Précision sur les données et coude de la vraisemblance pour trouver le nombre de classes dans un mélange

Christophe Biernacki[†] [*]

† INRIA Rhône-Alpes - ZIRST - 655, avenue de l'Europe

38330 Montbonnot Saint Martin, France

Résumé

Le choix du nombre de classes pour un modèle de mélange gaussien s'appuie souvent sur la détection visuelle d'un coude dans la courbe du maximum de vraisemblance. Cutler et Windham ont rendu la détection de ce coude automatique au travers du critère EL, mais la détermination du seuil reste très empirique. Notre point de vue est d'utiliser une notion de précision sur les données, information souvent disponible chez l'expert du domaine de collecte, pour traduire cette notion de coude. Des applications sur des jeux de données réelles soulignent l'intérêt de cette démarche.

Mots-clés

Classification automatique, modèle de mélange gaussien, données réelles.

Abstract

To choose the number of clusters in Gaussian model-based clustering is often realized by looking for a threshold in the curve of the maximum likelihood. Cutler and Windham have suggested the EL procedure to detect automatically this threshold, but this procedure uses an empirical rule of thumb. In this paper, we suggest to use precision on data, information generally available, to define the threshold. This procedure has been applied to real data sets and has shown encouraging results.

Keywords

Cluster analysis, Gaussian mixture model, real data sets.

[*] L'auteur avait commencé la rédaction de cet article lors de sa thèse à l'Université de Technologie de Compiègne, laboratoire UMR CNRS 6599, BP 529, 60205 Compiègne Cedex, France.

1 Introduction

Le choix du nombre de classes en classification automatique est, encore aujoud'hui, une question non résolue, et de nouvelles propositions pour résoudre ce problème sont toujours d'actualité.

Dans cet article, nous proposons d'utiliser la précision sur les données pour trouver le nombre de classes lorsque la classification automatique s'appuie sur une hypothèse de mélange gaussien. Dans ce contexte probabiliste, le premier réflexe, après avoir visualisé graphiquement les données, est d'observer la courbe du maximum de vraisemblance pour plusieurs nombres de groupes. Cette courbe est croissante car le maximum de vraisemblance croît avec le nombre de composants du mélange mais, intuitivement, dès que le ¡¡bon¿¿ nombre de classes est dépassé, la courbe devrait croître beaucoup moins rapidement. La difficulté est alors de définir ce seuil de ¡¡croissance moins rapide¿¿. Notre technique permet de rendre relativement objectif le choix du seuil en utilisant une information très souvent disponible : la précision sur les données.

La section 2 rappelle l'hypothèse de mélange gaussien utilisée en classification automatique et décrit le critère EL de Cutler et Windham pour trouver un seuil dans la courbe du maximum de vraisemblance. L'utilisation de la précision des données pour trouver ce seuil est ensuite détaillée dans la section 3. La section 4 compare le critère EL originel à notre version objective de EL sur deux jeux de données réelles. La section 5 conclut cette étude.

2 Mélange gaussien et critère EL

Nous rappelons l'utilisation du mélange gaussien en classification automatique, puis nous décrivons le critère EL de Cutler et Windham, critère proposé pour trouver le nombre de classes en détectant un seuil dans la courbe du maximum de vraisemblance.

2.1 L'hypothèse de mélange gaussien

La classification automatique qui nous intéresse consiste à partitionner n données $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ de $\mathbb{R}^d$ en K groupes. L'hypothèse de mélange gaussien est alors classiquement utilisée dans ce contexte (Bock [3]). Chaque $\boldsymbol{x}_i$ $(1 \leq i \leq n)$ est une réalisation indépendante de la densité de mélange $f(\boldsymbol{x}|\boldsymbol{\theta})$ composée de K gaussiennes. Cette densité est donnée par

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \phi(\boldsymbol{x}|\boldsymbol{a}_k),$$

où $p_1, \ldots, p_K$ sont les proportions du mélange avec $p_k > 0$ pour $1 \leq k \leq K$ et $\sum_{k=1}^{K} p_k = 1$, et $\phi(\boldsymbol{x}|\boldsymbol{a})$ est une loi normale de paramètre $\boldsymbol{a} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ avec $\boldsymbol{\mu}$ le centre (c'est-à-dire l'espérance mathématique) et $\boldsymbol{\Sigma}$ la matrice de variance. Le paramètre du mélange est noté $\boldsymbol{\theta} = (p_1, \ldots, p_{K-1}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_K)$.

La probabilité conditionnelle à $\boldsymbol{x}_i$ $(1 \leq i \leq n)$, associée au paramètre $\boldsymbol{\theta}$, que cette observation provienne de la composante k $(1 \leq k \leq K)$ est notée $t_k(\boldsymbol{x}_i)$ et se calcule par

$$t_k(\boldsymbol{x}_i) = \frac{p_k \phi(\boldsymbol{x}_i|\boldsymbol{a}_k)}{f(\boldsymbol{x}_i|\boldsymbol{\theta})}.$$

La partition est déduite en affectant chaque $\boldsymbol{x}_i$ à la classe fournissant la probabilité conditionnelle la plus grande.

Dans cette situation, on peut obtenir un estimateur $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ en maximisant la log-vraisemblance

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln(f(\boldsymbol{x}_i|\boldsymbol{\theta}))$$

par l'algorithme EM de Dempster et al. [6].

2.2 Le critère EL

Le maximum de log-vraisemblance, $L(\hat{\boldsymbol{\theta}})$, ne peut être employé directement comme critère pour choisir le nombre de classes car c'est une fonction croissante de ce paramètre. De nombreux critères proposent de pénaliser le maximum de log-vraisemblance par une fonction (souvent linéaire) du nombre de degrés de liberté pour contrer cette croissance systématique. Notant respectivement $L(K) = L(\hat{\boldsymbol{\theta}})$ et $dl(K)$ le maximum de log-vraisemblance et le nombre de degrés

