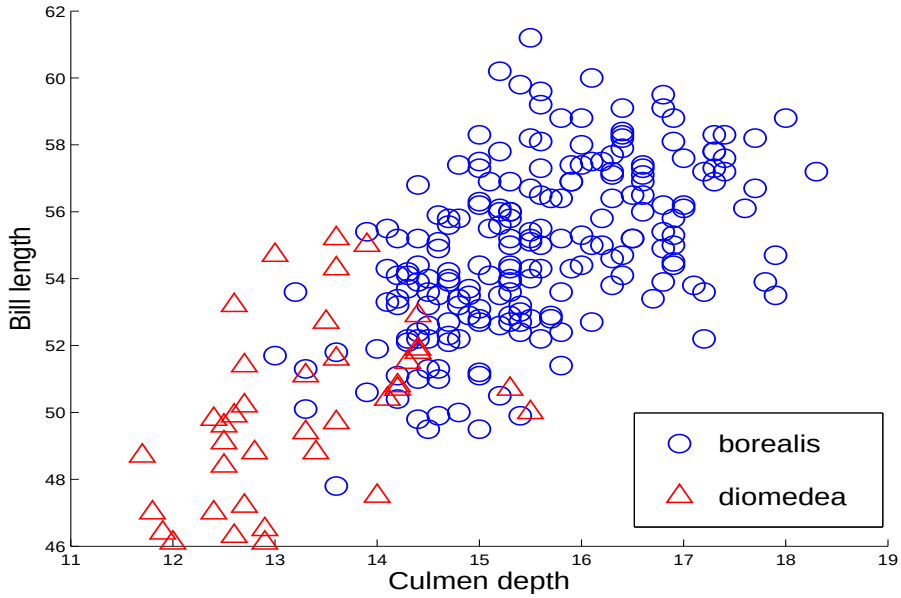
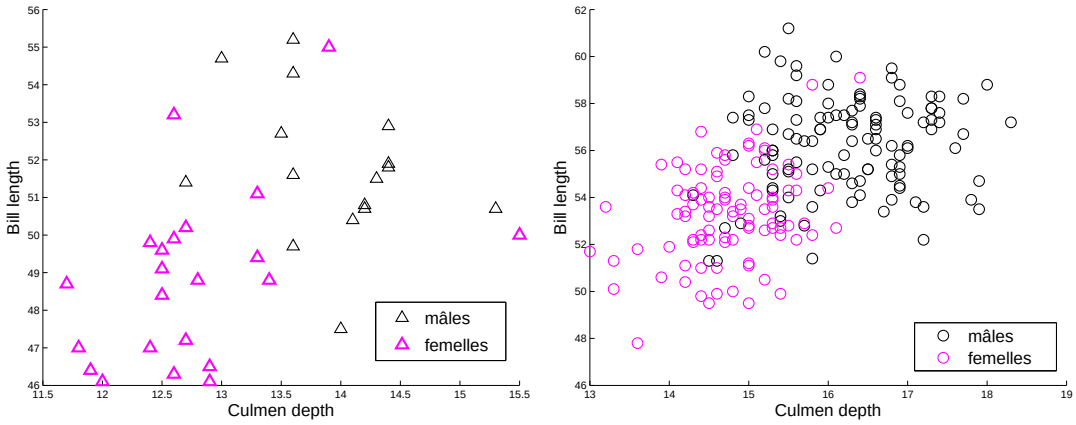


FIGURE 5.2: *Deux sous-espèces de Calonectris diomedea (Puffins cendrés)*



(a) La partition recherchée des *diomedea* (b) La partition recherchée des *borealis*

FIGURE 5.3: *Un chevauchement similaire des mâles et des femelles dans les deux populations d'oiseaux ?*

sous-espèces proviennent de populations distinctes, liées par un chevauchement homogène de leurs classes (le paramètre du modèle est estimé par $\tilde{\text{EM}}$), (iii) les sous-espèces proviennent de populations distinctes, sans lien entre elles (le paramètre du modèle est estimé par EM).

Le tableau 5.4 indique la valeur de BIC et le taux d'erreur de classement des puffins, obtenus par les différents modèles intra-populations, sous chacune des hypothèses (i) à (iii).

On observe (sans surprise vu FIG. 5.2) que BIC rejette clairement l'hypothèse (i) selon laquelle *borealis* et *diomedea* proviendraient de la même population. Si l'on suppose les échantillons issus de populations distinctes (hypothèses (ii) et (iii)), les deux meilleurs modèles (dont la valeur de BIC est proche : 3045.2 vs. 3045.7) sont inférés selon $\tilde{\text{EM}}$. Cela corrobore l'hypothèse selon laquelle le chevauchement des mâles et des

modèle	$(\pi_1, \mathbf{\Sigma}_1)$	$(\pi_1, \mathbf{\Sigma}_k)$	$(\pi_k, \mathbf{\Sigma}_1)$	$(\pi_k, \mathbf{\Sigma}_k)$
(i)	3119.5 (45.49)	3123.1 (49.18)	3094.0 (43.03)	3121.1 (37.70)
(ii)	**3045.7** (**11.48**)	3096.4 (17.62)	**3045.2** (41.39)	3099.3 (23.36)
(iii)	3047.5 (12.30)	3100.9 (45.08)	3048.5 (36.07)	3102.5 (31.56)

TABLE 5.4: *BIC et (% d'erreur de classement) obtenus pour la classification des borealis et des diomedea.*

femelles est homogène parmi *borealis* et *diomedea*. On remarque d'autre part que l'un des deux meilleurs modèles, correspond également au meilleur taux d'erreur de classement (11.48 %).
Ainsi la classification simultanée basée sur un lien entropique global entre populations améliore, dans ce nouvel exemple ornithologique, les deux formes communes de classification représentées par les hypothèses (i) et (iii).

