UNIVERSITE TOULOUSE III-PAUL SABATIER
U.F.R. Mathématiques Informatique Gestion

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE TOULOUSE III
Discipline : Mathématiques Appliquées (Probabilités et Statistiques)

présentée et soutenue
par

Guillaume SAINT PIERRE

Le 26 septembre 2003

Titre :

IDENTIFICATION DU NOMBRE DE COMPOSANTS D'UN MELANGE GAUSSIEN PAR CHAINES DE MARKOV A SAUTS REVERSIBLES DANS LE CAS MULTIVARIE OU PAR MAXIMUM DE VRAISEMBLANCE DANS LE CAS UNIVARIE

Sous la direction de :
Bernard GAREL

Devant le Jury composé de

M. Philippe BESSE	Professeur à l'Université Toulouse III	Examinateur
M. Gilles CELEUX	Directeur de Recherche (INRIA)	Examinateur
M. Bernard GAREL	Professeur à l'I.N.P. (ENSEEIHT)	Directeur de thèse
M. Gérard GOVAERT	Professeur à l'U.T.C.	Rapporteur
M. Christian ROBERT	Professeur à l'Université Paris Dauphine	Rapporteur
Mme. Christine THOMAS-AGNAN	Professeur à l'Université Toulouse I	Examinateur

REMERCIEMENTS

En premier lieu, je tiens à remercier Bernard Garel pour m'avoir encadré durant cette thèse. Il a su faire preuve de compréhension et de patience afin de s'adapter au caractère parfois emporté qui est le mien. Son sens de la rigueur et son honnêteté ont constitué un cadre de travail idéal à la libre expression de mes choix mathématiques. je lui exprime toute mon estime.

Je remercie Christian Robert et Gérard Govaert d'avoir bien voulu s'acquitter de la tâche ingrate d'évaluer ce travail, chose qu'ils ont fait avec autant de rapidité que d'efficacité.

Je voudrais exprimer ma reconnaissance à Gilles Celeux de me faire l'honneur de participer au jury de cette thèse, ainsi que pour les multiples "coups de pouce" et conseils dispensés sans compter.

J'adresse mes remerciements sincères à Philippe Besse et Christine Thomas-Agnan pour avoir accepté de participer à ce jury.

Nombreux sont ceux qui m'ont aidé ou encouragé dans l'élaboration de ce travail. Je pense particulièrement à Christophe Ambroise et Michel Doisy, ainsi qu'à Guillaume Bouchard alias "Mister Switch". Ils m'ont apporté un peu de cette confiance qui fait parfois défaut au cours des longues années de thèse.

Il le sait, je ne le remercierai jamais assez , je veux parler de l'empereur du LaTex, du magicien des alpha-stables, de mon ami Ludovic d'Estampes...

Merci aussi à toute l'équipe de l'INRIA Rhônes-Alpes pour m'avoir accueilli chaleureusement à plusieurs reprises.

Je remercie le Laboratoire de Statistique et Probabilités de l'Université Paul Sabatier pour m'avoir acceuilli dans ses locaux.

Ayant principalement travaillé ces dernières années dans les locaux du groupe Statistique et probabilités du LEN7 (Laboratoire d'Electronique de l'ENSEEIHT), je tiens à exprimer ma reconnaissance à Thierry Bosch pour laisser quelques "matheux" naviguer librement au sein de son laboratoire.

Je décerne une mention spéciale à Xavier Barthélémy, aussi bon en Fortran qu'en dosage de Ti-punch, ainsi qu'à Boris Lenseigne, aussi passionné de C++ que de death-metal.

Je dois aussi féliciter mon catalan préféré Jean-Michel Loubes pour sa recette du café-guronzan, et Renaud Camalès qui préfère nettement le rhum. Je pense bien sûr à tous mes co-bureaux qui m'ont supporté, à Franck Goussanou pour toutes les crises de rire, ainsi que Elie et Cécile que je vois trop peu souvent.

J'ai des remerciements tout particuliers à adresser aux anciens thésards de Mathématiques pures, à Tof "le verrou", au Pennach le beau catalan, aux Marco que j'aimerais revoir, à Jihad, Nico et Marianne, Vincent, et bien sûr Richard et "Le Trille".

"Gardarem lou pasatge !", vous me manquez.

Evidemment, je ne peux oublier toute l'équipée du laboratoire MIP. Certaines parties de "Jungle Speed" restent parmi les grands souvenirs de nombreux thésards. Merci à vous Pierre et Marie, et à vous tous : Mathias, Abderahim, Sandrine, Marion, Pascal et Sandrine, Jihen, Claudia, et bien sûr les deux Sophie. Bonne route à vous tous !

"Maximum respect" à tous mes amis qui m'ont soutenu et supporté quand j'en avais besoin. Parfois cela fait du bien d'oublier un peu, n'est ce pas Marino ?

"Spéciale dédicace" à Dodo et Céline pour m'avoir recueilli, à Franck pour ses petits cartons, à Guillaume et Nono pour les couinches interminables, à Vanessa pour son sourire magique, au Castor pour ses escargots statisticiens, au Nain pour les analyses psychanalitiques et à Vad pour les "board-games" hystériques.

Merci enfin à tout les "gars" de l'ENSEEIHT : David Dupuy pour ses fines analyses "rugbalistiques", Steve pour son sens de la rigueur, Patrick pour son franc-parler, David Daurenjou car il est toujours prêt à faire la fête, Denis qui essaie de garder son calme au milieu de tout ça, et Caroline qui se demande ce qui l'attend.

Je dois aussi féliciter "Tonton" qui nous élève au Kebab depuis 4 ans, pour sa bonne humeur aussi agréable que son café.

Merci à tout mes élèves pour m'avoir convaincu qu'enseignant pouvait être un très beau métier, à tout les "footeux" pour avoir supporté ma nervosité, aux "Ximaz" pour m'écouter couiner de jour en jour, et aussi aux "cochons" pour les tournois de caps.

L'amitié fidèle des Balmanais m'a apporté la sérénité. Merci pour tout JF & Valérie, Renaud, Stéphane & Valérie ainsi que Christophe.

Une autre dédicace très spéciale aux "dinosaures africains", mes amis de toujours : Bénédicte, Olivier et Laetitia. Qu'il est long le chemin des rizières de Bouaké aux marches de l'amphitéatre Laurent Schwartz...

Ces remerciements ne seraient pas complets si je n'évoquais pas le soutien constant de mes parents, la solidarité de Philippe & Céline et Aude, et les compétences de Joseph.

Un grand merci à "Mamie", toujours présente pour les grandes occasions, comme pour les plus petites.

Famille, collègues, ou amis, les mots et la place manquent pour exprimer l'ampleur de ma reconnaissance. Il est de ces moments où l'on se sent redevable à la terre entière, puisse la vie me faire le plaisir de vous rendre un peu de ce bonheur que vous m'avez transmis.

Guillaume.

Table des matières

Liste des figures

Liste des tableaux

Introduction

1 Introduction à l'analyse des mélanges gaussiens

Depuis la première tentative d'analyse d'un modèle de mélange par Pearson [1894], l'étude des mélanges de lois est devenue un domaine à part entière de la statistique moderne. De nombreux ouvrages de références existent sur le sujet, le plus récent est McLachlan et Peel [2000], faisant un "état de l'art" des différentes approches développées jusqu'à maintenant. Pour un point de vue détaillé des techniques d'analyses non-bayésiennes, on pourra se référer à Titterington et al. [1985] ainsi qu'à McLachlan et Basford [1988].

On considère ici l'étude des modèles où les données observées $y^{(n)} = (y_1, ..., y_n)$ sont considérées indépendantes et distribuées selon un mélange de lois fini, à k composantes, de densité

$$g(y; \pi, \theta) = \pi_1 f(y; \theta_1) + ... + \pi_k f(y; \theta_k),$$

où $\pi = (\pi_1, ..., \pi_k)$ sont les proportions du mélange de somme égale à un, et $\theta = (\theta_1, ..., \theta_k)$ le vecteur des paramètres de la loi f considérée. Lorsque f représente une loi normale de paramètre (μ_k, σ_k^2) on parlera de mélange gaussien. Ce type de modèle intervient dans une grande quantité de phénomènes ayant une signification physique. Lorsque c'est le cas, les composants du mélange ont une réelle importance et l'inférence sur k devient alors l'un des buts principaux de l'analyse. Les modèles de ce type peuvent permettent d'estimer une densité associée à un échantillon. Le nombre de composants du mélange permet donc d'obtenir une estimation plus ou moins lisse, même si cela n'a pas de signification pratique.

La seconde approche importante utilisant abondamment les modèles de mélanges sera appellée ici l'approche "classificatrice". Elle considère des données où l'interprétation physique des classes (c'est-à-dire des composants du mélange) constitue le but de l'analyse. On cherche alors à ranger les données selon des catégories, constituées par exemple par les composants dont nous cherchons à déterminer le nombre et les paramètres

Cette thèse s'attache dans un premier temps à l'étude d'un test d'hypothèses basé sur le maximum de vraisemblance, permettant un choix entre $k = 1$ et $k = 2$ pour un mélange de densités quelconques. Dans un second temps, nous étudions les modèles de mélanges gaussiens en se plaçant dans un cadre bayésien. Nous développons dans le cas multivarié l'algorithme de Monte Carlo par chaînes de Markov à sauts réversibles de Green [1995] permettant d'obtenir une approximation des lois a posteriori. Les méthodes bayésiennes tiennent une place de plus en plus importante, allant de pair avec la puissance des ordinateurs permettant de circonvenir à la complexité des calculs. Robert et Casella [1999] et Robert [1996a] constituent des ouvrages de références pour une introduction détaillée aux méthodes d'analyses bayésiennes. De nombreux articles paraissent régulièrement, traitants d'analyses de modèles de mélanges gaussiens par des techniques bayésiennes. Parmi les plus récents nous citerons par exemple Stephens [2000a] ou Celeux et al. [2000], abordant tout deux des sujets communs à cette thèse.

Cependant, relativement peu de travaux considèrent le cas des mélanges gaussiens multivariés, pourtant très utiles pour modéliser des images ou d'autres données complexes. Inspirés par l'approche présentée par Stephens [1997] de ce type de problèmes, nous tenterons de concilier la technique des sauts réversibles développée par Richardson et Green [1997] dans le cas des mélanges gaussiens univariés, avec la complexité induite par le cadre multivarié.

Nous précisons maintenant quelques aspects théoriques des modèles de mélanges de lois, et donnons brièvement les éléments de base de l'analyse statistique bayésienne.

2 Inférence sur les mélanges de lois

La formulation paramétrique des mélanges de lois a permis l'émergence de nombreuses techniques d'Inférence ayant pour but essentiel l'estimation. La première méthode utilisée pour ce faire, basée sur l'étude des moments, connaît un regain d'intérêt récent avec des auteurs comme Craigmile et Titterington [1997] par exemple. Titterington et al. [1985] fournissent quant à eux une revue détaillée concernant les méthodes à base de distance minimale. Il existe aussi des méthodes graphiques, mais les plus utilisées sont les méthodes bayésiennes ainsi que celles utilisant le maximum de vraisemblance. Ce sont ces deux derniers aspects que nous présentons plus particulièrement ici.

2.1 L'approche par maximum de vraisemblance

Cette approche s'intéresse à l'estimation souvent notée $\left(\hat{\pi}, \hat{\theta}\right)$ des paramètres (π, θ) en essayant de maximiser la vraisemblance

$$l\left(y^{(n)}, \pi, \theta\right) = \prod_{i=1}^{n} \left[\pi_1 f\left(y_i; \theta_1\right) + ... + \pi_k f\left(y_i; \theta_k\right)\right]. \tag{1}$$

Les estimateurs obtenus servent alors à évaluer les quantités auxquelles on s'intéresse. Les propriétés des estimateurs du maximum de vraisemblance sont maintenant bien connues et sont détaillées par exemple dans McLachlan et Peel [2000]. Cette approche permet en outre de développer des tests basés sur la statistique du rapport du maximum de vraisemblance (T.R.M.V). Certains auteurs comme Garel [1996] se sont penchés sur ce problème. Garel et Goussanou [2002] s'en servent pour tester l'homogéneité contre un mélange à 3 composants. Une analyse plus détaillée des techniques existantes se trouve au chapitre 2 traitant de ce sujet.

De multiples critères de type AIC ou BIC découlent aussi de cette approche, permettant de déterminer le nombre de composants en maximisant la vraisemblance, pondérée par le nombre de paramètres du modèle.

Bien que très populaire depuis l'utilisation généralisée de l'algorithme EM (Dempster et al. [1977]), la méthode du maximum de vraisemblance doit faire face à de gros problèmes de mise en oeuvre, autant sur le plan pratique que théorique.

Considérons par exemple le cas où $\mathcal{N}\left(y_i; \mu, \sigma^2\right)$ représente la densité d'une loi normale univariée de moyenne μ et de variance σ^2, la vraisemblance d'un mélange fini de ces lois s'écrit :

$$l\left(y^{(n)}, \pi, \mu, \sigma^2\right) = \prod_{i=1}^{n} \left[\pi_1 \mathcal{N}\left(y_i; \mu_1, \sigma_1^2\right) + ... + \pi_k \mathcal{N}\left(y_i; \mu_k, \sigma_k^2\right)\right]. \tag{2}$$

Dans certains cas (cf exemple 1), celle-ci peut alors tendre vers l'infini. Le point de l'espace des paramètres correspondant représente donc une singularité de la surface de vraisemblance, ce qui rend encore plus difficile la détection du maximum. On est alors obligé de se restreindre à des maximums locaux parmi lesquels il devient difficile de trouver un critère de choix.

Exemple 1 Dans le cas d'un mélange gaussien de la forme :

$$\pi\mathcal{N}\left(\mu_1, \sigma_1^2\right) + (1-\pi)\mathcal{N}\left(\mu_2, \sigma_2^2\right),$$

la vraisemblance d'un échantillon i.i.d. $(Y_1, ..., Y_n)$ est proportionnelle à :

$$\prod_{i=1}^{n} \left[\pi\mathcal{N}\left(y_i; \mu_1, \sigma_1^2\right) + (1-\pi)\mathcal{N}\left(y_i; \mu_2, \sigma_2^2\right)\right]. \tag{3}$$

Celle-ci contient 2^n termes. En developpant le produit (3) sous la forme

$$\pi^n \prod_{i=1}^{n} \mathcal{N}\left(y_i; \mu_1, \sigma_1^2\right) + \pi^{n-1}(1-\pi)\mathcal{N}\left(y_1; \mu_2, \sigma_2^2\right) \prod_{i=2}^{n} \mathcal{N}\left(y_i; \mu_1, \sigma_1^2\right) + ...$$

et en faisant tendre σ_2 vers 0 lorsque $\mu_2 = y_1$, on constate que la vraisemblance n'est plus bornée en σ_2.

Considérons l'approche utilisant le T.R.M.V,. développée par Garel [2001] afin de déterminer le nombre de composants d'un mélange fini. Là-aussi, le comportement de la vraisemblance devient problématique. Examinons le cas d'un test d'un échantillon issu d'un mélange à deux composantes contre un échantillon homogène avec $\mu \in [-a, a]$ et $a \in \mathbb{R}$.

Le problème consiste alors à tester l'hypothèse nulle

$$H_0 : g(y; \mu, p) = \mathcal{N}(y; 0, 1),$$

postulant une homogénéité, contre l'hypothèse alternative

$$H_1 : g(y; \mu, p) = (1-p)\mathcal{N}(y; 0, 1) + p\mathcal{N}(y; \mu, 1),$$

postulant un véritable mélange. Sous H_0, les vraies valeurs des paramètres sont représentés par l'ensemble $\{\mu = 0, p \in [0, 1]\} \cup \{p = 0, \mu \in [-a, a]\}$. Les résultats classiques sont mis en défaut puisque une des valeurs du paramètre p correspondant à l'homogénéité $(p = 0)$, se situe sur la frontière de l'espace des paramètres. La surface de vraisemblance fait de plus preuve d'une instabilité particulière lorsqu'on étudie celle-ci sous l'hypothèse H_1 alors que les données sont issues d'une $\mathcal{N}(0, 1)$.

Remarque 1 *: On réservera tout au long de la thèse la notation $\mathcal{N}\left(\mu, \sigma^2\right)$ pour les lois, et la notation $\mathcal{N}\left(y; \mu, \sigma^2\right)$ lorsqu'on fera référence à la densité correspondante, de variable y.*
Il en sera de même pour les autres lois.
Remarquons de plus que nous appellons "mélange simple" un mélange gaussien à deux composants de la forme : $(1-p)\mathcal{N}(y; 0, 1) + p\mathcal{N}(y; \mu, 1)$.

Les figures 1 et 2 permettent de situer les estimateurs du maximum de vraisemblance des vraies valeurs des paramètres. Tous les problèmes liés à l'approche par maximum de vraisemblance , même si celle-ci a fait ses preuves sur quelques mélanges simples, ont encouragé le développement de l'approche bayésienne.

On peut signaler aussi que ces problèmes de maximisation des fonctions de vraisemblance sont liés de manière très étroite à la notion d'identifiabilité.

2.2 Le modèle à données manquantes

Chaque y_i composant les données $y^{(n)}$ provient d'un des composants du mélange (par exemple le $i^{ème}$). On associe à chaque y_i une variable indicatrice $z_i \in \{1, ..., k\}$ telle que

$$\mathbb{P}(z_i = j) = \pi_j \quad \text{pour} \quad j = 1...k. \tag{4}$$

C'est-à-dire que z_i est issue d'un tirage parmi k catégories, de probabilités respectives $\pi_1, ..., \pi_k$. En notant $z^{(n)} = (z_1, ..., z_n)$, on a alors

$$\mathcal{L}oi\left(y_i \mid z^{(n)}\right) \sim f\left(y_i \mid \theta_{z_i}\right).$$

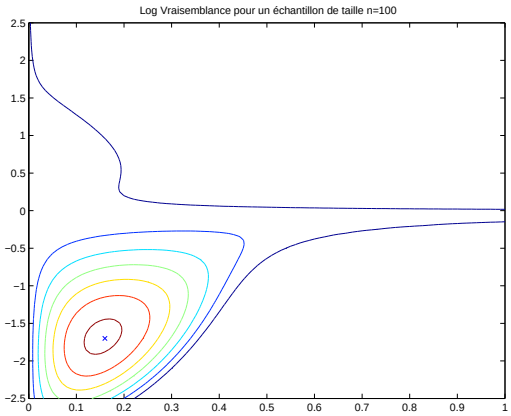


FIG. 1 – Surface de log-vraisemblance d'un échantillon homogène de taille 100 sous l'hypothèse d'un mélange simple. En abscisse se trouve la proportion p, en ordonnée la moyenne μ.