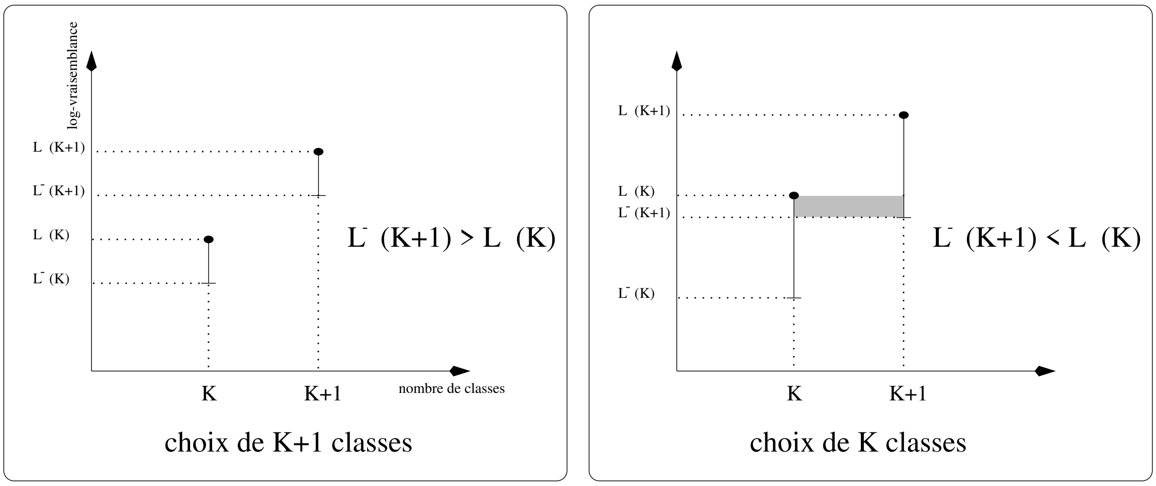


FIG. 1 – *Choix entre K et $K+1$ classes par le critère EL.*

de liberté pour K classes, les critères AIC (*An Information criterion*) de Akaike [1] et BIC (*Bayesian Information Criterion*) de Schwarz [7] appartiennent à cette catégorie de critères et s'expriment par :

$$\mathrm{AIC}(K) = -2L(K) + 2dl(K) \quad \text{et} \quad \mathrm{BIC}(K) = -2L(K) + \ln(n)dl(K).$$

Le nombre de classes retenu par chaque critère est alors celui qui le minimise. Mais, dans cet article, nous délaissons ce type de critères pour privilégier le choix d'un seuil de diminution de la croissance dans la courbe du maximum de vraisemblance en vue de trouver le nombre de classes.

À ce titre, Cutler et Windham [5] proposent le critère EL (*Elbow Likelihood*). Partant de $K^0 = 1$ $(q = 0)$, l'itération q de leur procédure s'énonce :

- si $K^q = n$ ou $L^-(K^q + 1) \leq L(K^q)$ alors $\hat{K} = K^q$ et arrêt ;
- sinon $K^{q+1} = K^q + 1$ et itérer,

avec $L(K) = L(\hat{\boldsymbol{\theta}})$ pour K classes et $L^-(K^q + 1) = L(K^q + 1) - 0.01|L(1)|$. Cette procédure est illustrée sur la figure 1.

L'inconvénient du critère EL est la détermination très empirique de $L^-(K)$. De plus, Cutler et Windham ont constaté une nette surestimation du nombre de composants par cette méthode. Nous pourrions aisément étendre le critère EL en remplaçant le seuil arbitraire 0.01 par un seuil α que choisirait l'utilisateur avec, par exemple, $0.01 \leq \alpha \leq 0.1$, c'est-à-dire poser $L^-(K^q+1) = L(K^q+1) - \alpha|L(1)|$. Ainsi, l'augmentation de α diminuerait le nombre de classes sélectionné et contribuerait à réduire la surestimation constatée. Pourtant, nous ne retenons pas cette optique pour les trois raisons suivantes : tout d'abord le caractère empirique de la procédure de Cutler et Windham n'est que très partiellement levé, ensuite le choix de α ne semble pas simple pour un utilisateur (α représente un pourcentage d'amélioration de la vraisemblance et est donc une notion abstraite pour un non-statisticien) et enfin on obtiendrait généralement des résultats différents, avec une même valeur α, suite à des changements d'échelle des données car les valeurs de la vraisemblance y sont très sensibles.

3 Critère EL et précision sur les données

L'objectif de cette section est d'utiliser la précision des données pour déterminer plus objectivement le terme $L^-(K)$ du critère EL. Nous présentons l'hypothèse de précision sur les données, puis l'expression de l'estimateur du paramètre du mélange avec cette précision et, enfin, la définition retenue pour $L^-(K)$. Nous concluons par un exemple illustrant l'intérêt de notre proposition.

3.1 L'hypothèse de précision sur les données

L'hypothèse de précision sur les données consiste à supposer qu'un expert du domaine relatif aux données $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ introduise un aléa sur celles-ci. Cet expert remplace ces n données par n données aléatoires $(\boldsymbol{x}_1+\boldsymbol{U}_1,\ldots,\boldsymbol{x}_n+\boldsymbol{U}_n)$, avec $\boldsymbol{U}_i$ $(1 \leq i \leq n)$ un vecteur aléatoire de loi

$$\pi_\gamma(\boldsymbol{u}) = \prod_{j=1}^{d} \mathcal{U}_{[-\gamma^j,\gamma^j]}(\boldsymbol{u}^j)$$

où $\mathcal{U}_{[a,b]}(x)$ est la distribution uniforme sur $[a,b]$, $\boldsymbol{u} = (\boldsymbol{u}^1,\ldots,\boldsymbol{u}^d)$ et $\boldsymbol{\gamma}^j \geq 0$ est la précision sur la variable j $(1 \leq j \leq d)$. L'ensemble des précisions sur toutes les variables, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}^1,\ldots,\boldsymbol{\gamma}^d)$, est fourni par l'expert. Cette substitution de données exprime l'indifférence du spécialiste entre toute donnée $\boldsymbol{x}_i$ $(1 \leq i \leq n)$ et toutes les valeurs contenues dans l'hyperparallèpipède de demi-côtés γ^j $(1 \leq j \leq d)$ centré en $\boldsymbol{x}_i$.

Typiquement, la détermination de la précision $\boldsymbol{\gamma}$ peut se faire de deux façons différentes. En premier lieu, l'expert peut déduire $\boldsymbol{\gamma}$ des caractéristiques de précision de l'instrument de mesure qui a produit $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$. Par exemple, si $(x_1,\ldots,x_n)$ sont des temps de course en seconde chronométrés au 10ème de seconde près, toute valeur $[x_i - 0.05, x_i + 0.05]$ $(1 \leq i \leq n)$ est admissible comme temps réel de course. L'expert peut aussi déduire la valeur de $\boldsymbol{\gamma}$ à partir de son propre ordre de grandeur. Par exemple une température corporelle au 10ème de degré suffit au médecin pour son diagnostic, même si le thermomètre utilisé est beaucoup plus précis.