De façon générale et en dehors du contexte de la classification des puffins, l'hypothèse (5.13) ne diminue que de $H-1$ unités la dimension du paramètre. Ainsi par exemple, dans le cas de deux groupes, supposer que l'entropie des composantes est homogène réduit d'une unité seulement la dimension du paramètre (quelle que soit la dimension de l'espace). On peut donc craindre le manque de parcimonie du paramètre inféré par $\tilde{\text{E}}$M (et de façon générale, le manque de parcimonie induit par l'hypothèse (5.13)). Mais l'exemple des puffins montre que même dans un espace de relativement grande dimension (la taille du paramètre dans le modèle (ii) & $(\pi_1, \mathbf{\Sigma_1})$ est $\nu = 35$) l'information qu'apporte $\tilde{\text{E}}$M n'est pas négligeable puisqu'elle permet d'améliorer BIC ainsi que l'erreur de classement.

Combinaison d'un lien global et conditionnel entre populations

A la section 4.1 le lien entre populations est conditionnel : il s'agit d'une contrainte sur le paramètre des composantes des différents mélanges. Jusqu'ici, à la section 5.3, le lien n'est pas conditionnel mais global : la contrainte porte sur l'entropie totale des composantes de chaque mélange. Ces deux types de contraintes, loin de s'opposer, peuvent au contraire se combiner.

Considérons à nouveau la classification de *borealis* et *diomedea* abordée au paragraphe précédent. Nous pouvons envisager par exemple (comme le faisaient certains modèles de Section 4.1) que la matrice des covariances est homogène ($\mathbf{\Sigma}^1$) ou libre ($\mathbf{\Sigma}^h$) parmi les mâles (resp. parmi les femelles) des deux espèces. Nous pouvons supposer également que la proportion de mâles et de femelles est identique (π^1) dans les deux espèces ou libre (π^h). Qu'elles portent sur les matrices de covariances ou le poids des composantes, les hypothèses précédentes sont de type conditionnel. On peut leur associer une contrainte globale en supposant que l'entropie des composantes est homogène et en estimant le paramètre du modèle par $\tilde{\text{E}}$M au lieu d'EM. Le tableau 5.5 compare les valeurs de BIC et du taux d'erreur de classement obtenus par combinaison des différentes

hypothèses intra et inter-populations, lorsque le paramètre est inféré par ẼM (entropie homogène) et par EM (entropie libre). Il existe une contradiction évidente à supposer les proportions homogènes dans chaque mélange, et libres d'un mélange à l'autre. Aussi les modèles combinant les hypothèses π_1 et π^h, ne sont-ils pas inférés dans TAB. 5.5.

algorithme	modèle	$(\pi_1, \mathbf{\Sigma}_1)$	$(\pi_1, \mathbf{\Sigma}_k)$	$(\pi_k, \mathbf{\Sigma}_1)$	$(\pi_k, \mathbf{\Sigma}_k)$
EM	$(\pi^1, \mathbf{\Sigma}^1)$	3017.8 (13.11)	3048.1 (45.90)	3019.4 (40.57)	3028.2 (46.31)
	$(\pi^1, \mathbf{\Sigma}^h)$	3047.5 (12.30)	3100.9 (45.08)	3043.1 (38.52)	3086.8 (39.75)
	$(\pi^h, \mathbf{\Sigma}^1)$	×	×	3019.0 (39.75)	3045.7 (48.77)
	$(\pi^h, \mathbf{\Sigma}^h)$	×	×	3048.5 (36.07)	4102.5 (31.56)
ẼM	$(\pi^1, \mathbf{\Sigma}^1)$	**3015.0** (13.11)	3042.0 (12.70)	**3014.2** (38.93)	3044.5 (25.40)
	$(\pi^1, \mathbf{\Sigma}^h)$	3045.7 (**11.48**)	3096.4 (17.62)	3043.4 (38.52)	3096.6 (23.77)
	$(\pi^h, \mathbf{\Sigma}^1)$	×	×	3016.7 (39.75)	3050.1 (42.21)
	$(\pi^h, \mathbf{\Sigma}^h)$	×	×	3045.2 (41.39)	3099.3 (23.36)

TABLE 5.5: *BIC et (% d'erreur de classement) obtenus dans la classification des puffins, par combinaison de contraintes intra-population et de liens interpopulations de plusieurs types : conditionnel et global.*

On observe que l'algorithme ẼM infère les deux meilleurs modèles ($BIC = 3014.2$ vs. 3015.0) et que l'erreur apparente de ces deux modèles ($\tau = 38.93$ vs. 13.11) est au moins aussi bonne que l'erreur correspondante, relative à EM.

5.4 Bilan et perspectives

Il arrive que des échantillons soient décrits par des variables dont le pouvoir discriminant est proche. Il peut être judicieux alors, de les modéliser par des mélanges dont les composantes se chevauchent de façon semblable. Dans ce chapitre nous avons traduit cette idée sous deux formes : en supposant d'abord que l'erreur de Bayes des mélanges est identique (Section 5.2) puis que l'entropie globale des composantes est homogène (Section 5.3).
Sur des données simulées nous avons montré que l'hypothèse d'un chevauchement identique des composantes, permettait d'accélerer la convergence des estimateurs du maximum de vraisemblance (lorsque cette hypothèse est une contrainte du vrai modèle).
Sur des données réelles ornithologiques, nous avons montré que même si elle n'est pas exactement une contrainte du vrai modèle, l'homogénéité du chevauchement des classes pouvait permettre d'améliorer le modèle inféré et la qualité du classifieur.
L'hypothèse d'un chevauchement homogène des classes constitue une containte globale entre les mélanges, qui peut être combinée avec les modèles de liens conditionnels affines

envisagés à la section 4.1. Mais considérée en tant que telle, cette contrainte permet également d'étendre le champ de la classification simultanée à des échantillons décrits par des variables n'ayant pas la même signification. Des échantillons peuvent en effet être décrits par des variables différentes, mais sémantiquement proches et dont le pouvoir discriminant est similaire.