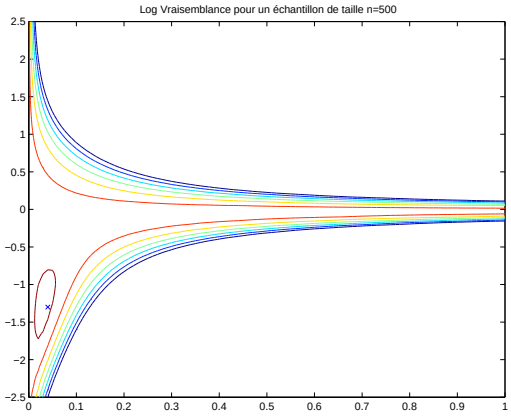


FIG. 2 – Surface de log-vraisemblance d'un échantillon homogène de taille 500 sous l'hypothèse d'un mélange simple. En abscisse se trouve la proportion p, en ordonnée la moyenne μ.

Lorsqu'on connaît les z_i, le composant auquel appartient y_i est déterminé. La connaissance de $z^{(n)}$ suffit donc à éliminer totalement la structure de mélange. On est alors en mesure d'écrire la vraisemblance des données complétées $x^{(n)} = \left(y^{(n)}, z^{(n)}\right)$:

$$\begin{aligned} l\left(x^{(n)} \mid \pi, \theta\right) &\propto \prod_{i=1}^{n} \pi_{z_i} f\left(y_i \mid \theta_{z_i}\right) \\ &= \prod_{j=1}^{k} \left(\prod_{i \diagup z_i = j} \pi_j f\left(y_i \mid \theta_j\right) \right). \end{aligned}$$

Dans une approche classificatrice, on traitera les z_i comme des variables à estimer afin de reconstruire les classes de l'échantillon. On peut signaler ici, que c'est cette même approche qui est utilisée pour l'algorithme EM, celui-ci permettant alors de maximiser la vraisemblance des données complètes ci-dessus.

2.3 Identifiabilité des mélanges

Estimer les paramètres d'une densité n'a de sens que lorsque ces paramètres sont identifiables. En général, une famille de densités paramétriques $f\left(y;\theta\right)$ est identifiable si des valeurs distinctes des paramètres, déterminent des éléments distincts de la famille de densité définie par

$$\left\{f\left(y;\theta\right) : \theta \in \Theta\right\},$$

où Θ représente l'espace des paramètres. On a l'égalité suivante

$$f\left(y;\theta\right) = f\left(y;\theta'\right), \tag{5}$$

si et seulement si $\theta = \theta'$. La notion d'identifiabilité pour les mélanges se doit d'être définie de manière légèrement différente. En effet, si l'on suppose que $f\left(y;\theta\right)$ est un mélange de deux densités $f_i\left(y;\theta_i\right)$ et $f_h\left(y;\theta_h\right)$ appartenant à la même famille paramétrique, alors l'égalité (5) est vérifiée même si les indices i et h sont échangés. On dit alors que l'identifiabilité des paramètres n'est pas respecté.

Suppossons que l'on ait un mélange de la forme

$$g\left(y;\pi,\theta\right) = \pi_1 f_1\left(y;\theta_1\right) + ... + \pi_k f_k\left(y;\theta_k\right),$$

où $f_j\left(y;\theta_j\right)$ représente la densité associée au j-ième composant du mélange. On peut alors interpréter la notion d'identifiabilité d'un mélange de la façon suivante : soit deux membres d'une même famille de densités d'un mélange

$$g\left(y;\pi,\theta\right) = \sum_{j=1}^{k} \pi_j f_j\left(y;\theta_j\right)$$

et

$$g\left(y;\pi^*,\theta^*\right) = \sum_{j=1}^{k^*} \pi_j^* f_j\left(y;\theta_j^*\right).$$

La famille de fonctions de densité est dite identifiable si on a l'égalité pour presque tout y (relativement à la mesure de référence) suivante

$$g\left(y;\pi,\theta\right) = g\left(y;\pi^*,\theta^*\right),$$

si et seulement si $k = k^*$ et que l'on peut permuter les différents composants pour avoir

$$\pi_j = \pi_j^* \text{ et } f_j(y;\theta_j) = f_j\left(y;\theta_j^*\right) \qquad (\ j = 1, ..., k).$$

Pour un point de vue détaillé sur l'identifiabilité des mélanges, on pourra se référer à Titterington et al. [1985] , section 3.1. Indiquons ici qu'on peut également caractériser l'identifiabilité de la famille $\{f(y;\theta) : \theta \in \Theta\}$ par la condition suivante :

$$\sum_{j=1}^{k} \pi_j f_j(y;\theta_j) = \sum_{j=1}^{k} \pi_j^* f_j\left(y;\theta_j^*\right) \ \lambda - ps \Longrightarrow \sum_{j=1}^{k} \pi_j \delta_{\theta_j} = \sum_{j=1}^{k} \pi_j^* \delta_{\theta_j^*}$$

où cette fois les π_j et π_j^* peuvent être nuls.

Même si le mélange est identifiable, les paramètres peuvent ne pas l'être. Si tous les $f_j(y;\theta_j)$ proviennent de la même famille paramétrique, alors le mélange est invariant sous les $k!$ permutations des paramètres associés. Ceci signifie notamment que si $\left(\hat{\pi}, \hat{\theta}\right)$ correspond à un maximum local de la vraisemblance $l\left(y^{(n)}, \pi, \theta\right)$, celle-ci aura au moins $k!$ maxima locaux correspondants à la même valeur. Si l'on note Ω_0 le sous-ensemble des paramètres tel que $g(y;\pi,\theta) = g(y;\pi_0,\theta_0)$, où (π_0, θ_0) désigne la vraie valeur du paramètre, alors Ω_0 n'est pas réduit à un seul point. Une analyse précise de la complexité topologique liée à ce type de problèmes a été réalisée par Ghosh et Sen [1985]. Les problèmes d'identifiabilité peuvent être contournés par l'imposition d'une contrainte sur les paramètres du mélange appelée condition de séparation. Cette dernière est en général imposée sur les proportions du mélange, c'est-à-dire

$$\pi_1 \leqslant \pi_2 \leqslant ... \leqslant \pi_k,$$

mais dans le cas des mélanges gaussiens univariés, on peut aussi considérer que

$$\mu_1 \leqslant \mu_2 \leqslant ... \leqslant \mu_k.$$

L'imposition d'une telle contrainte n'est pas nécessaire lorsque l'estimation d'un modèle de mélange est effectuée via l'algorithme EM. Cependant, la non-identifiabilité peut poser de gros problèmes dans un cadre bayésien utilisant l'analyse de la loi a posteriori. Ce problème particulier est appelé le "label switching". Les contraintes d'ordre sont souvent inappropriées dans un cadre multivarié, et ont de plus tendance à contraindre les paramètres de manière préjudiciable à l'analyse. Richardson et Green [1997] ont ainsi pu mettre en évidence des comportements différents selon le type de contraintes imposées (sur les moyennes ou sur les proportions). Selon Celeux et al. [2000], la troncature imposée par la condition d'identifiabilité peut aussi contraindre les paramètres sans respecter le support de la loi a posteriori, ce qui justifie l'émergence de nouvelles méthodes tentant de s'affranchir de ces conditions.

2.4 L'approche bayésienne

2.4.1 Quelques notions de statistique bayésienne

Notons tout d'abord que nous ne considérons que le cas paramétrique, consistant à supposer que y_i suit une loi de densité $f(y_i;\theta)$. La vraisemblance de l'échantillon est alors

$$l\left(y^{(n)}, \theta\right) = \prod_{i=1}^{n} f(y_i;\theta) \quad \text{avec } \theta = (\theta_1, ..., \theta_k).$$

L'hypothèse d'indépendance permet d'effectuer l'analyse théorique en n'étudiant qu'une donnée y suivant la loi de densité f. Pour cette section nous abandonnerons donc la notation $y^{(n)}$ en notant

$$l(y,\theta) = f(y,\theta).$$

De plus, nous emploierons fréquemment la notation «y suit la loi f» ou $y \sim f$ au lieu de «y suit la loi de densité f».

Définition 1 : On appelle modèle statistique bayésien la donnée d'un modèle statistique paramétré $f(y, \theta)$, et d'une **loi a priori** sur les paramètres, $\pi(\theta)$.

La loi π représente ainsi un moyen efficace de résumer l'information a priori disponible sur le paramètre θ, ainsi que l'incertitude sur la valeur de cette information, et donc de permettre ainsi une appréciation quantitative de θ.

La statistique bayésienne permet alors d'effectuer une «inversion» de l'approche classique par la vraisemblance, en observant la **loi a posteriori** des paramètres

$$\pi(\theta \mid y) = \frac{l(y, \theta)\, \pi(\theta)}{\int l(y, \theta)\, \pi(\theta)\, d\theta}.$$

Cette loi a posteriori représente l'actualisation de l'information a priori $\pi(\theta)$, au vu de l' information apportée par les observations, représentée par $l(y, \theta)$. On est donc amené à calculer le terme

$$m(y) = \int l(y, \theta)\, \pi(\theta)\, d\theta$$

appelé **loi marginale** de la donnée y. On remarquera que la loi a posteriori est proportionnelle à la loi jointe de (θ, y)

$$\varphi(\theta, y) = l(y, \theta)\, \pi(\theta),$$

ce que l'on notera le plus souvent

$$\pi(\theta \mid y) \propto l(y, \theta)\, \pi(\theta).$$

En statistique bayésienne, le paramètre de la loi de y est donc vu comme une variable aléatoire au même titre que y. On peut donc parler de la loi du couple $\varphi(\theta, y)$, mais cela induit une perception différente du rôle du paramètre θ que l'on concrétise en notant $f(y, \theta)$ comme une densité conditionnelle, c'est-à-dire :

$$f(y \mid \theta) = f(y, \theta).$$

Une des critiques faites à la statistique bayésienne est justement due à cette proportionnalité induisant une influence de la loi a priori utilisée sur l'analyse de la loi a posteriori. En effet, lors de l'inférence sur le paramètre θ, on utilise le plus souvent la loi a posteriori dont on calcule le mode, la moyenne, la médiane ou la variance et d'autres moments d'ordre plus élevés. Il devient alors évident que le choix de la loi a priori n'est pas neutre du point de vue de l'inférence.

Plusieurs approches sont alors disponibles. Le plus souvent on utilise des lois dites "conjuguées" ayant l'avantage de simplifier les calculs analytiques. L'idée consiste à choisir l'a priori dans une famille qui reste invariante en passant à la loi posteriori.

Définition 2 Une famille $\mathcal{F}$ de lois est dite **conjuguée** si, pour tout $\pi \in \mathcal{F}$, la loi a posteriori $\pi(\theta \mid y)$ appartient également à $\mathcal{F}$.

Les lois a priori conjuguées ne peuvent être obtenues qu'avec les familles de lois exponentielles, c'est-à-dire les densités de la forme :

$$f(y \mid \theta) = C(\theta)\, h(y) \exp\left[R(\theta).T(y)\right].$$

La famille de lois conjuguée est alors donnée par (cf par exemple Robert [1996a], page 97) :

$$\pi(\theta \mid \nu, \lambda) = K(\nu, \lambda)\, e^{\theta.\nu - \lambda\psi(\theta)}, \tag{6}$$

la loi a posteriori étant $\pi(\theta \mid \nu + y, \lambda + 1)$. On peut trouver dans Robert et Casella [1999], page 31, quelques exemples de lois conjuguées usuelles. On utilisera les lois conjuguées dans toute la

suite, celles-ci étant à la fois pratiques et justifiées dans le cadre qui nous occupe. Pour plus de précisions sur cet aspect précis que sont les lois conjuguées, on pourra se référer à Raiffa et Schlaifer [1961], qui les ont introduites, ainsi qu'à Diaconis et Ylvisaker [1979] .

L'utilisation de ces lois est cependant très critiquée du coté des bayesiens eux mêmes. L'aspect informatif de ces lois influence assez clairement l'analyse des paramètres, et il est fréquent d'envisager une modélisation par des lois moins informatives (des lois a priori plus "plates" par exemple).

Cependant, l'utilisation d'un a priori impropre dans le cadre des mélanges de lois est proscrit. Nous rappelons qu'une loi $\pi(.)$ est dite **impropre** lorsque $I = \int \pi(\theta)\, d\theta$ est infini.

Dans ce cas, la loi a posteriori pour les modèles de mélanges sera elle aussi impropre. On pourra se référer à McLachlan et Peel [2000] (page 126) pour une analyse et des exemples détaillés sur ce problème, en gardant à l'esprit que la loi a posteriori peut tout de même être définie sous certaines conditions.

2.4.2 Application aux mélanges gaussiens et premiers problèmes

Le modèle utilisé pour écrire la densité d'une observation issue d'un mélange de lois gaussiennes est alors le suivant :

$$g(y \mid \pi, \theta) = \sum_{j=1}^{n} \pi_j \mathcal{N}(y \mid \theta_j), \tag{7}$$

avec

$$\begin{aligned} \theta &= (\theta_1, ..., \theta_k) \text{ et } \theta_j = \left(\mu_j, \sigma_j^2\right) \in \Theta, \\ \pi &= (\pi_1, ..., \pi_k) \text{ et } \sum_{j=1}^{k} \pi_j = 1. \end{aligned}$$

L'approche classique par les lois conjuguées abordée plus haut consiste à utiliser pour les paramètres θ_i la loi conjuguée (cf (6)) :

$$\pi(\theta_i \mid \nu_i, \lambda_i) \propto e^{\theta_i . \nu_i - \lambda_i \psi(\theta_i)},$$

avec ν_i de même dimension que θ_i. Ce type de loi est particulièrement adapté aux structures exponentielles, dont font partie les mélanges gaussiens. La loi conjuguée usuelle pour les proportions $(\pi_1, ..., \pi_k)$ est la loi de Dirichlet notée $\mathcal{D}_k(\alpha_1, ..., \alpha_k)$ de densité

$$f(\pi \mid \alpha_1, ..., \alpha_k) = \pi^D(\pi_1, ..., \pi_k) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\Gamma(\alpha_1) ... \Gamma(\alpha_k)} \pi_1^{\alpha_1 - 1} ... \pi_k^{\alpha_k - 1} \mathbb{1}_{[\sum_{i=1}^{k} \pi_i = 1]}.$$

Connaissant les lois a priori, on peut maintenant déterminer la loi a posteriori de $(\pi, \theta) = (\pi_1, ..., \pi_k, \theta_1, ..., \theta_k)$ conditionnellement aux données $y^{(n)} = (y_1, ..., y_n)$, qui s'écrit :

$$[\pi_1, ..., \pi_k, \theta_1, ..., \theta_k \mid y_1, y_2, ..., y_k] \propto$$

$$\pi^D(\pi_1, ..., \pi_k) \prod_{j=1}^{k} \pi\left(\theta_j \mid \mu_j, \lambda_j\right) \prod_{i=1}^{n} \left(\sum_{j=1}^{k} \pi_j f(y_i \mid \theta_j) \right).$$

Cette loi a posteriori se décompose alors en k^n termes, ce qui rend le calcul des espérances a posteriori rédhibitoire pour des tailles d'échantillons supérieures à quarante. Ce problème insurmontable nous oblige à modifier notre approche afin d'analyser les mélanges de distributions.

3 Structure du document

Le chapitre I présente un travail sur un test d'hypothèse permettant de décider entre l'homogéneité et un mélange simple, lorsque le paramètre est connu sous H_0.

Le chapitre II présente les méthodes de Monte Carlo par chaînes de Markov en précisant les modèles bayésiens nécessaires. Nous illustrerons le fonctionnement de l'algorithme de Gibbs lorsque le nombre de composants du mélange est fixé, dans le cas univarié et multivarié

Le chapitre III présente la méthode à sauts réversibles de manière détaillée et théorique.

Le chapitre IV s'attache quand à lui aux solutions développées par Stephens [2000b] et Celeux [1998] face au problème du label switching, très fréquent dans nos applications.

Le chapitre V constitue l'application de ce dernier algorithme au cas des mélanges gaussiens multivariés. On définit les mouvements modifiant la dimension du modèle et on détaille l'expression de leur probabilité d'acceptation.

Le chapitre VI présente les résultats de l'algorithme sur quelques jeux de données, en analysant précisément l'influence des lois a priori sur la qualité d'estimation des paramètres et du nombre de composants.

Le dernier chapitre est constitué par un article publié dans la Revue de Statistique Appliquée, faisant suite à un contrat effectué avec les Services Techniques de la Navigation Aérienne. Ce travail nous a permis de financer une partie de notre recherche.

En annexe se trouvent un récapitulatif des lois et notations utilisées, ainsi qu'une partie des programmes présentés dans le document.

Chapitre I

Test d'homogéneité contre mélange simple dans un cadre général.

1 Introduction

Le travail présenté ici fait suite à mon mémoire de DEA de Mathématiques appliquées présenté en Juin 1999 sous la direction de Bernard Garel. Il apporte plusieurs compléments à la thèse de Franck Goussanou soutenue en septembre 2001 à l'Université Paul Sabatier.

Tester le nombre de composants k d'un mélange gaussien constitue un problème majeur qui n'est pas encore complètement résolu. Nous avons déjà abordé le fait que les mélanges gaussiens sont principalement utilisés pour deux buts principaux.

En premier lieu, ils permettent de proposer un cadre de travail alternatif aux méthodes d'estimation à noyaux de la densité, en fournissant un moyen de modéliser une densité quelconque (Escobar et West [1995], Robert [1996b]).

Dans ce cadre, les critères de type AIC et BIC sont largement utilisés et ont été étudiés par Biernacki et al. [2000] et Solka et al. [1998]. L'influence du nombre de composants sur l'apparence de la densité n'a que peu d'impact lorsque ce nombre est grand, ce qui est souvent le cas dans ce contexte. Ces critères, mêmes s'ils surestiment parfois le nombre de composants, semblent donc convenir pour ce premier domaine (Roeder et Wasserman [1997]).

Pour le second domaine, c'est à dire l'approche classificatrice, le nombre de composants possède souvent une explication physique (réelle ou supposée), avec laquelle on ne peut négocier. Le nombre de composants du mélange est alors primordial, et ce chapitre s'intéresse précisément à ce domaine.

Si l'on excepte les techniques bayesiennes développées dans les chapitres suivants, les deux méthodes les plus usitées permettant de déterminer le nombre de composants utilisent principalement des techniques liées à la vraisemblance.

Les critères classiques de type AIC ou BIC, basés sur une vraisemblance pénalisée, sont moins demandeurs en calculs que la statistique du rapport des maximums de vraisemblance qui requiert parfois des méthodes de bootstrapping afin d'obtenir des P-values. Ils ne produisent cependant pas de quantités permettant d'évaluer la confiance dans le résultat obtenu.