Les données $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ étant remplacées par les n données perturbées $(\boldsymbol{x}_1 + \boldsymbol{U}_1,\ldots,\boldsymbol{x}_n + \boldsymbol{U}_n)$, nous ferons comme si ces dernières correspondaient à un échantillon aléatoire qui est une réalisation d'une densité $f(\boldsymbol{x}|\boldsymbol{\theta})$ d'un mélange de K lois gaussiennes. Mais quel estimateur choisir pour $\boldsymbol{\theta}$ dans ce contexte?

Notons $\hat{\boldsymbol{\theta}}_u$ l' estimateur de $\boldsymbol{\theta}$ avec l'échantillon $(\boldsymbol{x}_1 + \boldsymbol{U}_1,\ldots,\boldsymbol{x}_n + \boldsymbol{U}_n)$. Nous proposons de retenir, comme estimateur de $\boldsymbol{\theta}$, l'estimateur moyen $\mathrm{E}[\hat{\boldsymbol{\theta}}_u|\mathcal{X}]$, que l'on notera simplement $\mathrm{E}[\hat{\boldsymbol{\theta}}_u]$, où l'espérance est prise sur la distribution jointe des $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n$ conditionnellement à $\mathcal{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$. En faisant l'hypothèse que la précision retenue affecte peu les probabilités d'appartenance des points aux classes et, dans le cas particulier de classes de même matrices de variance, il est montré, dans l'annexe A, que

$$\mathrm{E}[\hat{p}_{k,u}] \simeq \hat{p}_k, \quad \mathrm{E}[\hat{\boldsymbol{\mu}}_{k,u}] \simeq \hat{\boldsymbol{\mu}}_k \text{ et } \quad \mathrm{E}[\hat{\boldsymbol{\Sigma}}_u] \simeq \hat{\boldsymbol{\Sigma}} + \beta \cdot \mathrm{diag}((\boldsymbol{\gamma}^1)^2,\ldots,(\boldsymbol{\gamma}^d)^2),$$

où $\beta > 0$ et où $\hat{p}_k$, $\hat{\boldsymbol{\mu}}_k$ et $\hat{\boldsymbol{\Sigma}}$ sont, rappelons-le, les estimateurs de p_k, $\boldsymbol{\mu}_k$ et $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ $(k = 1,\ldots,K)$, obtenus en maximisant la log-vraisemblance calculée à partir des $\boldsymbol{x}_i$ $(1 \leq i \leq n)$. Il ressort donc que, sous l'hypothèse que la norme $||\boldsymbol{\gamma}||$ de $\boldsymbol{\gamma}$ soit assez petite par rapport à la séparation des classes, l'estimateur moyen des proportions et des centres est identique à l'estimateur obtenu si $||\boldsymbol{\gamma}|| = 0$. Par contre, il n'en est pas de même avec la matrice de variance commune. En effet, la matrice de variance estimée $\mathbf{E}[\hat{\mathbf{\Sigma}}_u]$ est moins précise que $\hat{\mathbf{\Sigma}}$. L'introduction de la précision $\boldsymbol{\gamma}$ conduit donc à des classes plus dispersées autour de leur centre.

Remarque L'hypothèse que la précision influence peu l'appartenance aux classes, hypothèse traduite par la relation (4) de l'annexe A, a permis d'exprimer $\mathrm{E}[\hat{\boldsymbol{\theta}}_u]$ en fonction de $\hat{\boldsymbol{\theta}}$. Par contre, si $||\boldsymbol{\gamma}||$ est trop grand pour maintenir cette hypothèse, il reste bien sûr la possibilité de générer un certain nombre d'échantillons à partir de $(\boldsymbol{x}_1 + \boldsymbol{U}_1, \ldots, \boldsymbol{x}_n + \boldsymbol{U}_n)$, puis de maximiser la vraisemblance sur chacun d'entre eux et, enfin, d'estimer $\mathrm{E}[\hat{\boldsymbol{\theta}}_u]$ par moyennage de toutes les estimations. Ce procédé est néanmoins beaucoup plus long.

3.2 Un nouveau seuil pour le critère EL

L'hypothèse de précision posée et l'estimation moyenne des paramètres du mélange ainsi déduite, il reste à définir le terme $L^-(K)$ qu'utilisera le critère EL.

Nous proposons de choisir $L^-(K)$ comme l'espérance conditionnelle à $\mathcal{X}$ prise sur la loi jointe des $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ de la log-vraisemblance de l'estimateur moyen $\mathrm{E}[\hat{\boldsymbol{\theta}}_u]$ pour l'échantillon $(\boldsymbol{x}_1 + \boldsymbol{U}_1, \ldots, \boldsymbol{x}_n + \boldsymbol{U}_n)$. Formellement, on a

$$L^-(K) = \mathrm{E}[L(\mathrm{E}[\hat{\boldsymbol{\theta}}_u], \boldsymbol{x}_1 + \boldsymbol{U}_1, \ldots, \boldsymbol{x}_n + \boldsymbol{U}_n)],$$

où $L(\boldsymbol{\theta}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ est la log-vraisemblance de $\boldsymbol{\theta}$ en $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Nous vérifions par la proposition suivante que, si l'hypothèse de précision affecte peu les probabilités d'appartenance aux classes, nous avons $L^-(K) \leq L(K)$.

Proposition *Sous l'hypothèse que les probabilités d'appartenance aux classes des $\boldsymbol{x}_i$ et $\boldsymbol{x}_i + \boldsymbol{U}_i$ $(1 \leq i \leq n)$ calculées en $\boldsymbol{E}[\hat{\boldsymbol{\theta}}_u]$ soient très proches des probabilités d'appartenance des $\boldsymbol{x}_i$ calculées en $\hat{\boldsymbol{\theta}}$, c'est-à-dire (pour tout $i = 1, \ldots, n$)*

$$t_{k,\bar{u}}(\boldsymbol{x}_i) \simeq t_k(\boldsymbol{x}_i) \quad \textit{et} \quad t_{k,\bar{u}}(\boldsymbol{x}_i + \boldsymbol{U}_i) \simeq t_k(\boldsymbol{x}_i), \tag{1}$$

où $t_{k,\bar{u}}(\boldsymbol{x})$ est la probabilité conditionnelle calculée en $\boldsymbol{E}[\hat{\boldsymbol{\theta}}_u]$ que $\boldsymbol{x}$ appartienne à la classe k, alors on a l'inégalité

$$L^-(K) \leq L(K).$$

Preuve Remarquons qu'un calcul direct permet d'exprimer la log-vraisemblance L comme une somme de deux termes CL et e

$$L(\boldsymbol{\theta},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = CL(\boldsymbol{\theta},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) + e(\boldsymbol{\theta},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n), \tag{2}$$

avec, notant z_i le numéro de la classe où le point $\boldsymbol{x}_i$ ($1 \leq i \leq n$) est affecté, soit $z_i = \arg\max_k t_k(\boldsymbol{x}_i)$,

$$CL(\boldsymbol{\theta},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = \sum_{i=1}^{n} \ln(p_{z_i}\phi(\boldsymbol{x}_i|\boldsymbol{a}_{z_i})) \quad \text{et} \quad e(\boldsymbol{\theta},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = -\sum_{i=1}^{n} \ln t_{z_i}(\boldsymbol{x}_i).$$