Nous avons présenté un algorithme ẼM qui permet d'inférer simultanément le paramètre de plusieurs mélanges en supposant homogène l'entropie globale des composantes. Cet algorithme repose sur l'interprétation d'EM selon la formule d'Hathaway. Il consiste à relaxer l'homogénéité de l'entropie en tant que contrainte sur le paramètre et à maintenir cette sujétion sur la partition probabiliste de la donnée. Dans ce travail, l'usage d'ẼM se restreint aux mélanges gaussiens. Mais il s'agit d'un algorithme générique, très simple à mettre en oeuvre (il ne consiste qu'en une modification de l'étape E d'EM), que l'on peut implémenter dans un grand nombre de mélanges. Nous précisons à ce propos que dans tous les cas, les estimateurs vers lesquels converge ẼM sont asymptotiquement ceux du maximum de vraisemblance.

ẼM est né de l'impossibilité à optimiser (même numériquement) le critère de vraisemblance (5.11) sous la contrainte d'entropie homogène (5.15). Mais d'une part il existe d'autres mesures de la séparation des classes, que l'entropie selon (5.14). En s'inspirant de J.C. Bezdek ([11]) par exemple, on peut mesurer le chevauchement des classes dans la population h par :

$$PC(\boldsymbol{t}^h) = (1/n^h) \sum_{k=1}^{K} \sum_{i=1}^{n^h} \left(t_{i,k}^h\right)^2. \tag{5.17}$$

D'autre part, peut-être pourrait-on modifier le critère (5.11) de sorte (i) que l'inférence soit possible sous la contrainte d'entropie homogène et (ii) que le paramètre estimé soit proche du maximum de la vraisemblance.
Nous proposons par exemple de remplacer l'entropie (5.14) par (5.17), et le critère (5.10) par :

$$C(\boldsymbol{\psi}, \boldsymbol{w}) = \underbrace{\sum_{h=1}^{H} \sum_{i=1}^{n^h} \ln \left[\sum_{k=1}^{K} \pi_k^h f_k^h(\mathbf{x}_i^h; \boldsymbol{\alpha}_k^h) \right]}_{L(\boldsymbol{\psi})} - \underbrace{\sum_{h=1}^{H} \sum_{i=1}^{n^h} \sum_{k=1}^{K} \left(w_{i,k}^k - t_{i,k}^k\right)^2}_{\|\boldsymbol{w}-\boldsymbol{t}\|^2}. \tag{5.18}$$

La partition $\boldsymbol{w}$ déterminée à l'étape E, s'interpréterait alors comme la plus proche de $\boldsymbol{t}$, non plus au sens de l'écart de Kullback $KL(\boldsymbol{w}, \boldsymbol{t})$, mais au sens de la distance $\|\boldsymbol{w} - \boldsymbol{t}\|$. Mais d'une part le critère (5.18) optimisé par rapport au paramètre, ne serait plus à proprement parler la vraisemblance complétée, mais :

$$C(\boldsymbol{\psi}, \boldsymbol{w}) = LC(\boldsymbol{\psi}, \boldsymbol{w}) + E(\boldsymbol{w}) + KL(\boldsymbol{w}, \boldsymbol{t}) - \|\boldsymbol{w} - \boldsymbol{t}\|^2. \tag{5.19}$$

D'autre part on ignore si l'estimateur vers lequel converge l'algorithme qui optimise alternativement (5.18) et (5.19), est proche de celui du maximum de vraisemblance.

Formaliser un recouvrement identique des classes en supposant homogène l'erreur de Bayes, est possible dans le cas de mélanges gaussiens homoscédastiques, mais cela

conduit dans le cas hétéroscédastique à des difficultés calculatoires infranchissables. Dans l'espoir de pouvoir imposer l'homogénéité de l'erreur de classement dans le cas hétéroscédastique comme homoscédastique, on peut remplacer la mesure de séparation des classes par l'erreur empirique normalisée :

$$\epsilon(\boldsymbol{t}^h) = 1 - (1/n^h) \sum_{i=1}^{n^h} \max_{\{k\}} t_{i,k}^h, \tag{5.20}$$

et tenter comme pour $\tilde{\text{E}}$M, de transférer la contrainte d'erreur homogène, du paramètre à la partition probabiliste. Mais estimer la partition floue $\boldsymbol{w}$ la plus proche de $\boldsymbol{t}$ au sens de l'écart de Kullback en supposant $\epsilon(\boldsymbol{w}^h)$ homogène, conduit à des difficultés numériques auxquelles nous n'avons pas de solution.
Ainsi les modèles d'erreur homogène semlent pour le moment confinés au cas gaussien homoscédastique.

Appendices

Annexe B

Extension d'un résultat de probabilités

Théorème 1 (Extension d'un résultat de [35])**.** *X et Y sont deux variables aléatoires réelles, absolument continues, symétriques, dont le support est $\mathbb{R}$.*
S'il existe une application affine de $\mathbb{R}$ dans $\mathbb{R}$ qui transforme stochastiquement X en Y, alors il en existe forcément une seconde.
Dans ce cas, ces deux applications affines sont les seules applications de classe $\mathscr{C}^1$ de $\mathbb{R}$ dans $\mathbb{R}$ qui transforment stochastiquement X en Y.

Preuve. Supposons qu'il existe un couple $(a, b) \in \mathbb{R}^* \times \mathbb{R}$ tel que $Y \sim aX + b$. X étant symétrique il existe un réel ω tel que $(X - \omega)$ et $(\omega - X)$ sont identiquement distribuées. On en déduit que $(aX + b)$ et $(-aX + 2a\omega + b)$ sont identiquement distribuées.