L'approche adoptée ici est de formaliser le problème sous forme de test d'hypothèses paramétriques en se basant sur la statistique du rapport des maximums de vraisemblance. Le problème général est le suivant : on veut tester l'hypothèse nulle

$$H_0 : g = \sum_{i=1}^{k_1} \pi_i f(\cdot; \theta_i)$$

contre l'hypothèse alternative

$$H_1 : g = \sum_{i=1}^{k_2} \pi_i f(\cdot, \theta_i) \qquad \text{avec} \quad k_1 < k_2.$$

On note Θ_0 (resp. Θ_1) le sous-ensemble de l'espace des paramètres correspondant à H_0 (resp. H_1). La statistique du test du rapport des maximums de vraisemblance (TRMV) est donnée par

$$-2\log(\lambda_n) = 2\left\{\sup_{\theta\in\Theta_0\cup\Theta_1} L_n(X_1,\cdots,X_n,\theta) - \sup_{\theta\in\Theta_0} L_n(X_1,\cdots,X_n,\theta)\right\}.$$

On sait depuis les résultats de Wilks [1938] et de Chernoff [1954] que, sous des conditions de régularité, cette statistique sous H_0 suit asymptotiquement un chi-deux dont le degré de liberté est égal à la différence du nombre de paramètres entre les deux hypothèses. Cependant, pour les modèles de mélange, les conditions de régularité ne sont plus vérifiées puisque sous l'hypothèse nulle, les proportions sont sur la frontière de l'espace des paramètres, et les paramètres ne sont pas identifiables. Pour s'en convaincre, on peut considérer le problème de test suivant: tester

$$H_0 : g\left(y;\mu,\sigma^2\right) = \mathcal{N}\left(y;\mu_{10},1\right)$$

contre l'hypothèse alternative

$$H_1 : g\left(y;\mu,\sigma^2\right) = (1-\pi)\mathcal{N}\left(y;\mu_{10},1\right) + \pi\mathcal{N}\left(y;\mu_2,1\right).$$

L'hypothèse nulle est alors représentée (figure I.1) par la réunion des trois segments de droite suivants:

$$H_0^1 : \pi = 0, \mu_1 = \mu_{10}, \mu_2 \text{ quelconque}$$
$$H_0^2 : \pi = 1, \mu_2 = \mu_{10}, \mu_1 \text{ quelconque}$$
$$H_0^2 : \pi \in]0,1[, \mu_1 = \mu_2 = \mu_{10.}$$

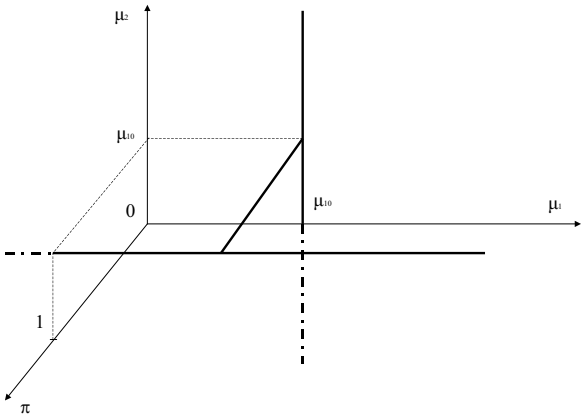


Figure I.1: représentation graphique de l'hypothèse nulle précédente.

On peut signaler aussi, que lorsque la vraie valeur du paramètre est connue sous H_0 et que l'espace des paramètres pour l'hypothèse alternative est non borné, Hartigan [1985] a montré qu'une statistique proche de celle du test du TRMV tend vers l'infini. Ce résultat, confirmé par Bickel et Chernoff [1995], nous impose de considérer que les paramètres sont contenus dans un compact de $\mathbb{R}^k$. Signalons que l'équivalence de la statistique de Hartigan avec celle du rapport des maximums de vraisemblance n'a été démontrée que très récemment (Liu et Shao [2003]).

Beaucoup de conjectures ont été avancées pour résoudre ce problème de test dans certains cas particuliers. Wolfe [1971] avance que la statistique du TRMV suit un χ^2_2 dans le cas d'un mélange portant sur les moyennes avec des variances communes, et un χ^2_4 dans le cas d'un mélange Gaussien portant à la fois sur les moyennes et les variances. Ce résultat ayant été mis en cause par plusieurs auteurs, notamment Everitt [1981] et McLachlan [1987], le problème est resté ouvert jusqu'à récemment.

Ghosh et Sen [1985] sont les premiers à écrire l'expression de la loi asymptotique dans le cadre d'un test d'homogéneité, en imposant une condition de séparation entre les paramètres. Berdaï et Garel [1996] ont affiné et étendu ce résultat au cas où la variance est elle aussi inconnue. En utilisant des techniques de reparamètrisation, Lemdani et Pons [1999] montrent que la statistique du TRMV est le sup du carré d'un processus gaussien. Dacunha-Castelle et Gassiat [1997] obtiennent le même résultat en utilisant une reparamètrisation localement conique. Leur travail, très théorique, nécessite cependant l'existence de différentielles à l'ordre 5. Encore une autre technique, nommée "sandwich method", est employée par Chen et Chen [2001]. Une avancée importante est cependant effectuée par Garel [2001] qui parvient à lever la condition de séparation introduite par Ghosh et Sen [1985] en utilisant une technique de Log-vraisemblance modifiée appliquée à sept types différents de mélanges gaussiens. Le cas général est traité dans Garel [2003]. Ces techniques, toujours appliquées aux mélanges gaussiens, ont été détaillées par Garel et Goussanou [2002] dans le cas d'un test d'homogénéité contre un mélange à trois composants.

Nous nous sommes intéressés, avec ce travail, à la généralisation de cette dernière approche aux cas des mélanges simples de lois quelconques, mais identiques. Nous donnons ici les hypothèses à effectuer sur la distribution en question dans le cas d'un test entre homogéneité et mélange simple, c'est-à-dire lorsque la vraie valeur du paramètre est connue sous H_0. Nous montrons ensuite que ces hypothèses sont vérifiées dans le cas des mélanges gaussiens, en détaillant le cas particulier où l'hypothèse alternative est celle d'un mélange gaussien portant sur les variances. En s'inspirant des récents travaux de Delmas [2001] nous abordons le calcul des valeurs critiques de ce test. Pour cela, nous comparons la méthode utilisant la borne due à Davies [1977] à celle utilisant la borne de Rice détaillée et améliorée par Delmas [2001].

L'obtention de nouvelles tabulations à l'aide d'un programme en Fortran nous permet d'appliquer cette théorie à des données épidémiologiques réelles.

2 Testing homogeneity against a simple mixture : the general case.

GUILLAUME SAINT PIERRE
Laboratoire de Statistique et Probabilités, Université Paul Sabatier, France

BERNARD GAREL
Stochastic models, LEN7, ENSEEIHT, France

FRANCK GOUSSANOU
Stochastic models, LEN7, ENSEEIHT, France

ABSTRACT. We consider the problem of testing homogeneity, that is testing no mixture, against a simple mixture. We give the expression of the likelihood ratio test statistic for the general case, under suitable conditions. We also give some percentiles points obtained using Davies bound. An application to Gaussian mixture of the variances is given. We illustrate the preceding results on a real data set.

Key words: Asymptotic distribution; Likelihood ratio test statistic; Mixture of two univariate normals.

2.1 Introduction

Because of their usefulness as an extremely flexible method of modeling a wide variety of random phenomena, finite mixture models have been arising strong and sustained interest for the last few decades, and especially these last few years.

From a theoretical point of view, statistical analysis of mixtures involves several topics such as stochastic processes, estimation and maximization, bayesian analysis and hidden Markov chains. An overview of the problems associated to finite mixture distribution can be found in the books by Titterington, Smith and Makov (1985) or, more recently, McLachlan and Peel (2000). In this paper we will concentrate on the testing problem.

For parametric hypotheses testing problems it is customary to use the likelihood ratio as a test statistic. Under standard regularity conditions, a classical result of Chernoff (1954) states that if the null hypothesis is true, the likelihood ratio test statistic (LRTS), asymptotically has a χ^2- distribution.

In the case of mixtures, however, the regularity conditions are not satisfied. The first reason is that the null hypothesis lies on the boundary of the parameter space, whereas the standard regularity conditions require it to be in the interior. Secondly, important problems of identifiability arise, which deeply modifies the maximization problem. Li and Sedransk (1988) gave topological study of those identifiability problems, introducing a bayesian type approach.

Moreover, classical results break down when the general model is used under homogeneity, causing maximum likelihood estimator's (MLE's) for some parameters to be inconsistent. The above caused Cheng and Traylor (1995) to identify the mixture model as one of the four non-regular parametric models. In the same way, as noticed by Lindsay (1995), the large sample behavior of the LRTS for testing homogeneity against a mixture has long been a mystery.

A lot of theoretical conjectures and simulation studies for specific cases have been published for this problem. An approximation of the asymptotic distribution of the LRTS is given by Wolfe (1971). It gives a χ^2_2 distribution for a normal mixture of the mean with common variance and a χ^2_4 distribution for a normal mixture with different means and different variances.

The aim of this paper is to study the problem of testing homogeneity against a simple mixture without separation condition in the general case.

In the Gaussian case, an asymptotic study has been performed by Hartigan (1985), showing that a statistic close to the LRTS tends to infinity with probability one if the mean parameter is unbounded. This has been further detailed by Bickel and Chernoff (1993). They showed that if the parameter space is unbounded, the LRTS approaches infinity as $\log(\log n))$.

With an assumption of a bounded parameters set, Ghosh and Sen (1985) gave the first version of the asymptotic distribution of the LRTS for testing homogeneity. Unfortunately they need to impose a separation condition between θ_1, the parameter of the first component, and θ_2 the parameter of the second component : $|\theta_2 - \theta_1| > \varepsilon > 0$.

Lemdani and Pons (1999) used a reparametrization to investigate the testing problem when the parameter is known under H_0. Another way to deal with the separation condition is presented by Chen and Chen (2001). They call it the sandwich method. These studies show that the good way to deal with the separation condition in the general case has not yet been found.

An important stage has been undertaken by Garel (2001) with the derivation of the LRTS for no fewer than seven Gaussian two component mixture cases without any separation condition, and Garel & Goussanou (2002) for the three component Gaussian mixture on the means.

Assuming that the true value of the parameter is known under H_0, we show that the separation condition can be removed with *a general density* and *assumptions on the first derivatives only.* We use the result of Redner (1981), saying that if W denotes a fixed neighborhood of the curves in the parameter space corresponding to H_0, then the probability that the maximum likelihood estimator is found in W tends to one when n goes to infinity.

In the next section, we give the assumptions and the main result. Then in section 3 we apply these results to the Gaussian simple mixture on the variances. Section 4 is devoted to the

computation of the percentiles points and illustrations on a real data set. Finally, the proofs are gathered in section 5.

2.2 Main results

We want to test the hypothesis H_0 of homogeneity against the hypothesis H_1 of a mixture of two densities. Let $X_1, \ldots, X_n$, be n independent and identically distributed random variables with density $f(x, p, \theta)$. Under H_0 we have

$$f(x, p, \theta) = g(x, \theta_{10}),$$

where θ_{10} is the true value of the parameter, that we assume equal to zero without loss of generality. Under H_1 we have

$$f(x, p, \theta) = (1-p)g(x, \theta_{10}) + pg(x, \theta), \tag{I.1}$$

where g is a probability density function satisfying some regularity conditions (see assumption 1). We call (I.1) a simple mixture. We assume that θ belongs to $\Theta =]-a, b[\subset \mathbb{R}$ and we denote by $\bar{\Theta}$ its closure, with $0 < a, b < +\infty$ and $p \in [0, 1]$. Let us remark that a single element of the null hypothesis of homogeneity is represented by two curves corresponding to $\{p = 0, \theta \in \Theta\}$ and $\{p \in [0, 1], \theta = 0\}$. All the expectations in the rest of the paper are considered under H_0. Let us denote by $\mathcal{F}$ the following family of distributions :

$$\mathcal{F} = \left\{(1-p)g(x, \theta_{10}) + pg(x, \theta) \mid x \in \mathbb{R},\ p \in [0, 1],\ \theta \in \bar{\Theta} \subset \mathbb{R}\right\}.$$

We assume that $\mathcal{F}$ is identifiable in the sense of Teicher (1963). This does not mean that the parameters are identifiable. If we denote $L_n(x_1, ..., x_n, p, \theta)$ the log likelihood function

$$L_n(x_1, ..., x_n, p, \theta) = \sum_{i=1}^{n} \log f(x_i, p, \theta),$$

the classical likelihood ratio test statistic is given here by

$$-2 \log \lambda_n = 2 \sup_{\theta \in \Theta} [L_n(\theta) - L_n(0)],$$

with

$$L_n(\theta) = \sup_{p \in [0,1]} L_n(p, \theta) \quad \text{and} \quad L_n(0) = \sum_{i=1}^{n} \log g(x_i, 0).$$

Before going further, we need to assume some conditions on the distribution g.

Assumption 1 : : For each $i \in \{1, ..., n\}$ and for each $\theta \in \bar{\Theta}$, $g(x_i, \theta)$ denotes any representative of the density of X_i with respect to the Lebesgue measure. We assume:

1. the existence of the first order derivative of g with respect to θ,
2. the continuity of $\frac{\partial g}{\partial \theta}(x, .) = g'_\theta(x, .)$ for almost every x.
3. for all $\theta \in \Theta$, $E\left|\frac{g'_\theta(X,\theta)}{g(X)}\right|^2 < +\infty$.

We will write $g(x,\theta_{10}) = g(x,0) = g(x)$ and $g'_\theta(x,0) = g'_\theta(x)$. Before stating the next assumption, we need to define $N_r(p,\theta)$ the open ball of radius r about $(p,\theta) \in [0,1]\times\Theta$, and $\bar{N}_r(p,\theta)$ its closure. Moreover, we denote $V_s(H_0)$ an open neighborhood of the two curves defining H_0. Remind that $V_s(H_0)$ can be viewed as a finite union of $N_s(p_{0i},\theta_{0i})$ with $(p_{0i},\theta_{0i}) \in \{p=0,\theta\in\Theta\}\cup\{p\in[0,1],\theta=0\}$.

Assumption 2 : : for each (p,θ) and for sufficiently small r and sufficiently large s, $\sup_{(p,\theta)\in N_r(p,\theta)} f(x,p,\theta)$ is measurable and

1. $\int \log\left[\max\left(1, \sup_{(p',\theta')\in\bar{N}_r(p,\theta)} f(x,p',\theta')\right)\right] g(x)\,dx < +\infty$,
2. $\int \log\left[\max\left(1, \sup_{(p',\theta')\notin V_s(H_0)} f(x,p',\theta')\right)\right] g(x)\,dx < +\infty$,
3. $\mathbb{E}\left|\log g(x)\right| < +\infty$.

This assumption is needed to apply Redner's result. In order to derive the LRTS for this general testing problem, we have to introduce the following notation. We define $Y_i(\theta)$ for $\theta \neq 0$ as

$$Y_i(\theta) = \frac{1}{\theta}\frac{\partial \log f}{\partial p}(X_i,0,\theta) = \frac{1}{\theta}\left[\frac{g(X_i,\theta)}{g(X_i)} - 1\right], \tag{I.2}$$

and $Y_i(0) = \frac{g'_\theta(X_i)}{g(X_i)}$. We notice that we can rewrite $Y_i(\theta)$ using a Taylor expansion. There exists $\eta \in]0,1[$ such that:

$$g(x,\theta) = g(x) + \theta g'_\theta(x,\eta\theta). \tag{I.3}$$

So we can find $\eta\in]0,1[$ such that $Y_i(\theta) = \frac{g'_\theta(X_i,\eta\theta)}{g(X_i)}$.

Assumption 3 : : For all $\theta\in\bar{\Theta}$, we assume $\mathbb{E}(Y_i(\theta)) = 0$, and for all $\theta'\in\bar{\Theta}$, $(\theta\neq\theta')$:

$$\left|Y_i(\theta) - Y_i(\theta')\right| \leqslant \left|\theta-\theta'\right| H(X_i),$$

where H is square integrable.

Assumption 3 implies $\mathbb{E}\left|Y_i(\theta) - Y_i(\theta')\right|^2 \leqslant K\left|\theta-\theta'\right|^2$. Moreover, as we have

$$\mathbb{E}\left|n^{-\frac{1}{2}}\sum_{i=1}^n Y_i(\theta) - n^{-\frac{1}{2}}\sum_{i=1}^n Y_i(\theta')\right|^2 = \frac{1}{n}\mathbb{E}\left|\sum_{i=1}^n\left(Y_i(\theta) - Y_i(\theta')\right)\right|^2 = \mathbb{E}\left|Y_1(\theta) - Y_1(\theta')\right|^2,$$

then $n^{-\frac{1}{2}}\sum_{i=1}^n Y_i(\theta)$ is tight (see Billingsley 1968).

Assumption 4 : : We assume that there exists a positive constant C such that for all $\theta\in\bar{\Theta}$ we have

$$\mathbb{E}(Y_1(\theta))^2 > C > 0.$$

Assumption 5 : : We assume that $\mathbb{E}\left(\sup_{\theta\in\bar{\Theta}} Y_1^2(\theta)\right) < +\infty$.

Using classical inequalities and the Corollary 7.10 of Ledoux and Talagrand (1991), we can show that Assumption 5 implies the uniform strong law of large numbers for $\frac{1}{n}\sum_{i=1}^{n} Y_i^2(\theta)$ in the sense that

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} Y_i^2(\theta) - \mathbb{E}\left(Y_1^2(\theta)\right)\right| \longrightarrow 0$$

almost surely under H_0. The last assumption is needed to control the likelihood ratio.

Assumption 6 : : We assume that

$$\lim_{\delta\to 0}\mathbb{E}\left[\sup_{\substack{|\theta|<\delta\\ p\in[0,1]}}\left|Y_1^2(\theta) - g_1(X_1,p,\theta)\right|\right] = 0$$

with

$$g_1(x_i,p,\theta) = \left[\frac{1}{\theta}\frac{g(x_i,\theta)-g(x_i)}{(1-p)\,g(x_i)+pg(x_i,\theta)}\right]^2.$$

Under the previous assumptions, we obtain the following result.

Theorem 1 : *The likelihood ratio test statistic for testing H_0 (homogeneity) against H_1 (simple mixture), without any separation condition on the parameter is given by*

$$-2\log\lambda_n = \sup_{\theta\in\bar{\Theta}\setminus\{0\}} T_n^2(\theta)\,\mathbb{1}_{[T_n(\theta)>0]} + o_p(1),$$

where

$$T_n(\theta) = n^{-\frac{1}{2}}\frac{\sum_{i=1}^{n}\left(\frac{g(X_i,\theta)}{g(X_i)}-1\right)}{\left(\mathbb{E}\left(\frac{g(X_1,\theta)}{g(X_1)}-1\right)^2\right)^{\frac{1}{2}}} \quad \text{if } \theta\neq 0,$$

and

$$T_n^2(0) = \lim_{\theta\longrightarrow 0} T_n^2(\theta) = n^{-1}\frac{\left[\sum_{i=1}^{n}\left(\frac{g'_\theta(X_i)}{g(X_i)}\right)\right]^2}{\mathbb{E}\left(\frac{g'_\theta(X_1)}{g(X_1)}\right)^2},$$

where $o_p(1)$ is a quantity which tends to 0 as n tends to infinity uniformly with respect to $\theta\in\Theta$. The process $T_n(\theta)$ is discontinuous at $\theta = 0$.