En effet, on a

$$\begin{aligned} CL(\boldsymbol{\theta},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) + e(\boldsymbol{\theta},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) &= \sum_{i=1}^{n} \ln\left(\frac{p_{z_i}\phi(\boldsymbol{x}_i|\boldsymbol{a}_{z_i})}{t_{z_i}(\boldsymbol{x}_i)}\right) \\ &= \sum_{i=1}^{n} \ln(f(\boldsymbol{x}_i|\boldsymbol{\theta})) \\ &= L(\boldsymbol{\theta},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n). \end{aligned}$$

En posant $e = e(\hat{\boldsymbol{\theta}},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$, la relation $t_{k,\bar{u}}(\boldsymbol{x}_i + \boldsymbol{U}_i) \simeq t_k(\boldsymbol{x}_i)$ de (1) permet d'établir que

$$e(\mathrm{E}[\hat{\boldsymbol{\theta}}_u],\boldsymbol{x}_1 + \boldsymbol{U}_1,\ldots,\boldsymbol{x}_n + \boldsymbol{U}_n) \simeq e,$$

d'où, utilisant (2),

$$L^{-}(K) \simeq \mathrm{E}[CL(\mathrm{E}[\hat{\boldsymbol{\theta}}_u],\boldsymbol{x}_1 + \boldsymbol{U}_1,\ldots,\boldsymbol{x}_n + \boldsymbol{U}_n)] + e. \tag{3}$$

D'autre part, la fonction $\ln\phi(\boldsymbol{x}|\boldsymbol{a})$ étant concave, l'inégalité de Jensen permet d'écrire pour tout $\boldsymbol{\theta}$ et tout $i = 1,\ldots,n$

$$\begin{aligned} \mathrm{E}[\ln(p_{z_i}\phi(\boldsymbol{x}_i + \boldsymbol{U}_i|\boldsymbol{a}_{z_i}))] &\leq \ln(p_{z_i}\phi(\mathrm{E}[\boldsymbol{x}_i + \boldsymbol{U}_i]|\boldsymbol{a}_{z_i})) \\ &= \ln(p_{z_i}\phi(\boldsymbol{x}_i|\boldsymbol{a}_{z_i})), \end{aligned}$$

avec l'espérance prise sur la loi de $\boldsymbol{U}_i$. On en déduit que

$$\mathrm{E}[CL(\mathrm{E}[\hat{\boldsymbol{\theta}}_u],\boldsymbol{x}_1 + \boldsymbol{U}_1,\ldots,\boldsymbol{x}_n + \boldsymbol{U}_n)] \leq CL(\mathrm{E}[\hat{\boldsymbol{\theta}}_u],\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n),$$

ce qui entraîne, en utilisant (3),

$$L^-(K) \leq CL(\mathrm{E}[\hat{\boldsymbol{\theta}}_u],\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) + e.$$

Utilisant la relation $t_{k,\bar{u}}(\boldsymbol{x}_i) \simeq t_k(\boldsymbol{x}_i)$ de (1), on a

$$e(\mathrm{E}[\hat{\boldsymbol{\theta}}_u],\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) \simeq e,$$

d'où

$$L(\mathrm{E}[\hat{\boldsymbol{\theta}}_u],\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) \simeq CL(\mathrm{E}[\hat{\boldsymbol{\theta}}_u],\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) + e,$$

et donc

$$L^-(K) \leq L(\mathrm{E}[\hat{\boldsymbol{\theta}}_u],\boldsymbol{x}_1,\dots,\boldsymbol{x}_n).$$

Enfin, on conclut en rappelant que $L(K)$ est le maximum de log-vraisemblance.

Cette proposition est bien sûr importante pour la cohérence de notre approche.

Le calcul pratique de $L^-(K)$ peut aisément se réaliser par une méthode de Monte-Carlo. On génère m échantillons indépendants $(\boldsymbol{u}_{1,\ell},\dots,\boldsymbol{u}_{n,\ell})$ $(\ell = 1,\dots,m)$ et on a

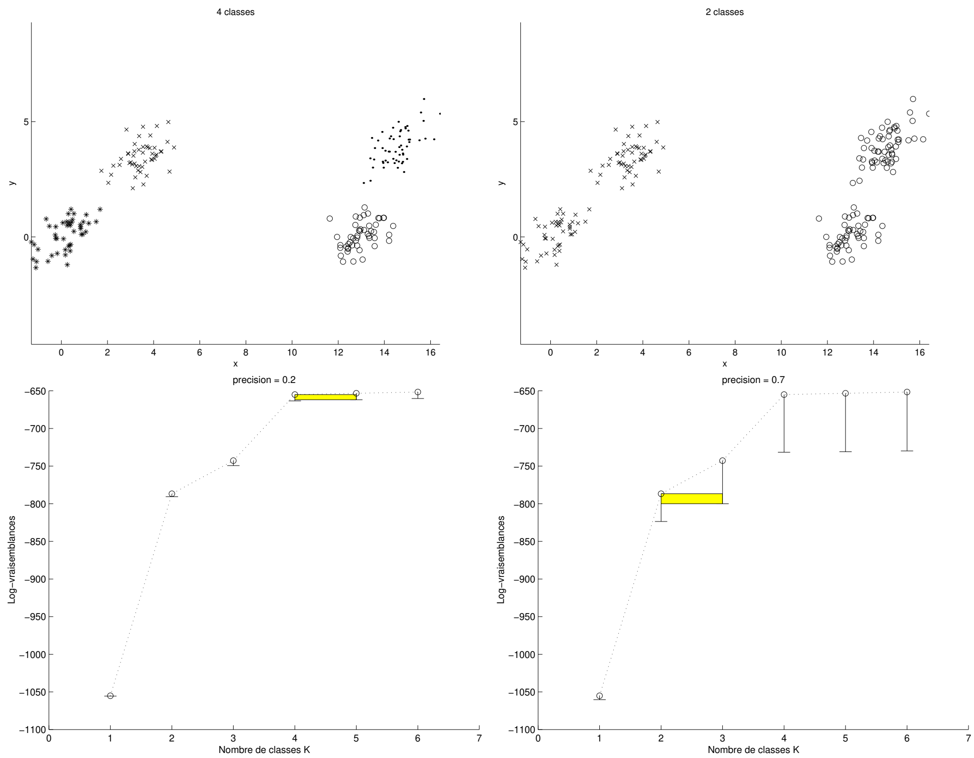
$$L^-(K) \simeq \frac{1}{m}\sum_{\ell=1}^{m} L(\mathrm{E}[\hat{\boldsymbol{\theta}}_u],\boldsymbol{x}_1+\boldsymbol{u}_{1,\ell},\dots,\boldsymbol{x}_n+\boldsymbol{u}_{n,\ell}).$$