Soit ϕ une application de classe $\mathscr{C}^1$ de $\mathbb{R}$ dans $\mathbb{R}$ telle que $Y \sim \phi(X)$. Puisque Y est absolument continue, ϕ est strictement monotone. En effet si ϕ ne l'était pas, ϕ' s'annulerait en un point c et la densité de Y serait infinie en $\phi(c)$.
Par ailleurs comme le support de Y est $\mathbb{R}$, ϕ est surjective de $\mathbb{R}$ dans $\mathbb{R}$. Donc ϕ est une bijection de classe $\mathscr{C}^1$ de $\mathbb{R}$ dans $\mathbb{R}$.

Supposons que ϕ est strictement croissante sur $\mathbb{R}$ et notons F_Y la fonction de répartition de Y. Pour tout réel γ, $[X \leq \gamma]$ équivaut à $[\phi(X) \leq \phi(\gamma)]$ et à $[aX + b \leq a\gamma + b]$. Comme $\phi(X)$ et $(aX + b)$ sont identiquement distribuées à Y on en déduit $F_Y(\phi(\gamma)) = F_Y(a\gamma + b)$. Et puisque F_Y est une bijection de $\mathbb{R}$ dans $]0; 1[$, $\phi(\gamma) = a\gamma + b$.

Supposons que ϕ est strictement décroissante sur $\mathbb{R}$. Pour tout réel γ, $[X \geq \gamma]$ équivaut à $[\phi(X) \leq \phi(\gamma)]$ et à $[-aX + 2a\omega + b \leq -a\gamma + 2a\omega + b]$. Comme $\phi(X)$ et $(-aX + 2a\omega + b)$ sont identiquement distribuées à Y on en déduit $F_Y(\phi(\gamma)) = F_Y(-a\gamma + 2a\omega + b)$ et donc $\phi(\gamma) = -a\gamma + 2a\omega + b$.

□

Conséquence . *X et Y étant deux variables aléatoires réelles distribuées selon deux lois normales (resp. deux lois de Student de même degré de liberté, deux lois de Cauchy), il existe exactement deux applications de classe $\mathscr{C}^1$ de $\mathbb{R}$ dans $\mathbb{R}$ qui transforment stochastiquement X en Y et elles sont toutes deux affines.*

Preuve. C'est une conséquence immédiate du théorème précédent puisque le groupe affine de $\mathbb{R}$ ($\mathrm{GA}(\mathbb{R})$) agit transitivement sur la famille des lois normales univariées (non dégénérées) (resp. sur la famille des lois de Student de même dégré de liberté, sur la famille des lois de Cauchy).

□

Conclusion générale et perspectives

Mélanges gaussiens parcimonieux d'interprétation statistique vs. modèles géométriques

Les mélanges gaussiens sont fréquemment utilisés pour classifier des données continues hétérogènes, et de façon plus générale pour les modéliser. Les raisons de leur succès sont nombreuses : la loi normale est un modèle répandu, ses propriétés font partie d'une culture statistique élémentaire, on estime son paramètre par maximum de vraisemblance de façon explicite, etc. D'autre part les mélanges gaussiens sont suffisamment souples pour permettre d'approcher un grand nombre de lois, d'autres lois que la loi normale et même des lois dicrètes.

G. Celeux et G. Govaert définissent dans [28], une famille de mélanges gaussiens parcimonieux d'interprétation géométrique, basés sur une décomposition spectrale des matrices de covariances conditionnelles (voir Section 1.1.3). A leur suite C. Bouveyron propose dans [21], des modèles gaussiens d'inspiration similaire, mais spécifiques aux données de grande dimension (voir Section 1.1.4). L'application MIXMOD (`http://www.mixmod.org/`), développée sous la direction de F. Langrognet, modélise et classifie des données quantitatives grâce aux mélanges de G. Celeux et G. Govaert, et des données en grande dimension grâce aux modèles de C. Bouveyron.

Or les modèles gaussiens parcimonieux d'interprétation géométrique souffrent de défauts majeurs. Ils ne sont pas stables par exemple, par projection dans les plans canoniques. La représentation en dimension deux d'un modèle de dimension supérieure est donc incorrecte. D'autre part, les modèles géométriques ne sont pas (en général) invariants à la modification des unités de mesure. En particulier, le modèle inféré n'est pas le même suivant que son paramètre est estimé sur les données réduites ou sur les données brutes.

Nous définissons au chapitre 2, de nouveaux modèles gaussiens dits RTV, dont la parcimonie porte sur des paramètres non plus géométriques, mais statistiques. Il sont stables par projection dans les plans canoniques. Cela assure que n'importe lequel d'entre eux peut être représenté de façon fidèle en dimension réduite. D'autre part, le choix de l'un de ces modèles par un critère de vraisemblance comme AIC, BIC ou ICL, est invariant à la modification des unités de mesure. En particulier, le modèle inféré sur les données et sur les données réduites, est le même. Les modèles RTV remédient ainsi à l'ensemble des inconvénients liés aux mélanges de [28]. Nous conseillons pour cette raison de les substituer dans MIXMOD aux actuels modèles dédiés à la classification de données quantitatives.