Theorem 2 : *In $C[0,a]$ (continuous path processes on $[0,a]$) the process*

$$\tilde{T}_n(\theta) = \begin{cases} T_n(\theta) & \text{if } \theta\neq 0 \\ n^{-\frac{1}{2}}\dfrac{\sum_{i=1}^{n}\left(\frac{g'_\theta(X_i)}{g(X_i)}\right)}{\left(\mathbb{E}\left(\frac{g'_\theta(X_1)}{g(X_1)}\right)^2\right)^{\frac{1}{2}}} & \end{cases},$$

converges weakly to a centered Gaussian process with unit variance. The same result holds in $C[-a,0]$ for

$$\breve{T}_n(\theta) = \begin{cases} T_n(\theta) & \text{if } \theta\neq 0 \\ -n^{-\frac{1}{2}}\dfrac{\sum_{i=1}^{n}\left(\frac{g'_\theta(X_i)}{g(X_i)}\right)}{\left(\mathbb{E}\left(\frac{g'_\theta(X_1)}{g(X_1)}\right)^2\right)^{\frac{1}{2}}} & \end{cases}.$$

2.3 An application to simple Gaussian mixture on the variances

In this section, we will give an application of Theorem 1 to Gaussian mixtures of variances. We replace the generic function $g(x,\theta)$ by $g\left(x,\sigma^2-1\right)$, the density of a centered Gaussian random variable with variance σ^2. Notice that $\left(\sigma^2-1\right)$ replaces the previous parameter θ. The parameter σ^2 lies in a compact set $\bar{\Theta}=[a,A]\subset\mathbb{R}$ with $0<a<1<A<+\infty$. Under H_0, the true parameter is $\sigma^2_{10}=1$ and we write $g(x,\sigma^2_{10}-1)=g(x)$. The aim is to test the hypothesis of homogeneity H_0

$$f(x,p,\theta)=g(x)=\phi(x),$$

where ϕ denotes the probability density function of the standard Gaussian distribution, against the hypothesis H_1 of a simple mixture

$$f(x,p,\theta)=(1-p)g(x)+pg(x,\sigma^2-1). \tag{I.4}$$

In order to guarantee the square integrability of $\frac{\partial \log f}{\partial p}(x,p,\theta)$, we need to impose the following condition. There exists a positive constant c_1 such that $\sigma^2\leqslant 2-c_1$. Therefore we denote $2-c_1$ by A. In this case of Gaussian mixtures of the variances we have:

$$Y_i\left(\sigma^2\right)=\frac{1}{\sigma^2-1}\left[\frac{1}{\sigma}e^{\frac{X_i^2}{2}\left(1-\frac{1}{\sigma^2}\right)}-1\right],$$

and, using the same notation as in the second section,

$$g'_{\sigma^2}(X_i)=\left(\frac{X_i^2-1}{2}\right)g(X_i).$$

In order to obtain the main theorem for this case, we have to check the five needed assumptions. Assumption 1 represents the minimal regularity conditions to handle the distribution g which, of course, are satisfied in the Gaussian case. Redner (1981) showed that all the assumptions needed for his result are fulfilled by the family of normal distributions, and so, Assumption 2 is satisfied. At the end of this section, we prove that Assumptions 3, 5 and 6 are satisfied. Let us remark that Assumption 4 is fulfilled since for $\sigma^2\in[a,A]$

$$\mathbb{E}\left(Y_1\left(\sigma^2\right)\right)^2=\frac{\frac{1}{\sqrt{2\sigma^2-\sigma^4}}-1}{\left(\sigma^2-1\right)^2}=\frac{1}{2}+\frac{3}{8}\left(\sigma^2-1\right)^2+O\left(\sigma^2-1\right)^4\geqslant\frac{1}{2}.$$

Therefore, all of the six assumptions are satisfied in our situation and we have the following theorem:

Theorem 3 : *The likelihood ratio test statistic for testing H_0 (homogeneity) against H_1 (simple mixture), without any separation condition on the parameter is given by*

$$-2\log\lambda_n=\sup_{\sigma^2\in[a,1[\cup]1,A]}T_n^2\left(\sigma^2\right)\mathbb{1}_{[T_n(\sigma^2)>0]}+o_p(1),$$

where

$$T_n\left(\sigma^2\right)=n^{-\frac{1}{2}}\frac{\sum_{i=1}^n\left(\frac{1}{\sigma}e^{\frac{X_i^2}{2}\left(1-\frac{1}{\sigma^2}\right)}-1\right)}{\left(\frac{1}{\sqrt{(2\sigma^2-\sigma^4)}}-1\right)^{\frac{1}{2}}}\quad\text{if }\sigma^2\neq 1\ ,$$

and

$$\lim_{\sigma^2\longrightarrow 1}T_n^2\left(\sigma^2\right)=n^{-1}\left(\sum_{i=1}^n\frac{X_i^2-1}{\sqrt{2}}\right)^2,$$

where $o_p(1)$ *is a quantity which tends to* 0 *as* n *tends to infinity uniformly with respect to* $\sigma^2 \in \bar{\Theta}$. *Moreover,* $-2\log\lambda_n$ *converges to* $\sup_{\sigma^2\in[a,1[\cup]1,A]} T^2\left(\sigma^2\right)\mathbb{1}_{[T(\sigma^2)>0]}$, *where* T *is a centered Gaussian process with unit variance and the same autocovariance function* γ *as* T_n *:*

$$\gamma(\xi,\eta) = \frac{\frac{1}{\sqrt{\xi+\eta-\xi\eta}} - 1}{\sqrt{\frac{1}{\sqrt{(2\xi-\xi^2)}} - 1}\sqrt{\frac{1}{\sqrt{(2\eta-\eta^2)}} - 1}}, \tag{I.5}$$

and

$$\lim_{\substack{(\xi,\eta)\longrightarrow(1,1)\\(\xi-1)(\eta-1)>0}} \gamma(\xi,\eta) = 1 \qquad \lim_{\substack{(\xi,\eta)\longrightarrow(1,1)\\(\xi-1)(\eta-1)<0}} \gamma(\xi,\eta) = -1.$$

Proof ***of Assumption 3:***

First we observe that the Y_i *are centered :*

$$\mathbb{E}\left|Y_i\left(\sigma^2\right)\right| = \mathbb{E}\left[\frac{\frac{1}{\sigma}e^{\frac{X_i^2}{2}\left(1-\frac{1}{\sigma^2}\right)} - 1}{\left(\sigma^2-1\right)}\right] = \frac{1}{\sigma^2-1}\int_{\mathbb{R}}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x_i^2}{2\sigma^2}}dx_i - \frac{1}{\sigma^2-1} = 0.$$

By a Taylor expansion we have

$$y_i\left(\sigma_2^2\right) - y_i\left(\sigma_1^2\right) = \left(\sigma_2^2-\sigma_1^2\right)\frac{\partial y_i}{\partial\sigma^2}\left(\tilde{\sigma}^2\right),$$

with $\tilde{\sigma}^2$ *between* σ_1^2 *and* σ_2^2 *and*

$$\frac{\partial y_i}{\partial\sigma^2}\left(\sigma^2\right) = \frac{\left[\frac{1}{\sigma}e^{\frac{x_i^2}{2}\left(1-\frac{1}{\sigma^2}\right)}\right]\left[\left(-\frac{1}{2\sigma^2}+\frac{x_i^2}{2\sigma^4}\right)\left(\sigma^2-1\right)-1\right]+1}{\left(\sigma^2-1\right)^2}.$$

Let us denote

$$h\left(\sigma^2\right) = \left[\frac{1}{\sigma}e^{\frac{x_i^2}{2}\left(1-\frac{1}{\sigma^2}\right)}\right]\left[\left(-\frac{1}{2\sigma^2}+\frac{x_i^2}{2\sigma^4}\right)\left(\sigma^2-1\right)-1\right]+1.$$

We remark that $h(1) = 0$ *and* $h'(1) = 0$. *So we have*

$$\begin{aligned} h\left(\sigma^2\right) &= h(1) + \left(\sigma^2-1\right)h'(1) + \frac{\left(\sigma^2-1\right)^2}{2}h''\left(\bar{\sigma}^2\right) \\ &= \frac{\left(\sigma^2-1\right)^2}{2}h''\left(\bar{\sigma}^2\right), \end{aligned}$$

where $\bar{\sigma}^2$ *lies between* 1 *and* σ^2, *and* h', h'' *are the first and the second derivatives respectively of* h *with respect to* σ^2. *Finally we get*

$$\begin{aligned} \frac{\partial y_i}{\partial\sigma^2}\left(\tilde{\sigma}^2\right) &= \frac{h''\left(\bar{\sigma}^2\right)}{2} = \frac{1}{16}\frac{e^{\frac{x_i^2}{2}\left(1-\frac{1}{\bar{\sigma}^2}\right)}}{\bar{\sigma}} \\ &\quad\times\frac{15\bar{\sigma}^6+33x_i^2\bar{\sigma}^6-45x_i^2\bar{\sigma}^4-13x_i^4\bar{\sigma}^4+15x_i^4\bar{\sigma}^2-9\bar{\sigma}^8+x_i^6\bar{\sigma}^2-x_i^6}{\bar{\sigma}^{12}}, \end{aligned}$$

with $\bar{\sigma}^2$ *between* 1 *and* $\tilde{\sigma}^2$. *Remember that we have assumed* $\sigma^2 \leqslant A$ *and, since* σ^2 *lies in a compact set, we can find a positive and square integrable function* Ψ *proportional to* $\exp\left(\frac{x_i^2}{2}\left(1-\frac{1}{A}\right)\right)\left(1+x_i^2+x_i^4+x_i^6\right)$ *such that:*

$$\left|\frac{\partial Y_i}{\partial\sigma^2}\left(\tilde{\sigma}^2\right)\right| \leqslant \Psi\left(X_i\right).$$

Furthermore we get:

$$\left|Y_i\left(\sigma_2^2\right) - Y_i\left(\sigma_1^2\right)\right| \leqslant \left|\sigma_2^2 - \sigma_1^2\right| \Psi\left(X_i\right).$$

■

Proof *of Assumption 5:*

We have

$$\sup_{\sigma^2\in[a,A]} \left[\frac{\frac{1}{\sigma} e^{\frac{X_i^2}{2}\left(1-\frac{1}{\sigma^2}\right)} - 1}{\left(\sigma^2-1\right)}\right]^2 = \sup_{\sigma^2\in[a,A]} \frac{\frac{1}{\sigma^2} e^{X_i^2\left(1-\frac{1}{\sigma^2}\right)} - \frac{2}{\sigma} e^{\frac{X_i^2}{2}\left(1-\frac{1}{\sigma^2}\right)} + 1}{\left(\sigma^2-1\right)^2}.$$

Let us observe that the numerator of the preceding ratio and its derivative are equal to zero for $\sigma^2 = 1$. So, with the help of a second order Taylor expansion of this numerator around $\sigma^2 = 1$, we obtain:

$$\begin{aligned} &\sup_{\sigma^2\in[a,A]} \left[\frac{\frac{1}{\sigma} e^{\frac{X_i^2}{2}\left(1-\frac{1}{\sigma^2}\right)} - 1}{\left(\sigma^2-1\right)}\right]^2 \\ \leqslant\ & \frac{1}{2} \sup_{\tilde{\sigma}^2\in[a,A]} \left[e^{X_i^2\left(1-\frac{1}{\tilde{\sigma}^2}\right)} \left(\frac{2}{\tilde{\sigma}^6} - \frac{4X_i^2}{\tilde{\sigma}^8} + \frac{X_i^4}{\tilde{\sigma}^{10}}\right) + e^{\frac{X_i^2}{2}\left(1-\frac{1}{\tilde{\sigma}^2}\right)} \left(\frac{-3}{2\tilde{\sigma}^5} + \frac{3X_i^2}{\tilde{\sigma}^7} - \frac{X_i^4}{\tilde{\sigma}^9}\right)\right], \end{aligned}$$

where $\tilde{\sigma}^2$ lies between σ^2 and 1. As $\tilde{\sigma}^2 \in [a, A]$ with $A < 2$, we conclude with

$$\mathbb{E}\left\{\sup_{\tilde{\sigma}^2\in[a,A]} \left[e^{X_i^2\left(1-\frac{1}{\tilde{\sigma}^2}\right)} \left(\frac{2}{\tilde{\sigma}^6} - \frac{4X_i^2}{\tilde{\sigma}^8} + \frac{X_i^4}{\tilde{\sigma}^{10}}\right) + e^{\frac{X_i^2}{2}\left(1-\frac{1}{\tilde{\sigma}^2}\right)} \left(\frac{-3}{2\tilde{\sigma}^5} + \frac{3X_i^2}{\tilde{\sigma}^7} - \frac{X_i^4}{\tilde{\sigma}^9}\right)\right]\right\} < +\infty$$

which is obtained by an easy majorization.

■

Proof *of Assumption 6:*

Almost surely we have :

$$\lim_{\delta\to 0} \sup_{\substack{|\sigma^2-1|<\delta \\ p\in[0,1]}} \left|Y_1^2(\sigma^2) - \left[\frac{\frac{1}{\sigma} e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}} - e^{-\frac{1}{2}X_1^2}}{\left(\sigma^2-1\right)\left\{(1-p)\, e^{-\frac{1}{2}X_1^2} + p\frac{1}{\sigma} e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}}\right\}}\right]^2\right| = 0.$$

In order to apply Lebesgue's Theorem let us write :

$$\begin{aligned} &\sup_{\substack{|\sigma^2-1|<\delta \\ p\in[0,1]}} \left|Y_1^2(\sigma^2) - \left[\frac{\frac{1}{\sigma} e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}} - e^{-\frac{1}{2}X_1^2}}{\left(\sigma^2-1\right)\left\{(1-p)\, e^{-\frac{1}{2}X_1^2} + p\frac{1}{\sigma} e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}}\right\}}\right]^2\right| \\ \leqslant\ & \sup_{\sigma^2\in[a,A]} \left|Y_1^2(\sigma^2)\right| + \sup_{\substack{|\sigma^2-1|<\delta \\ p\in[0,1]}} \left[\frac{\frac{1}{\sigma} e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}} - e^{-\frac{1}{2}X_1^2}}{\left(\sigma^2-1\right)\left\{(1-p)\, e^{-\frac{1}{2}X_1^2} + p\frac{1}{\sigma} e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}}\right\}}\right]^2. \end{aligned} \tag{I.6}$$

In order to get an independent of δ and H_0-integrable majorization of the last preceding term we consider two cases : $p \in \left[0; \frac{1}{2}\right]$ et $p \in \left]\frac{1}{2}; 1\right]$.

First case: $p \in [0; \frac{1}{2}]$

$$\sup_{\substack{|\sigma^2-1|<\delta \\ p\in\left[0,\frac{1}{2}\right]}} \left[\frac{\frac{1}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}} - e^{-\frac{1}{2}X_1^2}}{(\sigma^2-1)\left\{(1-p)\,e^{-\frac{1}{2}X_1^2} + p\frac{1}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}}\right\}}\right]^2 \leqslant \sup_{\substack{|\sigma^2-1|<\delta \\ p\in\left[0,\frac{1}{2}\right]}} \left[\frac{\frac{1}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}} - e^{-\frac{1}{2}X_1^2}}{(\sigma^2-1)\left\{(1-p)\,e^{-\frac{1}{2}X_1^2}\right\}}\right]^2$$

$$\leqslant \sup_{|\sigma^2-1|<\delta} \frac{4}{(\sigma^2-1)^2}\left[\frac{\frac{1}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}} - e^{-\frac{1}{2}X_1^2}}{e^{-\frac{1}{2}X_1^2}}\right]^2 \leqslant 4 \sup_{\sigma^2\in[a,A]} Y_1^2(\sigma^2) \tag{I.7}$$

which does not depend on δ and is integrable by Assumption 5.
Second case: $p \in \left]\frac{1}{2}, 1\right]$
We here consider two different cases again.
First we deal with $\sigma^2 \geqslant 1$ and write the following inequalities:

$$\sup_{\substack{|\sigma^2-1|<\delta \\ p\in\left]\frac{1}{2},1\right]}} \left[\frac{\frac{1}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}} - e^{-\frac{1}{2}X_1^2}}{(\sigma^2-1)\left\{(1-p)\,e^{-\frac{1}{2}X_1^2} + p\frac{1}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}}\right\}}\right]^2 \leqslant \sup_{\substack{|\sigma^2-1|<\delta \\ p\in\left]\frac{1}{2},1\right]}} \left[\frac{\frac{1}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}} - e^{-\frac{1}{2}X_1^2}}{(\sigma^2-1)\frac{p}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}}}\right]^2$$

$$\leqslant \sup_{|\sigma^2-1|<\delta} \frac{4}{(\sigma^2-1)^2}\left[\frac{\frac{1}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}} - e^{-\frac{1}{2}X_1^2}}{\frac{1}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}}}\right]^2 \leqslant 4 \sup_{|\sigma^2-1|<\delta} \left[Y_1^2(\sigma^2)\sigma^2 e^{-X_1^2(1-\frac{1}{\sigma^2})}\right].$$

As we assumed $\sigma^2 \geqslant 1$ we get:

$$-X_1^2(1-\frac{1}{\sigma^2}) \leqslant 0,$$

hence:

$$\sup_{\substack{|\sigma^2-1|<\delta \\ p\in\left]\frac{1}{2},1\right]}} \left[\frac{\frac{1}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}} - e^{-\frac{1}{2}X_1^2}}{(\sigma^2-1)\left\{(1-p)\,e^{-\frac{1}{2}X_1^2} + p\frac{1}{\sigma}e^{-\frac{1}{2}\frac{X_1^2}{\sigma^2}}\right\}}\right]^2 \leqslant 4A \sup_{\sigma^2\in[a,A]} \left[Y_1^2(\sigma^2)\right]$$

which, up to the constant A, gives the same majorization as in (I.7).
Let us now consider the case $\sigma^2 < 1$. By a Taylor expansion of the numerator of the ratio inside the squared brackets around $\sigma^2 = 1$, there exists $\sigma_1^2 \in]\sigma^2, 1[$ such that:

$$\left[\frac{\frac{1}{\sigma}e^{-\frac{1}{2}\frac{x_1^2}{\sigma^2}} - e^{-\frac{1}{2}x_1^2}}{(\sigma^2-1)\left\{(1-p)\,e^{-\frac{1}{2}x_1^2} + p\frac{1}{\sigma}e^{-\frac{1}{2}\frac{x_1^2}{\sigma^2}}\right\}}\right]^2 = \left[\frac{1}{2\sigma_1}\left(\frac{x_1^2}{\sigma_1^4} - \frac{1}{\sigma_1^2}\right)\frac{e^{-\frac{x_1^2}{2\sigma_1^2}}}{\left\{(1-p)\,e^{-\frac{1}{2}x_1^2} + p\frac{1}{\sigma}e^{-\frac{1}{2}\frac{x_1^2}{\sigma^2}}\right\}}\right]^2.$$

As $p \in \left]\frac{1}{2}, 1\right]$ and $\sigma^2, \sigma_1^2 \in [a, A]$, we have:

$$\left[\frac{\frac{1}{\sigma}e^{-\frac{1}{2}\frac{x_1^2}{\sigma^2}} - e^{-\frac{1}{2}x_1^2}}{(\sigma^2-1)\left\{(1-p)\,e^{-\frac{1}{2}x_1^2} + p\frac{1}{\sigma}e^{-\frac{1}{2}\frac{x_1^2}{\sigma^2}}\right\}}\right]^2 \leqslant \frac{A}{a}\left(\frac{x_1^2}{a^4} + \frac{1}{a^2}\right)^2 e^{x_1^2\frac{\sigma_1^2-\sigma^2}{\sigma_1^2\sigma^2}}.$$

We can choose δ strictly less than $\frac{1}{4}$, and we have:

$$\frac{3}{4} < \sigma^2 < \sigma_1^2 < 1 \quad \textit{which implies that} \quad \sigma_1^2 - \sigma^2 < \frac{1}{4}.$$

Then, for all σ^2 satisfying the preceding inequality, we have

$$0 < \frac{\sigma_1^2 - \sigma^2}{\sigma_1^2 \sigma^2} < \frac{4}{9}.$$

Then we see that

$$\mathbb{E}\left[\frac{A}{a}\left(\frac{X_1^2}{a^4} + \frac{1}{a^2}\right)^2 \exp\left(\frac{4X_1^2}{9}\right)\right] < +\infty. \tag{I.8}$$

Gathering (I.6), (I.7) and (I.8) we apply the Lebesgue dominated convergence Theorem, and Assumption 6 is satisfied.