Dans les expériences numériques qui vont suivre, nous prendrons $m = 100$.

3.3 Un exemple

Le choix de la précision $\boldsymbol{\gamma}$ a un impact sur le nombre de classes retenu comme l'illustre la figure 2. Le jeu de données est composé d'une réalisation de quatre classes gaussiennes. Le lecteur peut aussi voir seulement deux classes (une a gauche et une à droite) s'il s'éloigne suffisamment de la figure. L'éloignement indique une perte de précision car on agglomère visuellement les paquets trop imbriqués en de plus gros paquets.

Notre méthode a un comportement similaire : on détecte quatre classes avec une imprécision $\gamma = \boldsymbol{\gamma}^1 = \boldsymbol{\gamma}^2 = 0.2$ (figure 2 (a)), mais on détecte deux classes avec une imprécision plus grande $\gamma = 0.7$ (figure 2 (b)). Ces deux sous-figures se lisent à l'aide de la figure 1 : La courbe du maximum de log-vraisemblance $L(K)$ est tracée en pointillé et chaque valeur $L^{-}(K)$ est représentée au bas du trait vertical partant de $L(K)$. La zone grisée indique l'endroit où le coude est détecté par le critère.



(a) 4 classes ($\gamma = 0.2$) (b) 2 classes ($\gamma = 0.7$)

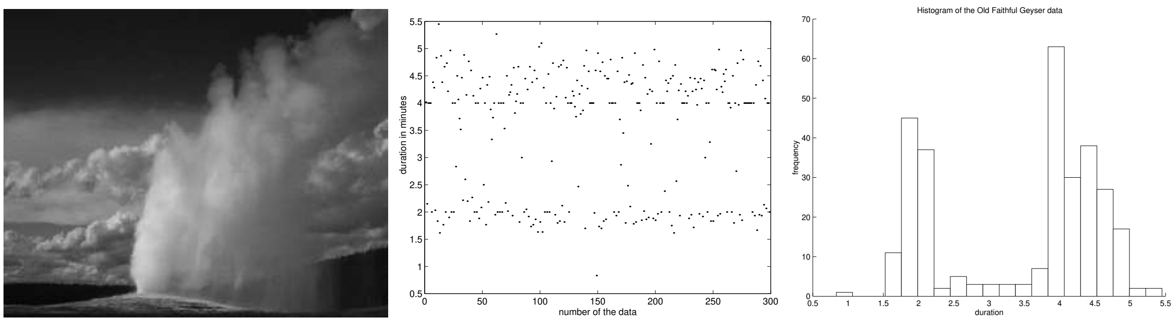
FIG. 2 – *Détection de deux (a) ou quatre classes (b) en fonction de la précision choisie (la légende des deux courbes de log-vraisemblance est donnée sur la figure 1).*

4 Applications

Cette section propose de comparer, sur deux jeux de données réelles, le critère EL utilisant l'hypothèse de précision et le critère EL originel, ainsi que les deux critères AIC et BIC de pénalisation de la log-vraisemblance. Dans chacune des deux situations, nous donnons la plage de précision qui permet de trouver le ¡¡bon¿¿ nombre de classes avec le nouveau critère.

4.1 Le geyser

¡¡The Old Faithful Geyser¿¿ (Yellowstone National Park, Wyoming, figure 3 (a)) est connu pour la grande régularité de ces éruptions. Ces durées peuvent être partitionnées en deux groupes : le groupe des 2 minutes et le groupe des 4 minutes. Azzalini and Bowman [2] ont mesuré 299 durées (voir figure 3 (b))) et l'histogramme de ces données, figure 3 (c), permet d'identifier clairement ces deux classes.



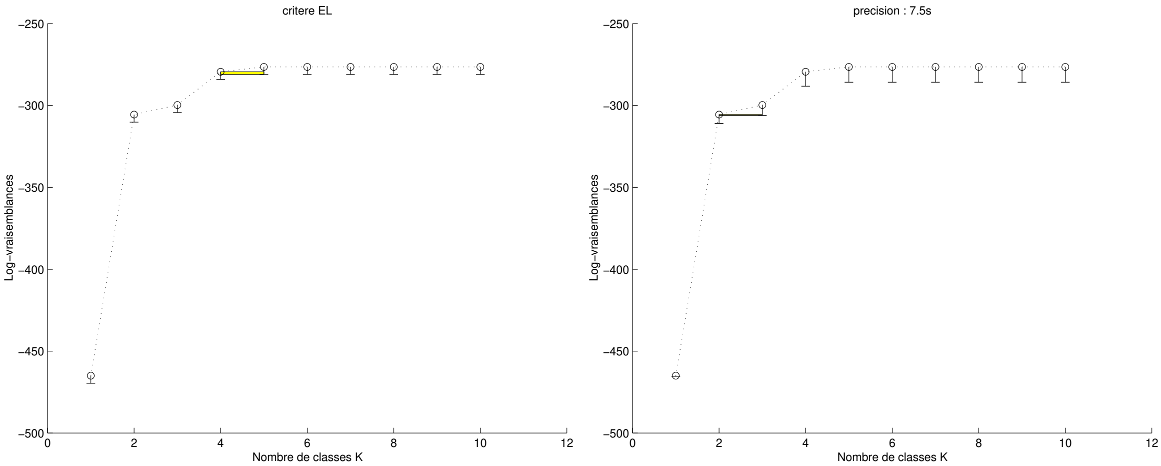
(a) (b) (c)