Les mélanges de Factor Analyzers : des modèles dédiés à la grande dimension, aux multiples propriétés de stabilité

Dans le domaine spécifique de la grande dimension, ni les modèles RTV ni ceux de

C. Bouveyron ne sont pleinement satisfaisants. Les modèles RTV ne sont pas orientés vers la formalisation du manque d'information dans certaines régions de l'espace. Les modèles de C. Bouveyron, eux, ne possèdent pas les propriétés d'invariance et de stabilité des modèles RTV.
Les mélanges de Factor Analyzers décrits dans [85] et dédiés à la grande dimension, allient tout à la fois les qualités des mélanges RTV et des modèles de [21]. D'une part ils permettent de formaliser le manque d'information dans certaines directions de l'espace. D'autre part ils sont stables par projection dans les plans canoniques. Enfin le choix d'un mélange de Factor Analyzers par un critère de vraisemblance, est invariant à la modification des unités de mesure. Donc, contrairement aux mélanges gaussiens de C. Bouveyron, les mélanges de Factor Analyzers ne sacrifient pas à la spécificité des données de grande dimension, les propriétés d'invariance et de stabilité des modèles RTV. Signalons de plus, que les mélanges de Factor Analyzers sont compatibles avec les contraintes inhérentes aux modèles RTV : on peut supposer les corrélations conditionnelles homogènes sans modifier la structure de Factor Analyzer du modèle. Nous ignorons cependant si l'inférence du paramètre du modèle sous une telle contrainte est réalisable.

Classification simultanée de plusieurs échantillons

Nous proposons dans la seconde partie de ce travail, une méthode novatrice de classification dite simultanée. Cette méthode repose sur un constat : dans de très nombreuses situations, la classification d'un échantillon revient à la classification de plusieurs échantillons, et tire avantage de cette transformation. En s'inspirant de méthodes statistiques dédiées initialement à un seul, puis étendues à plus d'un échantillon, la classification simultanée se propose de classifier de façon conjointe plusieurs échantillons, en supposant que les populations dont ils sont issus recèlent un lien.

Lien entre populations conditionnelles

Lorsque les échantillons sont décrits par des variables de même signification, et que l'on cherche le même nombre de groupes dans chacun d'eux (des groupes dont on donne la même interprétation), il peut être judicieux de supposer une transformation stochastique affine des populations conditionnelles. Nous mettons cela en évidence au chapitre 4, en définissant des modèles de mélanges dont les composantes sont deux à deux liées par une contrainte paramétrique. Nous envisageons d'abord ce type de lien dans le contexte classique de mélanges gaussiens, puis nous l'étendons aux mélanges de Student pour classifier des données susceptibles de comporter du bruit ou des outliers. Nous montrons dans de nombreux cas de données réelles, que ce type de lien (i) améliore de façon évidente la qualité du modèle au regard d'un critère de choix de modèle comme BIC ou ICL, (ii) améliore la qualité du classifieur en terme d'erreur de classement, (iii) permet d'interpréter le lien stochastique comme une évolution des populations conditionnelles.

Lien global entre mélanges

Mais la classification simultanée peut être mise en oeuvre dans un contexte moins drastique. Il arrive que les échantillons soient décrits par des variables qui n'ont pas le même sens, mais dont le pouvoir discriminant est proche. C'est le cas notamment, lorsque les descripteurs sont différents d'un échantillon à l'autre, mais analogues d'un point de vue sémantique. Nous envisageons au chapitre 5, une forme nouvelle de la classification simultanée, dédiée à ce type de contexte et basée sur un chevauchement similaire des classes dans les populations.
Nous supposons d'abord que les échantillons à classifier sont générés par des mélanges gaussiens dont l'erreur de Bayes est homogène. Mais cette hypothèse conduit, dans l'estimation du paramètre, à des difficultés numériques importantes, et elle ne peut pas être envisagée d'un point de vue pratique, en dehors du cas homoscédastique.
Nous supposons alors comme homogène, l'entropie globale des composantes. Nous définissons un algorithme générique $\tilde{\text{EM}}$, spécifique à cette contrainte. Il repose sur l'interprétation d'EM selon R.J. Hathaway ([60]) et peut être facilement implémenté dans une grande variété de mélanges (pas uniquement dans des mélanges gaussiens).
Nous montrons sur des données réelles que l'hypothèse de recouvrement homogène des classes peut être utile même si ce n'est pas une hypothèse exacte du vrai modèle. Elle permet, elle aussi, d'améliorer la qualité du modèle, la qualité du classifieur, et d'établir un lien interprétable entre populations.

Les deux formes précédentes de la classification simultanée ne s'excluent pas et peuvent au contraire être combinées. Il est possible en effet d'imposer à des mélanges gaussiens un recouvrement homogène de leurs classes et un lien conditionnel entre leurs composantes. Nous avons d'ailleurs montré sur des données réelles, l'intérêt que peut présenter l'association de ces deux types de contraintes.

Perspectives de la classification simultanée

Ce travail est un essai en faveur de la classification simultanée et non pas un recueil méthodologique exhaustif. Nous avons présenté quelques formes possibles de la classification simultanée, et expliqué le succès récurrent de la méthode par un argument théorique : établir un lien réaliste entre des populations améliore le compromis biais-variance du modèle. Mais la classification simultanée ne se limite pas aux modèles proposés ici et nous encourageons toutes les initiatives qui consistent à formaliser une information véritable, commune à des échantillons, par un lien formel entre populations.

Une extension immédiate de la classification simultanée consiste à envisager un lien affine conditionnel entre mélanges de la famille RTV.
La section 4.1 ne présente en guise de modèles intrapopulations, que quatre sortes de mélanges gaussiens. On peut souhaiter dans chaque population, disposer d'un échelonnement plus fin de la parcimonie. Or les modèles intrapopulations doivent être invariants à la modification des unités de mesure, pour être compatibles avec un lien affine conditionnel. On peut donc étendre la classification simultanée par lien affine aux modèles RTV, mais pas aux mélanges gaussiens d'interprétation géométrique.