■

2.4 Percentiles points and application to a real data set

In order to implement the test, we need to compute some percentiles points. We have to deal with the supremum of a non-stationary Gaussian process, which is difficult. Nowadays, we are able to compute very precise values using upper bounds such as the Davies bound or recent works using Rice formulae (Delmas 2001). In the following subsection we will compute the percentiles points using Davies bound for some intervals and some different levels. The second subsection is devoted to the same computations using Rice's method instead.

2.4.1 Percentiles points using Davies bound

We give here an overview of Davies' work (Davies, 1977) in order to generalize its result to a process $Z \in C^1([-T_1, 0[\cup]0, T_2])$ with $Z(0^+) = -Z(0^-)$. First, let $Z \in C^1([L, U])$ be a centered Gaussian process with unit variance and autocovariance function $cov(Z(\theta_1), Z(\theta_2)) = \gamma(\theta_1, \theta_2)$. We want to compute $\mathbb{P}_0\left(\sup_{L \leqslant \theta \leqslant U} Z(\theta) > c\right)$ for any constant $c > 0$, where $\mathbb{P}_0$ denote the probability associated to the null Hypothesis $\{\theta = 0\}$. Davies showed that

$$\mathbb{P}_0\left(\sup_{L \leqslant \theta \leqslant U} Z(\theta) > c\right) \leqslant \Phi(-c) + \frac{1}{2\pi} e^{-\frac{c^2}{2}} \int_L^U \left(\gamma_{11}(\theta)\right)^{1/2} d\theta,$$

where Φ is the cumulative distribution function of the standard normal distribution, and

$$\gamma_{11}(\theta) = \left.\frac{\partial^2 \gamma(\alpha, \beta)}{\partial \alpha \partial \beta}\right|_{(\theta,\theta)}.$$

Now we assume $Z \in C^1([-T_1, 0[\cup]0, T_2])$ with $Z(0^+) = -Z(0^-)$ almost surely. To compute $\mathbb{P}_0\left(\sup_{-T_1 \leqslant \theta \leqslant T_2} Z(\theta) > c\right)$ we need to use $U_c^Z[0, T_2]$ the number of upcrossings of the level c of the process Z on $[0, T_2]$, and $D_c^Z[-T_1, 0]$ the number of downcrossings of the level c of the process Z on $[-T_1, 0]$. We have

$$\begin{aligned} U_c^Z[0, T_2] &= \#\{t \in [0, T_2] : Z(t) = c, Z'(t) > 0\}, \\ D_c^Z[-T_1, 0] &= \#\{t \in [-T_1, 0] : Z(t) = c, Z'(t) < 0\}. \end{aligned}$$

The event $\left\{\sup_{L \leqslant \theta \leqslant U} Z(\theta) > c\right\}$ can be written as a union of three different sets:

$$\begin{aligned} &\left\{\sup_{L \leqslant \theta \leqslant U} Z(\theta) > c\right\} = \\ &\left\{Z\left(0^+\right) > c\right\} \cup \left\{Z\left(0^-\right) > c\right\} \cup \left\{\left(U_c^Z[0, T_2] + D_c^Z[-T_1, 0]\right) \mathbb{1}_{[|Z(0^+)| \leqslant c]} \geqslant 1\right\}. \end{aligned}$$

Using $\mathbb{P}\{\xi \geqslant 1\} \leqslant \mathbb{E}(\xi)$ for any positive random variable, we obtain:

$$\begin{aligned} \mathbb{P}\left\{\sup_{-T_1 \leqslant t \leqslant T_2} Z(t) > c\right\} &= \mathbb{P}\left[Z\left(0^+\right) > c\right] + \mathbb{P}\left[Z\left(0^-\right) > c\right] \\ &\quad + \mathbb{P}\left[\left(U_c^Z[0, T_2] + D_c^Z[-T_1, 0]\right) \mathbb{1}_{[|Z(0^+)| \leqslant c]} \geqslant 1\right]. \\ &\leqslant 2\Phi(-c) + \mathbb{E}\left(U_c^Z[0, T_2]\right) + \mathbb{E}\left(D_c^Z[-T_1, 0]\right). \end{aligned}$$

Assuming $T_1 = T_2$ and Z a symmetric process, we obtain that:

$$\mathbb{E}\left(U_c^Z\left[0,T\right]\right) = \mathbb{E}\left(D_c^Z\left[-T,0\right]\right).$$

Finally:

$$\mathbb{P}\left\{\sup_{-T\leqslant t\leqslant T} Z\left(t\right) > c\right\} \leqslant M\left(c\right),$$

where

$$M\left(c\right) = 2\Phi\left(-c\right) + \frac{1}{\pi}e^{-\frac{c^2}{2}}\int_0^T \left(\gamma_{11}(\theta)\right)^{1/2} d\theta.$$

Applying this result to the process T with γ given by (I.5), we obtain:

$$\gamma_{11}(\sigma^2) = -\frac{\frac{3}{4}(1-\sigma^2)^2 - \frac{1}{2}(3-4\sigma^2+2\sigma^4)}{\left(\frac{1}{\sqrt{2\sigma^2-\sigma^4}}-1\right)(2\sigma^2-\sigma^4)^{\frac{5}{2}}} - \frac{1}{4\left(\frac{1}{\sqrt{2\sigma^2-\sigma^4}}-1\right)^2}\frac{(\sigma^2-1)^2}{(2\sigma^2-\sigma^4)^3}$$

and around $h = 0$:

$$\left[\gamma_{11}(1+h)\right]^{\frac{1}{2}} = 3^{\frac{1}{2}}\left[\frac{1}{2} + \frac{11}{24}h^2 + \frac{257}{576}h^4 + O(h^6)\right].$$

Therefore the approximate percentile point for the test with $\sigma^2 \in [a, A]$, and α level, is solution of:

$$M(c) = \alpha.$$

We use the secant method to solve this equation, and we compute the quantiles for different levels α, and different sets $[a, A]$. The results are shown in Table I.1. Remember that they need to be applied to centered and reduced data only.

[a,A] \ **1-α**	**0.90**	**0.95**	**0.99**
[0.1,1.9]	3.9658	5.2550	8.3188
[0.2,1.9]	3.8432	5.1232	8.1722
[0.3,1.9]	3.7610	5.0350	8.0733
[0.4,1.9]	3.6971	4.9650	7.9942
[0.5,1.9]	3.6418	4.9046	7.9256
[0.6,1.9]	3.5918	4.8500	7.8632
[0.7,1.9]	3.5451	4.7988	7.8045
[0.8,1.9]	3.5004	4.7496	7.7477
[0.9,1.9]	3.4566	4.7012	7.6917

Table I.1: Percentiles points for the LRTS for different sets, obtained using Davies bound for different levels.

2.4.2 Percentiles points using Rice method

We consider now a symmetric centered Gaussian process Z with unit variance such as in the previous section. Following Delmas (2001) we have:

$$\begin{aligned}\mathbb{P}\left\{\sup_{-T\leqslant t\leqslant T} Z\left(t\right) > c\right\} &= \mathbb{P}\left[Z\left(0^+\right) > c\right] + \mathbb{P}\left[Z\left(0^-\right) > c\right] \\ &\quad + \mathbb{P}\left[\left(U_c^Z\left[0,T\right] + D_c^Z\left[-T,0\right]\right)\mathbb{1}_{[|Z(0^+)|\leqslant c]} \geqslant 1\right] \\ &\leqslant 2\Phi\left(-c\right) + 2\mathbb{E}\left(U_c^Z\left[0,T\right]\mathbb{1}_{[|Z(0^+)|\leqslant c]}\right),\end{aligned}$$

where

$$\mathbb{E}\left(U_c^Z\left[0,T\right]\mathbb{1}_{[|Z(0^+)|\leqslant c]}\right) = \frac{1}{2\pi}e^{-\frac{c^2}{2}}\int_0^T \left(\gamma_{11}(\theta)\right)^{1/2} d\theta - F\left(\gamma\left(0,T\right), \gamma_{1,0}\left(T,0\right)\right),$$

with

$$F\left(\gamma\left(0,T\right),\gamma_{1,0}\left(T,0\right)\right) = \phi\left(c\right)\int_0^T \frac{\left(\gamma_{11}(\theta)\right)^{1/2}}{\sqrt{2\pi}}\bar{\Phi}\left[\frac{c}{v}\left(\gamma_{11}(\theta)\right)^{1/2}\sqrt{\frac{1-\gamma\left(0,\theta\right)}{1+\gamma\left(0,\theta\right)}}\right] + \frac{\gamma_{1,0}\left(\theta,0\right)}{\sqrt{1-\gamma\left(0,\theta\right)}}\phi\left(c\sqrt{\frac{1-\gamma\left(0,\theta\right)}{1+\gamma\left(0,\theta\right)}}\right)\bar{\Phi}\left(-\frac{c\gamma_{1,0}\left(\theta,0\right)}{v\left(1+\gamma\left(0,\theta\right)\right)}\right)d\theta,$$

$$\gamma_{1,0}\left(\theta,0\right) = \lim_{\substack{\theta'\to 0\\ \theta'>0}} \frac{\partial\gamma(\alpha,\beta)}{\partial\alpha}\Big|_{(\theta,\theta')} \qquad \bar{\Phi} = 1-\Phi,$$

and

$$v^2 = \gamma_{11}(\theta) - \frac{\gamma_{1,0}^2\left(\theta,0\right)}{1-\gamma^2\left(0,\theta\right)},$$

$$\gamma\left(0,\theta\right) = \lim_{\substack{\theta'\to 0\\ \theta'>0}} \gamma\left(\theta',\theta\right).$$

As before, we compute the quantiles for different levels α and different sets $[a,A]$. We need to be very careful calculating all the previous terms because of numerical instabilities associated to $F\left(\gamma\left(0,T\right),\gamma_{1,0}\left(T,0\right)\right)$. The use of Taylor expansion is needed around the singularity. We used Maple software for obtaining theoretical expressions and Fortran 77 for the quantile's computation.

[a,A] \ **1-α**	**0.90**	**0.95**	**0.99**
[0.1,1.9]	3.9511	5.2441	8.3133
[0.2,1.9]	3.8310	5.1140	8.1674
[0.3,1.9]	3.7507	5.0268	8.0690
[0.4,1.9]	3.6868	4.9572	7.9901
[0.5,1.9]	3.6317	4.8971	7.9217
[0.6,1.9]	3.5818	4.8424	7.8593
[0.7,1.9]	3.5351	4.7912	7.8005
[0.8,1.9]	3.4902	4.7418	7.7436
[0.9,1.9]	3.4461	4.6932	7.6874

Table I.2: Percentiles points for the LRTS for different sets, obtained using Rice bound for different levels.

The impact of the penalizing term $F\left(\gamma\left(0,T\right),\gamma_{1,0}\left(T,0\right)\right)$ when approximating $\mathbb{P}\left\{\sup_{-T\leqslant t\leqslant T} Z\left(t\right) > c\right\}$ is tiny. Davies bound seems to give a very precise approximation but Rice method gives more accurate P values.

In order to get an idea of the true level of the test, we performed simulations for a nominal level 5%. We replicate 5000 times a random sample of $n = 250$ $\mathcal{N}\left(0,1\right)$ values. We calculate the percentage of rejections of H_0 both for Davies and Rice methods. See Table I.3.

2.4.3 Application to a real data set

We study a real data set of 542 observations representing the blood chloride values obtained during routine analysis under the assumption of a proportion (p) of "healthy" values, together with a proportion $(1-p)$ of "unhealthy" values. The data, represented in the following table come from Naylor and Smith (1983).

Even under a mixture assumption, we note that the mean is constant and approximately equal to 100. Naylor and Smith have identified a majority component corresponding to the "healthy" population. Unfortunately, its variance is smaller than a half the total variance. In order to apply our results, we need to assume "unhealthy" values, with biggest variance, to be the known component under a contamination model. The "healthy" values are considered following

[a,A]	**Davies**	**Rice**
[0.1,1.9]	5.07	5.08
[0.2,1.9]	4.88	4.92
[0.3,1.9]	4.79	4.79
[0.4,1.9]	4.72	4.73
[0.5,1.9]	4.75	4.75
[0.6,1.9]	4.77	4.77
[0.7,1.9]	4.58	4.58
[0.8,1.9]	4.54	4.55
[0.9,1.9]	4.63	4.67

Table I.3: Simulated level of rejection of H_0 when H_0 is true with 10 000 replications of n=250 random samples and a 5% nominal level.

Level	88	89	90	91	92	93	94	95	96	97	98	99	100
Frequency	2	3	4	5	7	5	13	13	27	36	40	72	68
Level	101	102	103	104	105	106	107	108	109	111	113	115	
Frequency	80	47	43	33	19	6	6	4	5	2	1	1	

Table I.4: Blood chloride level (mmol/liter) for 542 individuals (NRI Data)

the contamination density. From Naylor and Smith's work, we found the standard deviation of the "unhealthy" component to be approximately 5. We can write the testing problem as

$$\begin{aligned} H_0 &: \mathcal{N}(100, 25) \\ H_1 &: (1-p)\mathcal{N}(100, 25) + p\mathcal{N}\left(100, \sigma^2\right). \end{aligned}$$

Centering the data with respect to 100 and dividing by 5, we find the testing problem I.4 with p and σ^2 unknown. We computed the LRTS:

$$\sup_{\sigma^2 \in [0.1,1[\cup]1,1.9]} T_n^2\left(\sigma^2\right) \mathbb{1}_{[T_n(\sigma^2)>0]} \simeq 112.$$

The homogeneity Hypothesis is rejected at a 5% error rate. We notice that H_0 is rejected even at a 1% error rate.

2.5 Proof of the main result

In the proof of Theorem 1, we are going to use the following lemma.

Lemma 1 : We have

$$\lim_{\delta \to 0} \mathbb{E}\left[\sup_{\substack{|p|<\delta \\ \theta \in \bar{\Theta}}} \left|Y_1^2(\theta) - g_1(X_1, p, \theta)\right|\right] = 0 \tag{I.9}$$

and

$$\lim_{\delta \to 0} \mathbb{E}\left[\sup_{\substack{|p\theta|<\delta \\ (p,\theta) \in [0,1]\times\bar{\Theta}}} \left|Y_1^2(\theta) - g_1(X_1, p, \theta)\right|\right] = 0 \tag{I.10}$$

where

$$g_1(x_i, p, \theta) = \left[\frac{1}{\theta} \frac{g(x_i, \theta) - g(x_i)}{(1-p)\, g(x_i) + p g(x_i, \theta)}\right]^2.$$

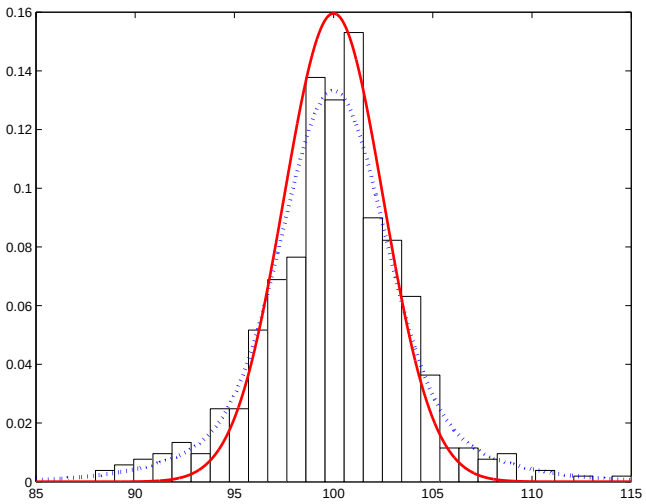


Figure I.2: Data histogram, predicted density of the "healthy" population (plain line) and predicted density of the total population (dotted line). See Naylor and Smith (1983).

Proof of the lemma*1.*

Proof of (I.9) :
The variable θ belongs to a compact set $[-a,b]$. For all $x \in \mathbb{R}$ we have

$$\lim_{\delta \to 0} \sup_{\substack{|p|<\delta \\ \theta \in \Theta}} \left| Y_1^2(\theta) - g_1(X_1, p, \theta) \right| = 0.$$

We have also

$$\sup_{\substack{|p|<\delta \\ \theta \in \Theta}} \left| Y_1^2(\theta) - g_1(X_1, p, \theta) \right| \leqslant A_1 + A_2$$

with

$$\begin{aligned} A_1 &= \sup_{\theta \in \Theta} \left| Y_1^2(\theta) \right| \\ A_2 &= \sup_{\substack{|p|<\delta \\ \theta \in \Theta}} \left[\frac{1}{\theta} \frac{g(x,\theta) - g(x)}{(1-p)\, g(x) + p g(x,\theta)} \right]^2. \end{aligned}$$

We can assume that $\delta < \frac{1}{2}$ (and so $p < \frac{1}{2}$). So we have $(1-p)\, g(x) + p g(x,\theta) \geqslant \frac{g(x)}{2}$ and this implies $A_2 \leqslant 4A_1$. Then by Assumption 5, we obtain $\mathbb{E}(A_1 + A_2) < +\infty$. The result follows using Lebesgue's dominated convergence Theorem.