FIG. 3 – *The Old Faithful Geyser (c), durées d'éruption en minutes (b) et histogramme de ces données (c).*

Nous supposons que les classes ont même variance et nous calculons $\hat{\boldsymbol{\theta}}$ en initialisant l'algorithme EM 50 fois au hasard puis en ne retenant que la solution maximisant la log-vraisemblance. Le critère de Cutler et Windham détecte 4 classes (figure 4 (a)), tout comme les critères AIC et BIC (tableau 1). Le critère utilisant la précision trouve 2 classes si $\gamma \geq 6.9$s, c'est-à-dire pour des mesures à au moins 13.8s près (voir la figure 4 (b) avec $\boldsymbol{\gamma} = 7.5$s). Nous ne donnons pas de borne supérieure aux valeurs de γ permettant de détecter 2 classes car une précision à la minute près ($\gamma = 30$s) permet encore de trouver 2 classes et il nous semble improbable qu'un expert dépasse ce seuil de précision sur de telles données.

4.2 Les papillons

Ce jeu de données comporte 23 papillons, chacun étant mesuré en 4 endroits de son anatomie (longueurs des ailes, etc.), et a été présenté par Celeux et



(a) (b)

FIG. 4 – *Détection de 4 classes par le critère EL originel (a) et de 2 classes par le nouveau seuil avec $\gamma = 7.5s$ (b) (la légende des courbes de log-vraisemblance est donnée sur la figure 1).*

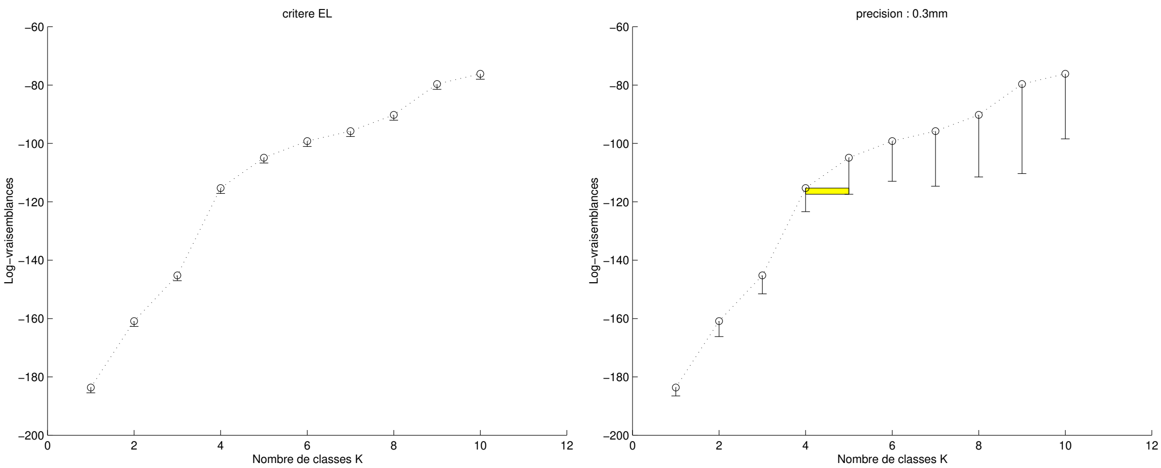
Critère	*K*									
	1	2	3	4	5	6	7	8	9	10
AIC	2120	1589	1507	**1337**	1340	1343	1346	1348	1347	1350
BIC	2136	1615	1543	**1384**	1396	1409	1422	1433	1443	1455

TAB. 1 – *Valeurs des critères AIC et BIC dans le cas du geyser.*

n°	Classe	$\boldsymbol{x}^1$ (mm)	$\boldsymbol{x}^2$ (mm)	$\boldsymbol{x}^3$ (mm)	$\boldsymbol{x}^4$ (mm)
1	2	22	35	24	19
2	3	24	31	21	22
3	1	27	36	25	15
4	3	27	36	24	23
5	2	21	33	23	18
6	4	26	35	23	32
7	1	27	37	26	15
8	3	22	30	19	20
9	3	25	33	22	22
10	1	30	41	28	17
11	2	24	39	27	21
12	1	29	39	27	17
13	1	29	40	27	17
14	3	28	36	23	24
15	2	22	36	24	20
16	3	23	30	20	20
17	1	28	38	26	16
18	1	25	34	23	14
19	1	26	35	24	15
20	2	23	37	25	20
21	1	31	42	29	18
22	3	26	34	22	21
23	2	24	38	26	21

TAB. 2 – *Les papillons.*

Robert [4]. Les papillons peuvent être groupés en quatre classes, l'une d'entre elles ne contenant qu'un seul individu (tableau 2).



(a) (b)

FIG. 5 – *Détection d'au moins 10 classes par le critère EL originel (a) et de 4 classes par le nouveau seuil avec $\boldsymbol{\gamma}^1 = \ldots = \boldsymbol{\gamma}^4 = 0.3mm$ (b) (la légende des courbes de log-vraisemblance est donnée sur la figure 1).*

Nous supposons que les classes ont même matrice de variance et nous calculons $\hat{\boldsymbol{\theta}}$ en initialisant l'algorithme EM 50 fois au hasard en ne conservant, de nouveau, que la meilleure solution au sens du critère de vraisemblance. Nous poserons que la précision est identique sur chaque composante, donc que $\boldsymbol{\gamma}^1 = \ldots = \boldsymbol{\gamma}^d = \gamma$. Le critère de Cutler et Windham détecte au moins 10 classes (figure 5 (a)), comme le critère AIC (tableau 3), et le critère BIC choisit 5 classes (tableau 3). Le critère utilisant la précision trouve 4 classes si $\gamma \in [0.29{,}0.54]$mm, c'est-à-dire pour des mesures à au moins 0.58mm près et à au plus 1.08mm près (voir la figure 5 (b) avec $\boldsymbol{\gamma} = 0.3$mm). Les données du tableau 2 sont fournies à 1mm près et nous remarquons alors que la précision $\gamma = 0.5$ associée permet de détecter 4 classes.