L'extension de la classification simultanée à la grande dimension doit être une direc-

tion prioritaires des investigations à venir.
Nous avons exhibé au chapitre 4, de nombreuses situations où la classification d'un seul échantillon passe par la classification de plusieurs échantillons. Cette transformation, nous l'avons vu, peut prendre deux formes. Dans certains cas comme celui des voitures (Section 3, Fig. 3.1), il existe une variable catégorielle qui permet, si l'on tient compte de l'information qu'elle recèle, de voir l'échantillon à classifier comme la réunion de plusieurs échantillons. Dans d'autres cas comme celui d'Old Faithful (Section 4.2.1, Fig. 4.3) l'échantillon à classifier peut être mis en relation avec un autre échantillon de même nature. Nous pensons que des situations similaires sont fréquentes dans le domaine de la grande dimension. Mettre en évidence l'existence de plusieurs échantillons à classifier au lieu d'un seul, apporterait alors une information, dans un contexte où par nature l'information est rare. D'autre part il est primordial de réduire la dimension du paramètre en grande dimension puisque l'information apportée par la donnée est faible. Or c'est précisément ce que font les modèles de classification simultanée en établissant un lien entre populations. Ainsi, on voit que la classification simultanée est particulièrement adaptée au contexte de la grande dimension. Elle permettrait en quelques sortes d'optimiser l'usage de l'information disponible.

Nous défendons dans ce qui précède le principe de la classification simultanée en grande dimension. Penchons-nous maintenant sur les formes qu'elle peut prendre dans ce contexte spécifique.

Pour établir un lien stochastique affine entre populations, il est nécessaire que les modèles conditionnels soient invariants à la modification des unités de mesure. Pour cette raison la classification simultanée par lien affine (Chapitre 4) est exclue dans les mélanges de C. Bouveyron, mais elle est possible dans les mélanges de Factor Analyzers. Nous proposons à la section 4.4 une forme de la classification simultanée en grande dimension, basée sur une transformation affine des vecteurs conditionnels qui portent l'aléa principal dans un mélange de Factor Analyzers. Ils ne sont pas encore implémentés mais nous plaçons beaucoup d'espoir dans ce type de modèles.

A notre avis, la seconde forme de la classification simultanée, basée sur un recouvrement homogène de données conditionnelles, présente également un intérêt dans le contexte de la grande dimension ; et cet intérêt est de deux sortes.

Supposer un chevauchement homogène des groupes permettrait d'améliorer la classification, même si cette contrainte n'est pas tout à fait exacte dans le vrai modèle. Cette hypothèse réduirait la variabilité du modèle estimé, ce qui est capital en grande dimension ; et le biais qu'elle induit ne serait pas discernable puisque l'information apportée par les données, en grande dimension, n'est pas suffisante.

Le second intérêt est d'ordre technique. Nous savons qu'en matière de mélanges, une disproportion entre la taille de la donnée et la dimension du paramètre, provoque une multiplication des extrema locaux de la vraisemblance. Les expériences menées sur l'algorithme $\tilde{\text{E}}$M dans des échantillons de petite taille, nous ont amenés à constater qu'$\tilde{\text{E}}$M permet d'éviter les extrema locaux vers lesquels converge trop souvent EM. De la même

façon, supposer un chevauchement homogène des classes en grande dimension, pourrait permettre de converger plus sûrement vers le maximum de la vraisemblance en introduisant dans l'estimation une contrainte proche du vrai modèle.

Ceci nous amène naturellement aux perspectives liées à l'algorithme ẼM lui-même. Asymptotiquement et lorsque les classes se chevauchent de façon homogène dans le vrai modèle, ẼM converge vers le paramètre $\boldsymbol{\theta}^*$ qui optimise la vraisemblance sous cette contrainte. A taille d'échantillon finie, le paramètre estimé par ẼM approche $\boldsymbol{\theta}^*$, mais nous ignorons dans quelle mesure. Il serait intéressant de disposer, même dans un cas très particulier, de l'estimateur du maximum de vraisembmance du paramètre de mélanges dont l'entropie des composantes est homogène. Cela permettrait en effet de comparer à taille d'échantillon finie, le paramère estimé par ẼM à l'estimateur du maximum de vraisemblance.

Bibliographie

[1] S. Aeberhard, D. Coomans, and O. de Vel. The performance of statistical pattern recognition methods in high dimensional settings. In *IEEE Signal Processing Workshop on Higher Order Statistics. Ceasarea*, pages 14–16. John Wiley, 1994.

[2] H. Akaike. Information theory as an extension of the maximum likelihood principle. In B. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, 1973.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 :716–723, 1974.

[4] J.A. Anderson. Logistic discrimination. In P.R. Krishnaiah and L. Kanal, editors, *Handbook of statistics*, pages 169–191, 1982.

[5] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814) :796–815, 2000.

[6] A. Atkinson and M. Riani. Exploratory tools for clustering multivariate data. *Computational Statistics & Data Analysis*, 52(1) :272–285, September 2007.

[7] J.D. Banfield and A.E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49 :803–821, 1974.

[8] J.P. Baudry. *Sélection de Modèle pour la Classification Non Spervisée. Choix du Nombre de Classes.* Thèse de doctorat, Université Paris-Sud 11, 2009.

[9] J.P. Baudry, A.E. Raftery, G. Celeux, K. Lo, and R. Gottardo. Combining mixture components for clustering. Technical report, 2008.

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

[11] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[12] C. Biernacki. *Choix de modèles en classification.* Thèse de doctorat, Université Technologie de Compiègne, 1997.

[13] C. Biernacki. Précision sur les données et coude de la vraisemblance pour trouver le nombre de classes dans un mélange. *Revue de Statistiques Appliquées*, 47(1) :47–62, 1999.

[14] C. Biernacki. Partition latente et dégénérescence dans les mélanges gaussiens. In *42èmes Journées de Statistique*, Marseille, France, 2010.

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

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

[17] C. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics and Data Analysis*, 51(2) :587–600, 2006.

[18] C. Biernacki and A. Lourme. Simultaneous Model-Based Clustering of Data Arising from Different Populations. In eum edizioni università di macerata, editor, *Classification and Data Analysis 2007, Book of Short Papers. Sixth Meeting of the Classification and Data Analysis Group of the Italian Statistical Society*, pages 199–202, 2007.