Proof of (I.10) :
Let us define the following sets:

$$\begin{aligned} A_1(\delta) &= \{(p,\theta) \in [0,1] \times [-a,b] \mid |p\theta| < \delta\}, \\ A_2\left(\sqrt{\delta}\right) &= \left\{(p,\theta) \in [0,1] \times [-a,b] \mid |\theta| < \sqrt{\delta}\right\}, \\ A_3\left(\sqrt{\delta}\right) &= \left\{(p,\theta) \in [0,1] \times [-a,b] \mid |p| < \sqrt{\delta}\right\}. \end{aligned}$$

Using : $p \geqslant \sqrt{\delta}$ and $|\theta| \geqslant \sqrt{\delta}$ implies $p|\theta| \geqslant \delta$, we see that $A_1(\delta) \subset A_2\left(\sqrt{\delta}\right) \cup A_3\left(\sqrt{\delta}\right)$. Then Equality (I.10) is deduced from equality (

refensemble2) and Assumption 6.

■

Proof of the theorem *1:*
Let us derive the partial derivatives of the log-density for the model:

$$\begin{aligned}
\frac{\partial \log f}{\partial p}(x_i,p,\theta) &= \left[\frac{g(x_i,\theta)-g(x_i)}{(1-p)g(x_i)+pg(x_i,\theta)}\right],\\
\frac{\partial \log f}{\partial p}(x_i,0,\theta) &= \frac{g(x_i,\theta)}{g(x_i)}-1,\\
\frac{\partial^2 \log f}{\partial p^2}(x_i,p,\theta) &= -\left[\frac{\partial \log f}{\partial p}(x_i,p,\theta)\right]^2,\\
\frac{\partial^2 \log f}{\partial p^2}(x_i,0,\theta) &= -\left[\frac{g(x_i,\theta)}{g(x_i)}-1\right]^2.
\end{aligned}$$

We can write a Taylor expansion of the log likelihood function with respect to $p\in]0,1[$ for all $\theta\neq 0$.

$$\begin{aligned}
L_n(x_1,...,x_n;p,\theta) &= L_n(0)+p\theta\sum_{i=1}^n\frac{1}{\theta}\left[\frac{g(x_i,\theta)}{g(x_i)}-1\right]\\
&\quad -\frac{1}{2}p^2\theta^2\sum_{i=1}^n\left[\frac{1}{\theta}\frac{\partial \log f}{\partial p}(x_i,\alpha p,\theta)\right]^2\\
&= L_n(0)+p\theta\sum_{i=1}^n\frac{1}{\theta}\left[\frac{g(x_i,\theta)}{g(x_i)}-1\right]\\
&\quad -\frac{1}{2}p^2\theta^2\sum_{i=1}^n g_1(x_i,\alpha p,\theta),
\end{aligned}$$

using the same notation than above and page 3 where $\alpha\in]0,1[$ depending on p and $(x_1,...,x_n)$.

We define a modified log-likelihood:

$$\tilde{L}_n(x_1,...,x_n;p,\theta)=L_n(0)+\sqrt{n}p\theta\sum_{i=1}^n\frac{y_i(\theta)}{\sqrt{n}}-\frac{n}{2}p^2\theta^2\sum_{i=1}^n\frac{y_i^2(\theta)}{n},$$

where $y_i(\theta)$ is a realization of $Y_i(\theta)$ (see (I.2)). Assumption 3 implies that

$$\sup_{\theta\in\bar{\Theta}}\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i(\theta)=O_p(1).$$

Assumption 4 implies that $\inf_{\theta\in\bar{\Theta}}\mathbb{E}\left(Y_1^2(\theta)\right)\geqslant C>0$ for all $\theta\in\bar{\Theta}$. Hence, using the consequence of Assumption 5, for given $\varepsilon>0$, we can find $K>0$ and n_0 such that for all $n>n_0$ we have:

$$\mathbb{P}\left[\begin{array}{c}\sup_{\theta\in\bar{\Theta}}\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i(\theta)<K\\ and\\ inf_{\theta\in\bar{\Theta}}\frac{1}{n}\sum_{i=1}^n Y_i^2(\theta)\geqslant\frac{C}{2}\end{array}\right]>1-\varepsilon.$$

Therefore, we can find $M>0$ such that, with a probability greater than $1-\varepsilon$, for $n>n_0$ and $|p\theta\sqrt{n}|\geqslant M$, we have $\tilde{L}_n(x_1,...,x_n;p,\theta)\leqslant L_n(0)$. Then the supremum of $2\left(\tilde{L}_n(x_1,...,x_n;p,\theta)-L_n(0)\right)$ is attained for $|p\theta\sqrt{n}|<M$. This supremum is given by:

$$\sup_{\theta\in\Theta}\frac{\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n y_i(\theta)\right]^2}{\frac{1}{n}\sum_{i=1}^n y_i^2(\theta)}\mathbb{1}_{\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\partial \log f}{\partial p}(x_i,0,\theta)\geqslant 0\right]}. \tag{I.11}$$

Using Assumption 5 and for the corresponding random variables we get

$$(I.11) = \sup_{\theta\in\bar{\Theta}} \frac{\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i(\theta)\right]^2}{\mathbb{E}\left(Y_1^2(\theta)\right)} \mathbb{1}_{\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f}{\partial p}(X_i,0,\theta)\geqslant 0\right]} + o_p(1),$$

which can be written as

$$\begin{aligned} &\sup_{\theta\in\bar{\Theta}\setminus\{0\}} \frac{\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{g(X_i,\theta)}{g(X_i)}-1\right]\right]^2}{\mathbb{E}\left[\frac{g(X_1,\theta)}{g(X_1)}-1\right]^2} \mathbb{1}_{\left[\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{g(X_i,\theta)}{g(X_i)}-1\right]}{\left\{\mathbb{E}\left[\frac{g(X_1,\theta)}{g(X_1)}-1\right]^2\right\}^{\frac{1}{2}}}\geqslant 0\right]} + o_p(1) \\ = &\sup_{\theta\in\bar{\Theta}\setminus\{0\}} T_n^2(\theta)\,\mathbb{1}_{[T_n(\theta)\geqslant 0]} + o_p(1), \end{aligned}$$

where $o_p(1)$ tends to 0 in probability as n tends to infinity, uniformly with respect to θ. Let us come back to the original likelihood function that can be rewritten:

$$\begin{aligned} L_n(x_1,...,x_n;p,\theta) &= L_n(0) + \sqrt{n}p\theta\sum_{i=1}^{n}\frac{y_i(\theta)}{\sqrt{n}} \\ &\quad -\frac{n}{2}p^2\theta^2\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2(\theta) + \frac{1}{n}\sum_{i=1}^{n}\left(g_1(x_i,\alpha p,\theta) - y_i^2(\theta)\right)\right]. \end{aligned}$$

Let $\mathcal{D} = \{(p,\theta) \mid f(x_i,p,\theta) = f_0\}$ with f_0 the true density under H_0, and $W(\mathcal{D})$ a fixed neighborhood of $\mathcal{D}$. Using Assumption 2 and Redner's result (1981), we have

$$\sup_{(p,\theta)\in[0,1]\times\bar{\Theta}} L_n(X_1,...,X_n;p,\theta) - \sup_{(p,\theta)\in W(\mathcal{D})} L_n(X_1,...,X_n;p,\theta) = o_p(1). \tag{I.12}$$

An arbitrary neighborhood of $\mathcal{D}$ can be included in a set $A_1(\delta)$, for a given δ. Using Lemma 1, uniformly with respect to $(p,\theta)\in A_1(\delta)$, the sum $\frac{1}{n}\sum_{i=1}^{n}\left(g_1(X_i,\alpha p,\theta) - Y_i^2(\theta)\right)$ tends to zero in probability as n tends to infinity and δ tends to zero. For δ small enough and for all $\varepsilon > 0$, there exist n_0 and K such that for all $n > n_0$:

$$\mathbb{P}\left[\begin{array}{c} \sup_{\theta\in\bar{\Theta}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i(\theta) < K \\ and \\ inf_{\theta\in\bar{\Theta}}\frac{1}{n}\sum_{i=1}^{n} g_1(X_i,\alpha p,\theta) \geqslant \frac{C}{2} \end{array}\right] > 1-\varepsilon$$

uniformly with respect to (p,θ) on $A_1(\delta)$. As above, we can find $M > 0$ such that $\sup_{(p,\theta)\in A_1(\delta)} L_n(x_1,...,x_n;p,\theta)$ is reached on $|p\theta\sqrt{n}| < M$. On this set we have:

$$\begin{aligned} \left|L_n - \tilde{L}_n\right| &= \frac{1}{2}p^2\theta^2 n\left|\frac{1}{n}\sum_{i=1}^{n} Y_i^2(\theta) - \frac{1}{n}\sum_{i=1}^{n} g_1(X_i,\alpha p,\theta)\right| \\ &\leqslant \frac{1}{2}M^2\left(\sup_{\substack{(p,\theta)\in[0,1]\times\bar{\Theta} \\ |p\theta|<\frac{M}{\sqrt{n}}}} \frac{1}{n}\sum_{i=1}^{n}\left|Y_i^2(\theta) - g_1(X_i,\alpha p,\theta)\right|\right) \\ &= o_p(1). \end{aligned} \tag{I.13}$$

Using (I.12) and (I.13) we obtain:

$$\sup_{\substack{(p,\theta)\in[0,1]\times\bar{\Theta} \\ |p\theta|<\frac{M}{\sqrt{n}}}} \left|L_n - \tilde{L}_n\right| = o_p(1).$$

We conclude that:

$$\left| \sup_{(p,\theta)\in[0,1]\times\bar{\Theta}} L_n\left(X_1,...,X_n;p,\theta\right) - \sup_{(p,\theta)\in[0,1]\times\bar{\Theta}} \tilde{L}_n\left(X_1,...,X_n;p,\theta\right) \right| = o_p\left(1\right).$$

Hence, the likelihood ratio test statistic is

$$\sup_{\theta\in\bar{\Theta}\setminus\{0\}} T_n^2\left(\theta\right) 1\!\!1_{[T_n(\theta)\geqslant 0]} + o_p\left(1\right).$$

We are going to deal now with the limit of $T_n^2(\theta)$ when θ tends to zero. Using the Taylor expansion (I.3) page 16 with $\eta\in\left]0,1\right[$ and $\theta\in\Theta$ we have:

$$\frac{g\left(x,\theta\right)}{g\left(x\right)} - 1 = \theta\frac{g'_\theta\left(x\right)}{g\left(x\right)} + \theta\left[\frac{g'_\theta\left(x,\eta\theta\right) - g'_\theta\left(x\right)}{g\left(x\right)}\right].$$

We rewrite

$$T_n\left(\theta\right) = \frac{sign(\theta)}{\sqrt{n}} \frac{\sum_{i=1}^n \left[\frac{g'_\theta(X_i,\eta\theta)-g'_\theta(X_i)}{g(X_i)} + \frac{g'_\theta(X_i)}{g(X_i)}\right]}{D\left(X_1,\theta,\eta\right)},$$

where

$$\begin{aligned} D\left(X_1,\theta,\eta\right) = \; & \left\{\mathbb{E}\left(\frac{g'_\theta(X_1)}{g(X_1)}\right)^2 + \mathbb{E}\left(\frac{g'_\theta(X_1,\eta\theta)-g'_\theta(X_1)}{g(X_1)}\right)^2\right. \\ & \left. +2\mathbb{E}\left[\frac{g'_\theta(X_1)}{g(X_1)}\left(\frac{g'_\theta(X_1,\eta\theta)-g'_\theta(X_1)}{g(X_1)}\right)\right]\right\}^{\frac{1}{2}}. \end{aligned}$$

By Assumption 3, we have $\mathbb{E}\left(\frac{g'_\theta(X_1,\eta\theta)-g'_\theta(X_1)}{g(X_1)}\right)^2 = \mathbb{E}\left(Y_1\left(\theta\right) - Y_1\left(0\right)\right)^2 \leqslant K\left|\theta\right|^2 < +\infty$. We deduce that $\mathbb{E}\left(\frac{g'_\theta(X_1,\eta\theta)-g'_\theta(X_1)}{g(X_1)}\right)^2$ tends to zero when θ tends to zero. By the Cauchy-Schwarz inequality we have

$$\mathbb{E}\left|\frac{g'_\theta\left(X_1\right)}{g\left(X_1\right)}\left(\frac{g'_\theta\left(X_1,\eta\theta\right) - g'_\theta\left(X_1\right)}{g\left(X_1\right)}\right)\right| \leqslant \left[\mathbb{E}\left(\frac{g'_\theta\left(X_1\right)}{g\left(X_1\right)}\right)^2\right]^{\frac{1}{2}} \left[\mathbb{E}\left(\frac{g'_\theta\left(X_1,\eta\theta\right) - g'_\theta\left(X_1\right)}{g\left(X_1\right)}\right)^2\right]^{\frac{1}{2}},$$

and

$$\lim_{\theta\longrightarrow 0} D\left(X_1,\theta,\eta\right) = \left\{\mathbb{E}\left(\frac{g'_\theta\left(X_1\right)}{g\left(X_1\right)}\right)^2\right\}^{\frac{1}{2}}.$$

Hence we have:

$$\lim_{\theta\longrightarrow 0} T_n^2\left(\theta\right) = \frac{1}{n}\frac{\left[\sum_{i=1}^n\left(\frac{g'_\theta(X_i)}{g(X_i)}\right)\right]^2}{\mathbb{E}\left(\frac{g'_\theta(X_1)}{g(X_1)}\right)^2}.$$

■

Proof of the theorem *2:*
We give the proof for $\mathcal{C}\left[0,a\right]$. Using (I.11) we have

$$\tilde{T}_n\left(\theta\right) = n^{-\frac{1}{2}}\frac{\sum_{i=1}^n Y_i\left(\theta\right)}{\left(\mathbb{E}\left(Y_1\left(\theta\right)\right)^2\right)^{\frac{1}{2}}} \quad \text{and} \quad \tilde{T}_n\left(0\right) = \lim_{\theta\longrightarrow 0^+} T_n\left(\theta\right) = n^{-\frac{1}{2}}\frac{\sum_{i=1}^n\left(\frac{g'_\theta(X_i)}{g(X_i)}\right)}{\left\{\mathbb{E}\left(\frac{g'_\theta(X_1)}{g(X_1)}\right)^2\right\}^{\frac{1}{2}}}.$$

We already stated that $\sum_{i=1}^{n} Y_i(\theta)$ *is tight. Moreover as* $\inf_\theta \mathbb{E}\left(Y_1(\theta)\right)^2 \geqslant C > 0$, *we obtain the tightness of* $\tilde{T}_n(\theta)$. *We now establish the convergence of the finite-dimensional distributions, i.e. those of* $\left(\tilde{T}_n(\theta_1), ..., \tilde{T}_n(\theta_m)\right)$ *with* $0 \leqslant \theta_1 \leqslant ... \leqslant \theta_m \leqslant a$. *For any* $a_j \in \mathbb{R}$ $(j = 1, ..., k)$, *we have*

$$\sum_{j=1}^{k} a_j \left(\frac{\sum_{i=1}^{n} Y_i(\theta_j)}{\left(\mathbb{E}\left(Y_1(\theta_j)\right)^2\right)^{\frac{1}{2}}} \right) = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{a_j Y_i(\theta_j)}{\left(\mathbb{E}\left(Y_1(\theta_j)\right)^2\right)^{\frac{1}{2}}} = \sum_{i=1}^{n} Z_i,$$

where for $\theta = 0$, $\frac{Y_i(\theta)}{\left(\mathbb{E}(Y_1(\theta))^2\right)^{\frac{1}{2}}}$ *is replaced by* $\frac{\frac{g'_\theta(X_i)}{g(X_i)}}{\left\{\mathbb{E}\left(\frac{g'_\theta(X_1)}{g(X_1)}\right)^2\right\}^{\frac{1}{2}}}$. *The variables* Z_i *are independent and so, using the Cramer-Wold device and the classical central limit theorem, we obtain the asymptotic normality of these distributions.*

■

References

Bickel P. and Chernoff H. (1995). *Asymptotic distributions of the likelihood ratio statistic in a prototypical non regular problem.* In *Statistics and Probability: a Raghu Raj Bahadur Festchrift (eds J.K. Ghosh, S.K. Mitra, K.R. Parthasaraty and B.L.S. Prakasa Rao)*, New York. Wiley. pp. 83–96.

Billingsley P. (1968). *Convergence of probability measures.* John Wiley & Sons Inc.

Chen H. and Chen J. (2001). *Large sample distribution of the likelihood ratio test for normal mixtures.* Statist. Probab. Lett., vol. 52, n°2. pp. 125–133.

Cheng R. C. H. and Traylor L. (1995). *Non-regular maximum likelihood problems.* J. Roy. Statist. Soc. Ser. B, vol. 57, n°1. pp. 3–44.

Chernoff H. (1954). *On the distribution of the likelihood ratio.* Ann. Math. Statistics, vol. 25. pp. 573–578.

Davies R. B. (1977). *Hypothesis testing when a nuisance parameter is present only under the alternative.* Biometrika, vol. 64, n°2. pp. 247–254.

Delmas C. (2001). *Distribution of the maximum of a random field and statistical applications (in french)*, Université Paul Sabatier, Toulouse III, France.

Garel B. (2001). *Likelihood ratio test for univariate Gaussian mixture.* J. Statist. Plann. Inference, vol. 96, n°2. pp. 325–350.

Garel B. and Goussanou F. (2002). *Removing separation conditions in a 1 against 3-components gaussian mixture problem.* In *Classification, Clustering and data Analysis, Sokolowski A. and Boch H.H.(Eds)*, Berlin. Springer. pp. 61–73.

Ghosh J. and Sen P. (1985). *On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results.* In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II, Le Cam, L.M., Olshen, R.A. (Eds)*, Wadsworth Statist./Probab. Ser., Monterey. Wadsworth. pp. 789–806.

Hartigan J. A. (1985). *A failure of likelihood asymptotics for normal mixtures.* In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II, Le Cam, L.M., Olshen, R.A. (Eds)*, Wadsworth Statist./Probab. Ser., Monterey. Wadsworth. pp. 807–810.

Ledoux M. and Talagrand M. (1991). *Probability in Banach spaces.* Springer-Verlag.

Lemdani M. and Pons O. (1999). *Likelihood ratio tests in contamination models.* Bernoulli, vol. 5, n°4. pp. 705–719.

Li L. A. and Sedransk N. (1988). *Mixtures of distributions: a topological approach.* Ann. Statist., vol. 16, n°4. pp. 1623–1634.

Lindsay B. G. (1995). *Mixture models: Theory, geometry and applications.* IMS.

McLachlan G. and Peel D. (2000). *Finite mixture models.* Wiley-Interscience, New York.

Naylor J. and Smith A. (1983). *A contamination model in clinical chemistry: an illustration of a method for the efficient computation of posterior distributions.* The Statistician, vol. 32. pp. 82–87.