Critère	K									
	1	2	3	4	5	6	7	8	9	10
AIC	395.2	359.7	338.4	288.6	282.5	278.8	277.1	278.9	267.2	**266.4**
BIC	411.1	381.2	365.6	321.5	**321.1**	323.1	327.0	334.6	328.5	333.4

TAB. 3 – *Valeurs des critères AIC et BIC dans le cas des papillons.*

5 Conclusion

Nous avons proposé une méthode de détection d'un seuil dans la courbe du maximum de vraisemblance pour choisir le nombre de classes dans le cadre du modèle de mélange gaussien. Nous utilisons la notion de précision sur les données, information très souvent disponible, et dont il serait dommage de se priver.

L'utilisation de la précision a été détaillée dans le cas particulier de classes de même matrice de variance. Deux jeux de données réelles ont permis d'appliquer cette procédure et de montrer son intérêt par rapport au critère EL de Cutler et Windham, procédure empirique de choix du seuil. La procédure a aussi été comparée aux critères AIC et BIC qui pénalisent la log-vraisemblance par la complexité du modèle. Sur l'ensemble de ces critères (EL, AIC et BIC), et des critères de choix du nombre de classes en général, la procédure proposée à l'avantage de pouvoir choisir différents nombres de classes par réglage de la précision sur les données voulue par l'utilisateur. Cette possibilité d'obtenir différents nombres de classes en fonction de la précision retenue met en évidence le fait que la classe est un concept très dépendant du domaine des données. Le caractère très concret de la précision rend ce réglage compréhensible par chacun donc assez objectif.

Il serait néanmoins interessant de prolonger les expérimentations sur d'autres jeux de données. Il serait aussi possible d'étendre sans difficulté ce principe à des classes de matrices de variance différentes et même à des classes non gaussiennes.

Remerciements L'auteur remercie les rapporteurs pour leurs conseils.

A Annexes

Cette section détaille le calcul de l'estimateur du paramètre du mélange sous l'hypothèse de précision.

Notons $\hat{\boldsymbol{\theta}}_u$ et $\hat{\boldsymbol{\theta}}$ les estimateurs de $\boldsymbol{\theta}$ avec respectivement les échantillons $(\boldsymbol{x}_1 + \boldsymbol{U}_1, \ldots, \boldsymbol{x}_n + \boldsymbol{U}_n)$ et $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Ce deuxième échantillon correspond au cas où l'expert n'introduit aucun aléa sur les données, c'est-à-dire $||\boldsymbol{\gamma}|| = 0$ où $||.||$ est une norme quelconque. Exprimons alors $\hat{\boldsymbol{\theta}}_u$ en fonction de $\hat{\boldsymbol{\theta}}$ dans le cas où les classes ont même matrice de variance puis déduisons l'expression de l'espérance de cet estimateur, l'espérance étant prise sur la distribution jointe des $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$, conditionnellement aux $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$. C'est cet estimateur moyen que nous utiliserons comme estimateur de $\boldsymbol{\theta}$ sous cette hypothèse de précision.

A.1 Estimateur $\hat{\boldsymbol{\theta}}_u$

Pour exprimer $\hat{\boldsymbol{\theta}}_u$, supposons que $||\boldsymbol{\gamma}||$ soit assez petite par rapport à la séparation des classes, c'est-à-dire que les probabilités d'appartenance aux classes de chaque point $\boldsymbol{x}_i$ et $\boldsymbol{x}_i + \boldsymbol{U}_i$ $(i = 1, \ldots, n)$ soient très proches. Cela se traduit par :

$$t_{k,u}(\boldsymbol{x}_i + \boldsymbol{U}_i) \simeq t_{ik} \tag{4}$$

où $t_{k,u}(\boldsymbol{x}_i + \boldsymbol{U}_i)$ est la probabilité que $\boldsymbol{x}_i + \boldsymbol{U}_i$ appartienne à la classe k, probabilité calculée en $\hat{\boldsymbol{\theta}}_u$, et $t_{ik} = t_k(\boldsymbol{x}_i)$ est la probabilité que $\boldsymbol{x}_i$ appartienne à la classe k, probabilité calculée en $\hat{\boldsymbol{\theta}}$.

L'estimateur de la proportion de la classe k ($k = 1,\ldots,K$) est donné par

$$\begin{aligned}
\hat{p}_{k,u} &= \frac{\sum_{i=1}^n t_{k,u}(\boldsymbol{x}_i + \boldsymbol{U}_i)}{n} \\
&\simeq \frac{\sum_{i=1}^n t_{ik}}{n} \\
&= \hat{p}_k.
\end{aligned}$$

De même, le centre de la classe k s'exprime par

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_{k,u} &= \frac{\sum_{i=1}^n t_{k,u}(\boldsymbol{x}_i + \boldsymbol{U}_i)(\boldsymbol{x}_i + \boldsymbol{U}_i)}{\sum_{i=1}^n t_{k,u}(\boldsymbol{x}_i + \boldsymbol{U}_i)} \\
&\simeq \frac{\sum_{i=1}^n t_{ik}(\boldsymbol{x}_i + \boldsymbol{U}_i)}{\sum_{i=1}^n t_{ik}} \\
&= \hat{\boldsymbol{\mu}}_k + \frac{\sum_{i=1}^n t_{ik}\boldsymbol{U}_i}{\sum_{i=1}^n t_{ik}}.
\end{aligned}$$

Enfin, nous limitant au cas de matrices de variance égales pour toutes les classes, la matrice de variance commune s'écrit

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_u &= \frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K t_{k,u}(\boldsymbol{x}_i + \boldsymbol{U}_i)(\boldsymbol{x}_i + \boldsymbol{U}_i - \hat{\boldsymbol{\mu}}_{k,u})(\boldsymbol{x}_i + \boldsymbol{U}_i - \hat{\boldsymbol{\mu}}_{k,u})' \\
&\simeq \frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K t_{ik}(\boldsymbol{x}_i + \boldsymbol{U}_i - \hat{\boldsymbol{\mu}}_{k,u})(\boldsymbol{x}_i + \boldsymbol{U}_i - \hat{\boldsymbol{\mu}}_{k,u})' \\
&\simeq \frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K t_{ik}\left(\boldsymbol{x}_i + \boldsymbol{U}_i - \hat{\boldsymbol{\mu}}_k - \frac{\sum_{\ell=1}^n t_{\ell k}\boldsymbol{U}_\ell}{\sum_{\ell=1}^n t_{\ell k}}\right)\left(\boldsymbol{x}_i + \boldsymbol{U}_i - \hat{\boldsymbol{\mu}}_k - \frac{\sum_{\ell=1}^n t_{\ell k}\boldsymbol{U}_\ell}{\sum_{\ell=1}^n t_{\ell k}}\right)' \\
&= \hat{\boldsymbol{\Sigma}} + \frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K t_{ik}\left\{\left(\boldsymbol{U}_i - \frac{\sum_{\ell=1}^n t_{\ell k}\boldsymbol{U}_\ell}{\sum_{\ell=1}^n t_{\ell k}}\right)\left(\boldsymbol{U}_i - \frac{\sum_{\ell=1}^n t_{\ell k}\boldsymbol{U}_\ell}{\sum_{\ell=1}^n t_{\ell k}}\right)'\right. \\
&\quad \left. + (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)\left(\boldsymbol{U}_i - \frac{\sum_{\ell=1}^n t_{\ell k}\boldsymbol{U}_\ell}{\sum_{\ell=1}^n t_{\ell k}}\right)' + \left(\boldsymbol{U}_i - \frac{\sum_{\ell=1}^n t_{\ell k}\boldsymbol{U}_\ell}{\sum_{\ell=1}^n t_{\ell k}}\right)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)'\right\}
\end{aligned}$$