[19] C.M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11 :443–482, 1999.

[20] C.M. Bishop and M. Svensén. Robust bayesian mixture modelling. *Neurocomputing*, 64 :235–252, 2005.

[21] C. Bouveyron. *Modélisation et classification des données de grande dimension : application à l'analyse d'images*. Thèse de doctorat, Université Grenoble 1, 2006.

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

[23] N.A. Campbell and R.J. Mahon. A multivariate study of variation in two species of rock crab of genus leptograpsus. *Australian Journal of Zoology*, 22 :417–425, 1974.

[24] M. Carreira-Perpinan. A review of dimension reduction techniques. Technical report, Dpt. of Computer Science, 1997.

[25] G. Celeux and J. Diebolt. The SEM Algorithm : A Probabilistic Teacher Algorithm derived from the EM Algorithm for the Mixture Problem. *Computational Statistics Quaterly*, 2 :73–82, 1985.

[26] G. Celeux and J. Diebolt. Comportement asymptotique d'un algorithme d'apprentissage probabiliste pour les mélanges de lois de probabilité. *Rapport de recherche INRIA*, 563, 1986.

[27] G. Celeux and G. Govaert. A classification em algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.*, 14(3) :315–332, 1992.

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

[29] A. Cutler and M.P. Windham. Information-based validity functionals for mixture analysis. In H. Bozdogan, editor, *Proceedings of the first US-Japan Conference o, the Frontiers of Statistical Modeling*, pages 149–170, 1993.

[30] F. D'Amico, Ú. Doyle, P. Smiddy, M. Jarry, A.C. Crook, and J. O'Halloran. Bib characteristics in the white-throated dipper cinclus cinclus and their possible role in assortative mating. *A paraître*.

[31] F. D'Amico and G. Hémery. Time-activity budgets and energetics of dipper cinclus cinclus are dictated by temporal variability of river flow. *Comparative Biochemistry and Physiology A Molecular and Integrative Physiology*, 148 (4) :811–820, 2007.

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

[33] A. Dasgupta and A.E. Raftery. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93 :294–302, 1998.

[34] A.P. Dawid. Properties of diagnostic data distributions. *Biometrics*, 32 :647–658, 1976.

[35] B. De Meyer, B. Roynette, P. Vallois, and M. Yor. On independent times and positions for Brownian motions. *Revista Matemática Iberoamericana (1985-2001)*, 18(3) :541–586, 2002.

[36] P. Demartines and J. Hérault. Curvilinear component analysis : a self-organizing neural network for non linear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1) :148–154, 1997.

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

[38] M. Di Zio, U. Guarnera, and R. Rocci. A mixture of mixture models for a classification problem : The unity measure error. *Comput. Stat. Data Anal.*, 51(5) :2573–2585, 2007.

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

[40] B.S. Everitt. A monte carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, 19 :79–89, 1981.

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

[42] B.W. Flury, M.J. Schmid, and A. Narayanan. Error rates in quadratic discrimination with constraints on the covariance matrices. *Journal of Classification*, 11 :101–120, 1994.

[43] I. Fodor. A survey of dimension reduction techniques. Technical report, Center of Applied Scientific Computing, 2002.

[44] F. Forbes, D. Wraith, S. Doyle, and E. Frichot. Multivariate robust clustering via mixture models, 2010. `http://carlit.toulouse.inra.fr/MSTGA/Reunion_juin2010/ForbesMSTGAmai2010.pdf`.

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

[46] J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405) :165–175, 1989.

[47] G. Galimberti and G. Soffritti. Model-based methods to identify multiple cluster structures in a data set. *Computational Statistics & Data Analysis*, 52(1) :520–536, September 2007.

[48] S. Girard. A nonlinear PCA based on manifold approximation. *Computational Statistics*, 15(2) :145–167, 2000.

[49] G. Govaert. *Data Analysis*. Wiley, 2009.

[50] J.C. Gower. Generalized procrustes analysis. *Psychometrika*, 40 :33–51, 1975.

[51] P.T. Green and C.M. Theobald. Sexing birds by discriminant analysis : further considerations. *Ibis*, 131(3) :442–447, 1989.

[52] J. Grifone. *Algèbre linéaire*. Cépaduès-Editions, 2002.

[53] P.J.F. Groenen and P.H. Franses. Visualizing time-varying correlations across stock markets. *Journal of Empirical Finance*, 7(2) :155–172, August 2000.

[54] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 :1157–1182, 2003.

[55] G.T. Hallgrimsson, S. Palsson, and R.W. Summers. Bill length : a reliable method for sexing purple sandpipers. *Journal of Field Ornithology*, 79 :87–92, 2008.

[56] A. Hardy. On the number of clusters. *Computational Statistics and Data Analysis*, 23 :83–96, 1996.

[57] A. Hardy and N. Kasoro. Une nouvelle méthode de classification pour les données intervalles. *Mthématiques et sciences humaines*, 187 :79–91, 2009.

[58] T. Hastie and W. Stuetzle. Principle curves. *Journal of the American Statistical Association*, 84 :502–516, 1989.

[59] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society series B*, 58 :155–176, 1996.

[60] R.J. Hathaway. Another interpretation of the em algorithm for mixture distributions. *Statistics and Probability Letters*, 4(2) :53–56, 1986.

[61] L.F. Hoogerheide, J.F. Kaashoek, and H.K. van Dijk. On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank : An application of flexible sampling methods using neural networks. *Journal of Econometrics*, 139(1) :154–180, July 2007.

[62] S. Ingrassia and R. Rocci. Constrained monotone em algorithms for finite mixture of multivariate gaussians. *Computational Statistics & Data Analysis*, 51(11) :5339–5351, 2007.

[63] J. Jacques. *Contributions à l'analyse de sensibilité et à l'analyse discriminante*. Thèse de doctorat, Université Grenoble 1, 2005.