Redner R. (1981). *Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions.* Ann. Statist., vol. 9, n°1. pp. 225–228.

Teicher H. (1963). *Identifiability of finite mixtures.* Ann. Math. Statist., vol. 34. pp. 1265–1269.

Titterington D. M., Smith A. F. M. and Makov U. E. (1985). *Statistical analysis of finite mixture distributions.* John Wiley & Sons Ltd, Chichester.

Wolfe J. (1971). *A Monte Carlo study of the sampling distribution ratio for mixtures of multinormal distributions.* Technical Bulletin STB, U.S. NAV. Pers. and Train. Res. San Diego, vol. 72, n°2.

3 Compléments

Nous avons présenté dans cet article, deux méthodes différentes permettant d'approximer

$$\mathbb{P}\left[\sup_{t\in[0,T]}|Z(t)|>c\right] \tag{I.14}$$

où Z est un processus gaussien, symétrique et de variance un. Nous avons comparé les différentes majorations de (I.14) en constatant une meilleure précision de la borne de Rice. Le gain qui en résulte est toutefois très relatif comme nous avons pu nous en apercevoir lors des simulations présentées (tableau I.3). Malgré tout, nous devons signaler un avantage indéniable de la méthode de Rice présentée et développée par Delmas [2001]. En effet, cette méthode nous permet d'obtenir une borne inférieure pour (I.14). Etudions brièvement cet aspect de la méthode de Rice. On a :

$$\mathbb{P}\left[\sup_{t\in[0,T]}|Z(t)|>c\right]=\mathbb{P}\left[|Z(0)|>c\right]+\mathbb{P}\left[\left(U_c^Z[0,T]+D_c^Z[-T,0]\right)\mathbb{1}_{[|Z(0)|\leqslant c]}\geqslant 1\right],$$

or, pour une variable aléatoire ξ positive, on a $\mathbb{P}\{\xi\geqslant 1\}\leqslant\mathbb{E}(\xi)$. C'est cette inégalité que nous avons utilisée pour la borne supérieure. Pour la borne inférieure, on se sert du fait que lorsque ξ est une variable aléatoire positive à valeurs entières on a :

$$\mathbb{E}(\xi)-\frac{1}{2}\mathbb{E}(\xi(\xi-1))\leqslant\mathbb{P}\{\xi\geqslant 1\}.$$

On note $\mathbb{E}\left[\xi^{[2]}\right]$ le deuxième moment factoriel $\mathbb{E}(\xi(\xi-1))$ de ξ. Delmas [2001] fait remarquer que

$$\mathbb{E}\left[\left(U_c^Z[0,T]+D_c^Z[-T,0]\right)\mathbb{1}_{[|Z(0)|\leqslant c]}^{[2]}\right]\leqslant\mathbb{E}\left[\left(U_c^Z[0,T]+D_c^Z[-T,0]\right)^{[2]}\right].$$

Nous obtenons donc

$$\begin{aligned} & \mathbb{P}\left[|Z(0)|>c\right]+\mathbb{E}\left[\left(U_c^Z[0,T]+D_c^Z[-T,0]\right)\mathbb{1}_{[|Z(0)|\leqslant c]}\right]-\frac{1}{2}\mathbb{E}\left[\left(U_c^Z[0,T]+D_c^Z[-T,0]\right)^{[2]}\right] \\ \leqslant\ & \mathbb{P}\left[\sup_{t\in[0,T]}|Z(t)|>c\right]. \end{aligned}$$

Nous avons détaillé précédemment la majoration de (I.14) que l'on peut rappeler ici :

$$\mathbb{P}\left[\sup_{t\in[0,T]}|Z(t)|>c\right]\leqslant\mathbb{P}\left[|Z(0)|>c\right]+\mathbb{E}\left[\left(U_c^Z[0,T]+D_c^Z[-T,0]\right)\mathbb{1}_{[|Z(0)|\leqslant c]}\right].$$

Dans le cas d'un processus gaussien Z non stationnaire centré, de variance un, on remarque que

$$\begin{aligned} \mathbb{E}\left[D_c^Z[-T,0]\,\mathbb{1}_{[|Z(0)|\leqslant c]}\right] &= \mathbb{E}\left[U_c^Z[0,T]\,\mathbb{1}_{[|Z(0)|\leqslant c]}\right] \\ \mathbb{E}\left[D_c^Z[-T,0]^{[2]}\right] &= \mathbb{E}\left[U_c^Z[0,T]^{[2]}\right], \end{aligned}$$

ce qui n'est plus vrai lorsque Z est décentré. Cela nous permet de simplifier l'encadrement fourni par la méthode de Rice :

$$\begin{aligned} & \mathbb{P}\left[|Z(0)|>c\right]+2\mathbb{E}\left[U_c^Z[0,T]\,\mathbb{1}_{[|Z(0)|\leqslant c]}\right]-\mathbb{E}\left[U_c^Z[0,T]^{[2]}\right]-\mathbb{E}\left[U_c^Z[0,T]\,D_c^Z[-T,0]\right] \\ \leqslant\ & \mathbb{P}\left[\sup_{t\in[0,T]}|Z(t)|>c\right] \qquad (\text{I.15}) \\ \leqslant\ & \mathbb{P}\left[|Z(0)|>c\right]+2\mathbb{E}\left[U_c^Z[0,T]\,\mathbb{1}_{[|Z(0)|\leqslant c]}\right]. \end{aligned}$$

Tous ces termes sont explicités en détail dans Delmas [2001] pour un cadre plus général. L'expression générale de ces termes y est donnée pour un processus gaussien, stationnaire ou non, ainsi que pour un processus non centré. Nous donnons quand à nous les expression nécessaires au calcul de l'encadrement (I.15) dans le cas particulier d'un processus centré et de variance un. On note :

$$\gamma_{1,0}(\theta,\theta')=\left.\frac{\partial\gamma(\alpha,\beta)}{\partial\alpha}\right|_{(\theta,\theta')} \qquad \gamma_{1,1}(\theta,\theta')=\left.\frac{\partial\gamma(\alpha,\beta)}{\partial\alpha\partial\beta}\right|_{(\theta,\theta')}$$

$$\gamma_{1,0}(\theta,0)=\lim_{\substack{\theta'\to 0\\ \theta'>0}}\left.\frac{\partial\gamma(\alpha,\beta)}{\partial\alpha}\right|_{(\theta,\theta')} \qquad \gamma_{1,1}(\theta)=\gamma_{1,1}(\theta,\theta)$$

$$v^2=\gamma_{1,1}(\theta)-\frac{\gamma_{1,0}^2(\theta,0)}{1-\gamma^2(0,\theta)} \qquad \sigma^2(\theta,\theta')=\gamma_{1,1}(\theta)-\frac{\gamma_{1,0}(\theta,\theta')^2}{1-\gamma(\theta,\theta')^2}$$

$$\rho=\frac{\gamma_{1,1}(\theta,\theta')}{\sigma(\theta,\theta')\sigma(\theta',\theta)}+\frac{\gamma(\theta,\theta')\gamma_{1,0}(\theta,\theta')\gamma_{1,0}(\theta',\theta)}{\sigma(\theta,\theta')\sigma(\theta',\theta)\left(1-\gamma(\theta,\theta')^2\right)}$$

$$J_{11}(a,b,\rho)=(\rho+ab)\,J_{00}+\sqrt{1-\rho^2}\phi(b)\,\phi\left(\frac{\rho b-a}{\sqrt{1-\rho^2}}\right)+a\phi(b)\,\bar{\Phi}\left[\frac{\rho b-a}{\sqrt{1-\rho^2}}\right]+b\phi(a)\,\bar{\Phi}\left[\frac{\rho a-b}{\sqrt{1-\rho^2}}\right],$$

avec, pour (X,Y) un couple de variables aléatoires gaussiennes centrées de variance $\begin{pmatrix}1&\rho\\ \rho&\rho\end{pmatrix}$:

$$\begin{aligned} J_{00} &= P[X>a,Y>b] \\ &= \frac{1}{\pi}\arctan\left(\sqrt{\frac{1+\rho}{1-\rho}}\right)-\int_0^1\left(a\phi(at)\,\bar{\Phi}\left[\frac{b-\rho at}{\sqrt{1-\rho^2}}\right]+b\phi(bt)\,\bar{\Phi}\left[\frac{a-\rho bt}{\sqrt{1-\rho^2}}\right]\right)dt \end{aligned}$$

$$b(\theta,\theta')=c\left(1-\gamma(\theta,\theta')\right)\frac{\gamma_{1,0}(\theta,\theta')}{\sigma(\theta,\theta')\left(1-\gamma(\theta,\theta')^2\right)} \qquad m_1(\theta,\theta')=c\left(-1-\gamma(\theta,\theta')\right)\frac{\gamma_{1,0}(\theta,\theta')}{\sigma(\theta,\theta')\left(1-\gamma(\theta,\theta')^2\right)}$$

Termes de la borne inférieure.	
$\mathbb{E}\left(U_c^Z[0,T]\,\mathbb{1}_{[Z(0)>c]}\right)$	$F\left(\gamma(0,T),\gamma_{1,0}(T,0)\right) = \phi(c)\int_0^T \frac{\left(\gamma_{1,1}(\theta)\right)^{1/2}}{\sqrt{2\pi}}\bar{\Phi}\left[\frac{c}{v}\left(\gamma_{1,1}(\theta)\right)^{1/2}\sqrt{\frac{1-\gamma(0,\theta)}{1+\gamma(0,\theta)}}\right] + \frac{\gamma_{1,0}(\theta,0)}{\sqrt{1-\gamma(0,\theta)}}\phi\left(c\sqrt{\frac{1-\gamma(0,\theta)}{1+\gamma(0,\theta)}}\right)\bar{\Phi}\left(-\frac{c\gamma_{1,0}(\theta,0)}{v(1+\gamma(0,\theta))}\right)d\theta$
$\mathbb{E}\left(U_c^Z[0,T]\,\mathbb{1}_{[Z(0)\leqslant c]}\right)$	$\frac{1}{2\pi}e^{-\frac{c^2}{2}}\int_0^T\left(\gamma_{1,1}(\theta)\right)^{1/2}d\theta - F\left(\gamma(0,T),\gamma_{1,0}(T,0)\right)$
$\mathbb{E}\left(U_c^Z[0,T]\,\mathbb{1}_{[\lvert Z(0)\rvert\leqslant c]}\right)$	$\frac{1}{2\pi}e^{-\frac{c^2}{2}}\int_0^T\left(\gamma_{1,1}(\theta)\right)^{1/2}d\theta - F\left(\gamma(0,T),\gamma_{1,0}(T,0)\right) - F\left(-\gamma(0,T),-\gamma_{1,0}(T,0)\right)$
$\mathbb{E}\left[U_c^Z[0,T]^{[2]}\right]$	$\int_0^T\int_0^T \sigma(\theta,\theta')\,\sigma(\theta',\theta)\,J_{11}\left(b(\theta,\theta'),b(\theta',\theta),\rho\right)\frac{\phi^2\left(c\diagup\left(\sqrt{1+\gamma(\theta,\theta')}\right)\right)}{\sqrt{1-\gamma^2(\theta,\theta')}}d\theta d\theta'$
$\mathbb{E}\left[U_c^Z[0,T]\,D_c^Z[-T,0]\right]$	$\int_0^T\int_0^T \sigma(\theta,\theta')\,\sigma(\theta',\theta)\,J_{11}\left(m_1(\theta,\theta'),m_1(\theta',\theta),\rho\right)\frac{\phi^2\left(c\diagup\left(\sqrt{1-\gamma(\theta,\theta')}\right)\right)}{\sqrt{1-\gamma^2(\theta,\theta')}}d\theta d\theta'$

De manière similaire à Delmas [2001], nous avons rencontré beaucoup de problèmes dans le calcul du terme $F\left(\gamma(0,T),\gamma_{1,0}(T,0)\right)$ dont l'expression exacte est donnée page 24. Ce terme se présente sous la forme d'une intégrale dont l'intégrand comporte presque uniquement des formes indéterminées qu'il nous a fallu développer à l'aide du logiciel Maple. En effet, une analyse un peu précise (cf figures I.3 et I.4) nous fait prendre conscience de la difficulté de l'intégration au voisinage de 1. Les irrégularités présentes sur le graphique (I.4) apportent la preuve qu'un logiciel comme Maple a énormément de difficultés à appréhender l'évolution de cette fonction , pourtant continue et dérivable au voisinage du point singulier. Le même calcul effectué en Fortran, double précision, montre les mêmes signes d'instabilité.

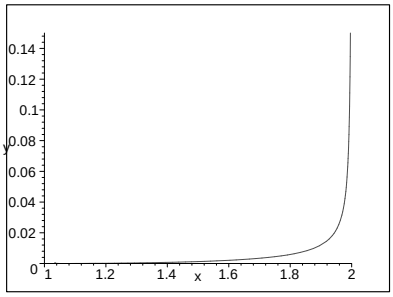


FIG. I.3 – Tracé de la fonction à intégrer, $F\left(\gamma(0,T),\gamma_{1,0}(T,0)\right)$, intervenant dans le calcul de la borne de Rice.

Nous pouvons constater cependant (figure I.5) que l'utilisation d'un développement limité des différents termes élimine presque totalement les irrégularités de l'intégrand. On pourra com-

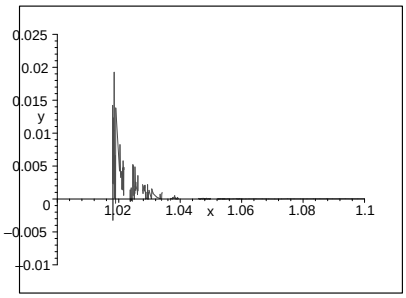


FIG. I.4 – Tracé de la fonction à intégrer, $F\left(\gamma\left(0,T\right),\gamma_{1,0}\left(T,0\right)\right)$, intervenant dans le calcul de la borne de Rice. Détail des perturbations intervenants au voisinage du point singulier.

prendre l'intérêt d'utiliser Maple en soulignant le fait que les développements nécessaires vont parfois jusqu'à l'ordre 15.

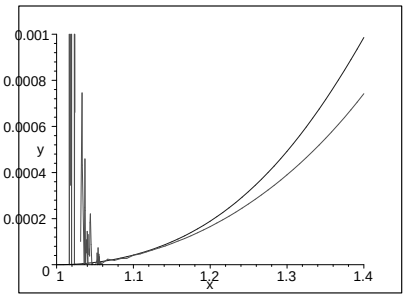


FIG. I.5 – Tracé de la fonction à intégrer, $F\left(\gamma\left(0,T\right),\gamma_{1,0}\left(T,0\right)\right)$, intervenant dans le calcul de la borne de Rice, comparé au tracé de son développement limité.

Le programme contenant les calculs détaillés du terme $F\left(\gamma\left(0,T\right),\gamma_{1,0}\left(T,0\right)\right)$ ainsi que les graphiques présentés ici est fourni en annexe. Comme on a pu le voir avec les tableaux I.1 et I.2, et comme le constate aussi Delmas [2001], la différence entre les valeurs critiques obtenues par la méthode de Davies ou par celle de Rice sont très proches. De même, selon Delmas [2001] (page 95 par ex), la borne inférieure que l'on vient de détailler dans le cas de la méthode de Rice fournit un encadrement d'amplitude très petite de la véritable valeur critique.

Dans le cas non étudié par Delmas [2001] d'un mélange gaussien sur les variances, et au vu des résultats obtenus, il est raisonnable de conjecturer que l'encadrement est aussi très précis. Les approximations des valeurs critiques fournies en I.1 et I.2 sont donc suffisantes. On constate cependant un phénomène de convergence assez lente de la statistique asymptotique du TRMV.

On peut remarquer pour finir qu'il semble possible de développer une borne inférieure en utilisant les techniques de Davies, plus simples et plus rapides à mettre en oeuvre.

Chapitre II

Introduction aux méthodes de Monte Carlo par chaînes de Markov

Les méthodes de Monte Carlo ont été développées pour calculer de manière numérique, et donc approchée, certaines intégrales dont l'expression théorique rendait les calculs rédhibitoires. Depuis que les ordinateurs sont devenus suffisamment puissants, ces méthodes de calcul sont maintenant largement utilisées, et l'on trouve un vaste choix de livres traitant de ce sujet. On pourra citer Gentle [2002] et Chen et al. [2000], ce dernier traitant de manière approfondie les méthodes de Monte Carlo appliquées à l'analyse Bayésienne. Les chaînes de Markov offrent quant à elles un moyen aisé de simuler des valeurs distribuées asymptotiquement selon la loi désirée. La combinaison de ces deux aspects donne lieu à ce que l'on appelle les méthodes de Monte Carlo par chaînes de Markov (MCMC). Il devient alors aisé de calculer certaines intégrales dépendant de lois à partir desquelles on ne peut simuler directement. Robert [1996a] et Robert et Casella [1999] constituent des ouvrages de référence très complets sur le sujet.

Ce chapitre a pour but d'introduire les quelques notions de bases liées aux techniques MCMC en prenant un exemple illustratif des difficultés rencontrées avant l'émergence de ces méthodes. Après un bref rappel théorique sur les chaînes de Markov, nous abordons la construction de certains algorithmes MCMC les plus utilisés.

En dernier lieu, et afin d'introduire les éléments graphiques des programmes utilisés dans les chapitres suivants, nous présentons des applications de ces algorithmes aux mélanges gaussiens univariés et multivariés.

1 Problèmes inhérents à l'approche bayésienne

Nous introduisons ici quelques unes des difficultés rencontrées par l'approche bayesienne ayant trouvées leurs solutions avec les méthodes MCMC. L'approche bayésienne considère que l'information apportée par les données y est résumée par une loi de probabilité $\pi(\theta \mid y)$, dite loi a posteriori, qui se déduit de la loi jointe $\pi(\theta) f(y \mid \theta)$ par calcul de la loi marginale de y notée $m(y)$:

$$\pi(\theta \mid y) = \frac{\pi(\theta) f(y \mid \theta)}{m(y)} \quad \text{avec} \quad m(y) = \int \pi(\theta) f(y \mid \theta) \, d\theta,$$

où $\pi(\theta)$ est appelée loi a priori sur le paramètre θ. On est alors amené à évaluer des estimateurs d'une fonction de θ, notée $h(\theta)$, sous une certaine fonction de coût $L(\delta, \theta)$, ceci conduisant à minimiser en δ le coût a posteriori :

$$\int L(\delta, \theta) \pi(\theta \mid y) \, d\theta.$$

Dans le cas particulier d'un coût quadratique

$$L(\delta, \theta) = \|h(\theta) - \delta\|^2,$$

l'estimateur de Bayes de la fonction $h(\theta)$ minimisant le coût a posteriori est alors

$$\begin{aligned} \delta^{\pi}(y) &= \mathbb{E}^{\pi}[h(\theta) \mid y] \\ &= \int h(\theta)\, \pi(\theta \mid y)\, d\theta. \end{aligned}$$

On est ainsi conduit à évaluer une intégrale par rapport à la mesure $\pi(\theta \mid y)\, d\theta$. Cet exemple nous permet d'illustrer les problèmes liés à l'utilisation de l'approche bayésienne, problèmes d'intégration essentiellement. En effet, les difficultés liées au calcul de $\delta^{\pi}(y)$ sont, d'une part que $\pi(\theta \mid y)$ n'est généralement pas connue sous forme explicite et, d'autre part, que l'intégration de $h(\theta)$ suivant $\pi(\theta \mid y)$ ne peut pas être faite analytiquement dans la plupart des cas.
Tout ceci permet donc de mettre en évidence les problèmes d'intégration comme faisant partie des principaux écueils de la statistique bayésienne ; un autre étant constitué par les problèmes de minimisation.