A.2 Estimateur moyen $\mathbb{E}[\hat{\boldsymbol{\theta}}_u]$

Calculons maintenant l'espérance, conditionnellement aux $\boldsymbol{x}_i$ ($i = 1,\ldots,n$), de l'estimateur $\hat{\boldsymbol{\theta}}_u$ sur la distribution jointe des $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n$.

L'estimateur moyen des proportions ($k = 1,\ldots,K$) s'énonce

$$\mathrm{E}[\hat{p}_{k,u}] \simeq \hat{p}_k.$$

De même, l'estimateur moyen des centres est donné par :

$$\mathrm{E}[\hat{\boldsymbol{\mu}}_{k,u}] \simeq \hat{\boldsymbol{\mu}}_k,$$

car pour tout $i = 1,\ldots,n$ on a $\mathbf{E}[\boldsymbol{U}_i] = \mathbf{0}$. Utilisant aussi le fait que pour tout $i = 1,\ldots,n$ et pour tout $j = 1,\ldots,n$ avec $i \neq j$ on a $\mathrm{E}[\boldsymbol{U}_i\boldsymbol{U}'_j] = \mathbf{0}$, l'estimateur moyen de la matrice de variance commune s'écrit

$$\mathrm{E}[\hat{\boldsymbol{\Sigma}}_u] \simeq \hat{\boldsymbol{\Sigma}} + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}\left(\mathrm{E}[\boldsymbol{U}_i\boldsymbol{U}'_i] - \frac{2t_{ik}}{\sum_{\ell=1}^{n} t_{\ell k}}\mathrm{E}[\boldsymbol{U}_i\boldsymbol{U}'_i] + \frac{\sum_{\ell=1}^{n} t_{\ell k}^2 \mathrm{E}[\boldsymbol{U}_\ell\boldsymbol{U}'_\ell]}{(\sum_{\ell=1}^{n} t_{\ell k})^2}\right).$$

Notant $\boldsymbol{U}$ une variable de même loi que les n variables $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n$, on a

$$\mathrm{E}[\hat{\boldsymbol{\Sigma}}_u] \simeq \hat{\boldsymbol{\Sigma}} + \mathrm{E}[\boldsymbol{U}\boldsymbol{U}']\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}\left(1 - \frac{2t_{ik}}{\sum_{\ell=1}^{n} t_{\ell k}} + \frac{\sum_{\ell=1}^{n} t_{\ell k}^2}{(\sum_{\ell=1}^{n} t_{\ell k})^2}\right).$$

Remarquant que $\mathrm{E}[\boldsymbol{U}] = 0$, $\mathrm{Cov}(\boldsymbol{U}^i,\boldsymbol{U}^j) = 0$ ($i \neq j$ avec $i,j = 1,\ldots,d$) et $\mathrm{Var}(\boldsymbol{U}^i) = \frac{(\boldsymbol{\gamma}^i)^2}{3}$ ($i = 1,\ldots,d$), on obtient

$$\mathrm{E}[\boldsymbol{U}\boldsymbol{U}'] = \frac{1}{3}\mathrm{diag}((\boldsymbol{\gamma}^1)^2,\ldots,(\boldsymbol{\gamma}^d)^2),$$

où $\mathrm{diag}(a_1,\ldots,a_d)$ est une matrice diagonale d'éléments diagonaux $a_1,\ldots,a_n$. D'où

$$\mathrm{E}[\hat{\boldsymbol{\Sigma}}_u] \simeq \hat{\boldsymbol{\Sigma}} + \frac{1}{3}\mathrm{diag}((\boldsymbol{\gamma}^1)^2,\ldots,(\boldsymbol{\gamma}^d)^2)\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}\left(1 - \frac{2t_{ik}}{\sum_{\ell=1}^{n} t_{\ell k}} + \frac{\sum_{\ell=1}^{n} t_{\ell k}^2}{(\sum_{\ell=1}^{n} t_{\ell k})^2}\right),$$

donc, après une dernière réorganisation, on a

$$\mathrm{E}[\hat{\boldsymbol{\Sigma}}_u] \simeq \hat{\boldsymbol{\Sigma}} + \frac{1}{3}\mathrm{diag}((\boldsymbol{\gamma}^1)^2,\ldots,(\boldsymbol{\gamma}^d)^2)\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \frac{t_{ik}}{(\sum_{\ell=1}^{n} t_{\ell k})^2}((\sum_{\ell=1}^{n} t_{\ell k} - t_{ik})^2 + \sum_{\ell=1,\ell\neq i}^{n} t_{\ell k}^2).$$

Références

[1] 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, Budapest, Akademiai Kiado, 1973.

[2] A. Azzalini and A.W. Bowman. A Look at Some Data on the Old Faithful Geyser. *Applied Statistics*, 39:357–365, 1990.

[3] H.H. Bock. Statistical Testing and Evaluation Methods in Cluster Analysis. In *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions*, pages 116–146, Calcutta, December 16-19 1981.

[4] G. Celeux and C. Robert. Une histoire de discrétisation (avec discussion). *La Revue de Modulad*, 11:7–44, 1993.

[5] A. Cutler and M.P. Windham. Information-Based Validity Functionals for Mixture Analysis. In H. Bozdogan, editor, *Proceedings of the first US-Japan Conference on the Frontiers of Statistical Modeling*, pages 149–170, Amsterdam, 1993. Kluever (Academic Publishers).

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data with the EM Algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[7] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6:461–464, 1978.