[64] J. Jacques and C. Biernacki. Extension of model-based classification for binary data when training and test populations differ. *Journal of applied Statistics, A paraître*, pages 109–130, 2009.

[65] I.T. Jolliffe. *Principal Component Analysis*. Springer, second edition, 2002.

[66] M.A. Jorgensen and L.A. Hunt. Mixture modelling clustering of data sets with categorical and continuous variables. In D.L. Dowe, K.B. Korbe, and J.J. Oliver, editors, *ISIS : Information, Statistics and Induction in Science*, pages 375–383. World Scientific Publishing, 1996.

[67] R.E. Kass and A.E. Raftery. Bayes factors and model uncertainty. *Journal of the American Statistical Association*, 90 :773–795, 1995.

[68] C. Keribin. Consistent estimation of the order of mixture models. *Sankhyã, Series A*, 62(1) :49–66, 2000.

[69] J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.*, 27 :887–906, 1956.

[70] S.W. Kieffer. Seismicity at old faithful geyser : an isolated source of geothermal noise and possible analogue of volcanic seismicity. *Journal of Volcanology and Geothermal Research*, 22(1-2) :59 – 95, 1984.

[71] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag New York, 1995.

[72] W. Kristof and B. Wingersky. Generalization of the orthogonal procrustes rotation procedure for more than two matrices. In *Proceedings of the 79th annual convention of the American psychological association*, 1971.

[73] E. Lebarbier and T. Mary-Huard. Le critère bic, fondements théoriques et interprétation. *Journal de la Société Française de Statistique*, 1 :39–57, 2006.

[74] B.G. Leroux. Consistent estimation of a mixing proportion. *Annals of Statistics*, 20 :1350–1360, 1992.

[75] B.G. Lindsay and M.L. Lesperance. A review of semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47 :29–99, 1995.

[76] A. Lourme and C. Biernacki. Classification simultanée à base de mélanges gaussiens pour des échantillons d'origines multiples. In *41èmes Journées de Statistique, SFdS*, Bordeaux, France, 2009.

[77] A. Lourme and C. Biernacki. Joint Clustering of Two Samples from Multiple Origins by Equalizing their Entropy. In Salvatore Ingrassia and Roberto Rocci, editors, *Classification and Data Analysis 2009, Book of Short Papers. Seventh Meeting of the Classification and Data Analysis Group of the Italian Statistical Society*, pages 541–544, 2009.

[78] A. Lourme and C. Biernacki. Simultaneous Gaussian Model-Based Clustering for Samples of Multiple Origins. 2010. Submitted to *Computational Statistics*.

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

[80] A. Lourme and C. Biernacki. Simultaneous t-Model-Based Clustering for Data Differing over Time Period : Application for Understanding Companies Financial Health. 2010. Submitted to *Case Studies in Business, Industry and Government Statistics*.

[81] J.S. Marron and M.P. Wand. Exact mean integrated sqared error. *Annals of Statistics*, 20 :712–736, 1992.

[82] C. Maugis. *Sélection de variables pour la classification non supervisée par mélanges gaussiens. Application à l'étude de données transcriptômes.* Thèse de doctorat, Université Paris-Sud 11, 2008.

[83] C. Maugis, G. Celeux, and M-L. Martin-Magniette. Variable Selection in Model-based Clustering : A General Variable Role Modeling. 2008.

[84] C. Maugis, G. Celeux, and M-L. Martin-Magniette. Variable selection for clustering with gaussian mixture models. *Biometrics*, 65 :701–709, 2009.

[85] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

[86] G.J. McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 36 :318–324, 1987.

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

[88] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2 edition, March 2008.

[89] A.T. Mendoza-Rosas and S. De la Cruz-Reyna. A mixture of exponentials distribution for a simple and precise assessment of the volcanic hazard. *Natural Hazards and Earth System Science*, 9(2) :425–431, 2009.

[90] K. Pearson. Contributions to the theory of mathematical evolution. *Philosophical Transactions of the royal Society of London A*, 185 :71–110, 1894.

[91] S. Richardson and P.J. Green. On bayesian analysis of mixtures with unknown number of components. *Journal of the Royal Statistical Society series B*, 59 :731–792, 1997.

[92] J.S. Rinehart. *Geysers and geothermal energy / John S. Rinehart*. Springer-Verlag, New York :, 1980.

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

[94] S. Roweis. Em algorithms for PCA and SPCA. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[95] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500) :2323–2326, 2000.

[96] B. Schölkopf, A. Smola, and K-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5) :1299–1319, 1998.

[97] N.J. Schork and B. Thiel. Practical bayesian density estimation using mixtures of normals. *Mixture distributions in human genetics*, 39 :155–178, 1996.

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

[99] W. Seidel, K. Mosler, and M. Alker. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, 52 :481–487, 2000.

[100] C. Soubiran. Kinematics of the galaxy's stellar populations from a proper motion survey. *Astronomy and astrophysics*, 274 :181–188, 1993.

[101] J.M. Ten Berge. Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, 42(2) :267–276, 1977.

[102] J.B. Tenenbaum, V. Silva, and J.C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500) :2319–2323, 2000.

[103] M. Tenenhaus and V.E. Vinzi. Pls regression, pls path modeling and generalized procrustean analysis : a combined approach for multiblock analysis. *Journal of Chemometrics*, 19 :145–153, 2005.

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

[105] M. Tomonari, K. Senya, and M. Sueharu. Clustering images with multinomial mixture models. In *Proc. of the 8th International Symposium on Advanced Intelligent System*, 2007.

[106] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S. Fourth Edition.* Springer, New York, 2002.

[107] J.K. Vermunt and J. Magidson. Hierarchical mixture models for nested data structures. *Birds of Western Paleartic Update*, 1 :75–98, 1997.

[108] J.H. Ward. Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, 58 :236–244, 1963.