C'est cette classe de problèmes que se proposent d'aborder les méthodes MCMC.

2 Intégration par la méthode de Monte Carlo et échantillonnage pondéré

Une première solution est donnée par les méthodes d'approximation d'intégrales de la forme

$$\mathbb{E}^{f}[h(y)] = \int h(y)\, f(y)\, dy, \tag{II.1}$$

appelées méthodes de Monte Carlo. Le principe est de proposer l'utilisation d'un échantillon $(y_1, ..., y_m)$ généré suivant la densité $f(y)$ afin approcher l'intégrale II.1 par la moyenne empirique

$$\bar{h}_m = \frac{1}{m} \sum_{j=1}^{m} h(y_j). \tag{II.2}$$

En effet, d'après la loi forte des grands nombres, $\bar{h}_m$ converge presque sûrement vers $\mathbb{E}^{f}[h(y)]$.
Une seconde solution consiste à approcher ce type d'intégrales sans avoir à simuler selon la loi de densité f. Ces méthodes sont dites d'échantillonnage pondéré.

Afin de mieux comprendre l'intérêt de ce genre de méthodes, nous allons étudier un exemple introductif tiré de Ripley [1987].

Exemple 2 : Soit à calculer pour une loi de Cauchy $\mathcal{C}(1)$, (voir la définition en Annexe A), la probabilité de dépasser la valeur 2

$$p = \int_2^{+\infty} \frac{1}{\pi(1 + x^2)} dx = \frac{1}{2} - \frac{arctg(2)}{\pi}.$$

On approche classiquement p par

$$\hat{p}_1 = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{[x_j > 2]},$$

à partir d'un échantillon $x_1, ..., x_m$ iid issu d'une $\mathcal{C}(1)$. La variance de $\hat{p}_1$ est alors $\frac{p(1-p)}{m}$, soit $\frac{0.127}{m}$ car $p = 0.15$. Lorsqu'on écrit p sous la forme suivante

$$p = \int_0^{1/2} \frac{y^{-2}}{\pi(1+y^{-2})} dy,$$

on remarque que c'est aussi l'espérance de $h(Y) = \frac{2}{\pi(1+Y^2)}$ avec Y suivant une loi uniforme sur $[0, 1/2]$. On peut donc évaluer p par

$$\hat{p}_2 = \frac{1}{m} \sum_{j=1}^{m} h(y_j),$$

les y_j provenant d'un échantillon de loi uniforme sur $[0, 1/2]$. Une intégration par partie sur la variance de $\hat{p}_2$, qui est $\left(\mathbb{E}\left[h^2\right] - \mathbb{E}\left[h\right]^2\right) / m$, montre que celle-ci est alors de $0.95.10^{-4} / m$. On a donc divisé la première variance par un facteur d'ordre 10^{-3}.

Définition 3 : On appelle échantillonnage pondéré toute méthode permettant d'approcher (II.1) à partir d'un échantillon $y_1, ..., y_m$ généré suivant une loi de densité g, par l'approximation

$$\bar{h}_m = \frac{1}{m} \sum_{j=1}^{m} \frac{f(y_j)}{g(y_j)} h(y_j).$$

Cette définition utilise une représentation alternative de (II.1) à savoir

$$\int h(y) \frac{f(y)}{g(y)} g(y) dy,$$

la convergence étant ainsi obtenue pour les mêmes raisons que pour les méthodes de Monte Carlo. Cette méthode est intéressante au sens où elle permet une latitude presque totale dans le choix de la loi g, celle-ci pouvant être sélectionnée parmi des lois facilement simulables, tout en autorisant l'utilisation d'un même échantillon (généré suivant g) non seulement pour diverses fonctions h, mais également pour diverses lois f. Remarquons qu'il est possible de déterminer la loi optimale g pour une fonction h et une loi f données. Pour plus de détails, on pourra se référer par exemple à Robert [1996a].

3 Méthodes de Monte Carlo par chaînes de Markov

3.1 Introduction

Nous avons vu dans la section précédente qu'il n'est pas nécessaire de simuler un échantillon suivant la loi f pour approcher l'intégrale $\int h(y) f(y) dy$, puisque la méthode d'échantillonnage pondéré permet de faire appel à une autre distribution que f. Nous allons étudier cette possibilité en montrant qu'il est possible d'obtenir un échantillon $y_1, ..., y_m$ de loi f, sans simuler directement suivant f.

A ce stade de l'étude des méthodes MCMC, il est nécessaire de rappeler quelques résultats et définitions propres aux chaînes de Markov dont nous aurons besoin par la suite.

3.2 Notions sur les chaînes de Markov

Une chaîne de Markov homogène à temps discret et d'espace d'état E continu est définie par une séquence de variables aléatoires $\left(X^{(0)}, X^{(1)}, ...\right)$, avec $X^{(t)} \in E$, vérifiant la **propriété de Markov** en temps. Celle-ci signifie que, étant donné l'état courant $X^{(t)}$ $(t \geqslant 0)$, la loi de $X^{(t+1)}$ est indépendante du passé de la chaîne $\left(X^{(0)}, ..., X^{(t-1)}\right)$, l'**homogénéité** signifiant l'invariance de cette propriété par rapport au temps.

Remarque 2 *: L'espace d'état peut être très complexe et contenir aussi bien des espaces discrets que des espaces continus.*

La loi d'une chaîne de Markov homogène en temps (discret) $\{X^{(t)}\}$ est alors caractérisée par l'état initial $X^{(0)}$ et par le noyau de transition

$$P(x, A) = \mathbb{P}\left(X^{(t+1)} \in A \mid X^{(t)} = x\right) \qquad \text{pour } A \subset E.$$

Définition 4 Un noyau de transition est une fonction P définie sur $E \times \mathcal{B}(E)$ ($\mathcal{B}(E)$ désignant la tribu Borélienne) à valeurs dans $[0, 1]$ telle que :

1. $\forall x \in E$, $P(x, .)$ est une mesure de probabilité,
2. $\forall A \in \mathcal{B}(E)$, $P(., A)$ est mesurable.

Ce n'est rien d'autre qu'une probabilité sur $\mathcal{B}(E)$, dépendant mesurablement du paramètre x. Lorsque l'espace E est discret, le noyau de transition est simplement une matrice de transition ayant pour éléments $P_{xy} = \mathbb{P}\left(X^{(t+1)} = y \mid X^{(t)} = x\right)$ pour $x, y \in E$. Dans le cas continu avce densité, le noyau représente la densité conditionnelle $P(x, x')$ de la transition. C'est-à-dire que

$$\mathbb{P}(X \in A \mid x) = \int_A P(x, x')\, dx'.$$

Remarque 3 *: On ne s'intéresse ici qu'aux chaînes de Markov homogènes et n'ayant pas de comportement périodique.*

Si $X^{(t)}$ est distribuée selon ν sur E, alors $X^{(t+1)}$ est distribuée selon νP donnée par

$$\nu P(A) = \int P(x, A)\, \nu(dx).$$

Une chaîne de Markov est de **mesure invariante** π si $\pi = \pi P$ c'est-à-dire :

$$X^{(t)} \sim \pi \Longrightarrow X^{(t+1)} \sim \pi.$$

Il nous faut déterminer les conditions nécessaires pour pouvoir obtenir un échantillon distribué selon la loi π, ceci dans le but d'approcher des intégrales par rapport à cette loi. On voit facilement qu'une propriété nécessaire est que la chaîne $X^{(t)}$ générée puisse atteindre tous les points du support de π. C'est ce que l'on appelle précisément la propriété d'irréductibilité, indiquant qu'il est alors possible pour la chaîne de rejoindre 2 états quelconques avec une probabilité non nulle.

Définition 5 : Une chaîne de Markov est dite **irréductible** (ou ϕ-irréductible) s'il existe une densité ϕ sur E telle que pour tout $A \subset E$ avec $\phi(A) > 0$

$$\mathbb{P}\left(\exists t > 0 \;\; X^{(t+1)} \in A \mid X^{(0)} = x\right) > 0 \qquad \text{pour tout } x \in E.$$

Lorsqu'il existe une mesure de probabilité invariante pour une chaîne irréductible, la chaîne est dite **positive**. Dans un cadre discret, la **récurrence d'un état** se définit comme une garantie de retour, c'est-à-dire que la probabilité d'y revenir est égale à un. On devine donc que cette notion est automatiquement satisfaite pour les chaînes irréductibles sur un espace d'état fini. On dit alors que la chaîne $X^{(t)}$ est récurrente et irréductible. Lorsque l'espérance du temps de retour a l'état x est finie, on dit que cet état est **récurrent positif**.

Lorsque la chaîne est à espace d'état continu, il faut utiliser à la place une notion sensiblement plus forte appelée **récurrence au sens de Harris** dont on pourra trouver les détails dans Robert [1996a]. Si l'on ajoute à cette propriété le fait que la chaîne est positive et apériodique, on obtient alors ce qui est appelé une chaîne **ergodique**, cette notion assurant la convergence de la chaîne vers une loi limite.

Nous allons maintenant introduire le concept de chaîne de Markov **réversible**, définition que l'on peut résumer par la condition suivante

$$\int_A P(x, B)\, \pi(x)\, dx = \int_B P(x, A)\, \pi(x)\, dx, \tag{II.3}$$

pour tout $x \in E$ et tout A et B inclus dans E.

Cette condition signifie que le taux auquel la chaîne se déplace de x à y (i.e. $\pi(x) P(x, y)$) lorsqu'elle est à l'équilibre (lorsque la convergence vers la distribution stationnaire est atteinte), est égal au taux auquel elle se déplace de y à x (i.e. $\pi(y) P(y, x)$). La condition ci-dessus est souvent appelée **condition de balance**. Les chaînes de Markov vérifiant cette condition sont appelées des chaînes réversibles.

La réversibilité permet de montrer des résultats de type théorème de la limite centrale (cf Th 3.29 de Robert [1996a]). Les conditions nécessaires à ce type de théorème sont assez contraignantes, mais la réversibilité permet de les obtenir. L'intérêt est que la réversibilité, généralement très contraignante, est facile à imposer dans la plupart des algorithmes de Monte Carlo par chaînes de Markov grâce à des étapes de simulation supplémentaires (cf Green [1995] ou Tierney [1994]).

Les chaînes réversibles sont très utiles car pour toute chaîne irréductible de noyau de transition P, si π est une distribution vérifiant (II.3), alors la chaîne est récurrente au sens de Harris, réversible et de loi stationnaire π.

De manière intuitive, on peut considérer qu'une chaîne de Markov récurrente irréductible de loi invariante π va nous permettre d'approcher les intégrales de la forme $\int_E h(y) f(y)\, dy$. C'est ce que précise le théorème suivant (Tierney [1996], p65) :

Théorème 4 : Soit $\left(X^{(0)}, ..., X^{(t-1)}\right)$ une chaîne de Markov irréductible sur E de noyau de transition P et de distribution invariante π. Soit h une fonction à valeurs réelles sur E $\pi-$intégrable. Alors

$$\mathbb{P}\left(\frac{1}{m}\sum_{j=1}^{m} h(x_j) \longrightarrow \int_E h(x)\, \pi(x)\, dx \,\middle|\, X^{(0)} = x^{(0)}\right) = 1$$

pour presque tout $x^{(0)}$.

Ce théorème implique notamment que sous certaines conditions de régularité pour h, la moyenne empirique de l'échantillon généré par une chaîne de Markov de distribution invariante $p\left(x \mid y^{(n)}\right)$, converge presque sûrement vers $\mathbb{E}\left(h(X) \mid y^{(n)}\right)$.

3.3 Quelques algorithmes MCMC

Le principe fondamental des méthodes MCMC consiste donc à utiliser une chaîne de Markov irréductible vérifiant le théorème précédent. Partant d'une valeur initiale arbitraire $x^{(0)} \in E$, on

génère une chaîne $\left(x^{(t)}\right)$ à partir d'un noyau de transition de loi stationnaire f, qui garantit de plus la convergence en loi vers f. Pour T "assez grand", on peut considérer $x^{(T)}$ comme étant distribué suivant f et obtenir ainsi un échantillon $(x^{(T)}, x^{(T+1)}...)$ qui est effectivement distribué suivant f, même si les $x^{(T+t)}$ ne sont pas indépendants.

Définition 6 : On appelle algorithme MCMC toute méthode produisant une chaîne de Markov $\left(x^{(t)}\right)$ ergodique dont la loi stationnaire est la distribution d'intérêt.

Une telle approche peut sembler peu naturelle, mais elle a le mérite de proposer une structure "universelle" de simulation applicable à des modèles trop complexes pour autoriser un traitement satisfaisant par une autre approche. Le problème de l'identification des mélanges gaussiens est d'ailleurs typique de ces modèles n'admettant ni traitement analytique, ni approximations numériques dans les cadres classiques (maximum de vraisemblance etc...).

Nous allons aborder deux grands types de méthodes MCMC : les méthodes de Metropolis-Hastings, très générales, et l'échantillonnage de Gibbs. Ce dernier est d'une importance pratique et historique considérable.

3.3.1 L'algorithme de Metropolis-Hastings

Il repose sur l'utilisation d'une densité conditionnelle $q\left(y \mid x\right)$ par rapport à la mesure dominante pour le modèle, et ne peut être mis en pratique que si $q\left(. \mid x\right)$ est soit simulable rapidement, soit disponible analytiquement. Remarquons que dans le cadre de cette thèse, la mesure dominante sera toujours la mesure de Lebesgue.

L'algorithme de Metropolis-Hastings associé à la loi objectif f et la loi conditionnelle q produit une chaîne de Markov $\left(x^{(t)}\right)$ générée par l'algorithme suivant :

étant donné $x^{(t)}$

1- Générer $y_t \sim q\left(y \mid x^{(t)}\right)$

2- Prendre $$x^{(t+1)} = \begin{cases} y_t \text{ avec probabilité } \rho\left(x^{(t)}, y_t\right) \\ x^{(t)} \text{ avec probabilité } 1 - \rho\left(x^{(t)}, y_t\right) \end{cases}$$

où $\rho\left(x^{(t)}, y_t\right) = \min\left[1, \frac{f(y_t)}{f\left(x^{(t)}\right)} \frac{q\left(x^{(t)} \mid y_t\right)}{q\left(y_t \mid x^{(t)}\right)}\right]$.

Remarque 4 *: q est appelée la loi instrumentale*

Il est nécessaire d'imposer des conditions minimales sur loi conditionnelle q pour que f soit effectivement la loi limite de la chaîne $\left(x^{(t)}\right)$ produite. Par exemple, si $\mathcal{E}$, support de f, que nous supposerons *connexe*, est systématiquement tronqué par q, c'est-à-dire s'il existe $A \subset \mathcal{E}$ tel que

$$\int_A f\left(x\right) dx > 0 \quad \text{et} \quad \int_A q\left(y \mid x\right) dy = 0 \qquad \forall x \in \mathcal{E},$$

l'algorithme de Metropolis-Hastings ne peut admettre f comme loi stationnaire puisque, partant de $x^{(0)} \notin A$, la chaîne $\left(x^{(t)}\right)$ ne visite jamais A. La condition nécessaire et minimale à imposer est alors :

$$\bigcup_{x \in \text{supp } f} \text{supp } q\left(. \mid x\right) \supset \text{supp } f,$$

en supposant de plus que supp f est connexe.

Sous la condition précédente, la chaîne de Markov définie par l'algorithme de Metropolis-Hastings admet le noyau de transition suivant :

$$P(x,A) = \mathbb{1}_{[x\notin A]} \int_A q(y \mid x)\, \rho(x,y)\, dy + \mathbb{1}_{[x\in A]} \left[1 - \int_A q(y \mid x)\, \rho(x,y)\, dy\right].$$

Ceci représente la probabilité de se déplacer vers l'ensemble $A \subset E$ en partant de n'importe quel x, le premier terme étant la probabilité d'arriver dans A en partant de $x \notin A$, le second, la probabilité de rester dans A alors qu'on y était déjà. Ce noyau de transition définit alors une chaîne réversible ayant donc f pour loi stationnaire, ce que l'on peut résumer par le théorème suivant :

Théorème 5 : Pour toute loi conditionnelle q, vérifiant la condition minimale, f est une loi stationnaire de la chaîne $\left(x^{(t)}\right)$ produite par l'algorithme de Metropolis-Hastings.

Preuve *On désire vérifier la condition de balance (II.3).*

On a

$$\begin{aligned}\int_B P(x,A)\, f(x)\, dx &= \int\int \mathbb{1}_{[x\in B]}\mathbb{1}_{[y\in A]} q(y \mid x)\, \rho(x,y)\, f(x)\, dxdy \\ &\quad + \int\int \mathbb{1}_{[x\in B]}\mathbb{1}_{[x\in A]} q(y \mid x)\, [1-\rho(x,y)]\, f(x)\, dxdy.\end{aligned}$$

On s'intéresse ici au premier terme. Définissons l'ensemble suivant

$$D = \{(x,y) \nearrow f(y)\, q(x \mid y) < f(x)\, q(y \mid x)\}.$$

Sur cet ensemble, $\rho(x,y) = \frac{f(y)q(x|y)}{f(x)q(y|x)}$, *sinon* $\rho(x,y) = 1$. *Le premier terme de l'égalité précédente s'écrit donc*

$$\int\int \mathbb{1}_{[x\in B]}\mathbb{1}_{[y\in A]}\mathbb{1}_{[(x,y)\in D]} f(y)\, q(x \mid y)\, dxdy + \int\int \mathbb{1}_{[x\in B]}\mathbb{1}_{[y\in A]}\mathbb{1}_{[(x,y)\in D^c]} q(y \mid x)\, f(x)\, dxdy$$

$$= \int\int \mathbb{1}_{[y\in B]}\mathbb{1}_{[x\in A]}\mathbb{1}_{[(x,y)\in D^c]} f(x)\, q(y \mid x)\, dxdy + \int\int \mathbb{1}_{[y\in B]}\mathbb{1}_{[x\in A]}\mathbb{1}_{[(x,y)\in D]} q(x \mid y)\, f(y)\, dxdy$$

$$= \int\int \mathbb{1}_{[y\in B]}\mathbb{1}_{[x\in A]} f(x)\, q(y \mid x)\, \rho(x,y)\, dxdy.$$

En effet, le changement de variable qui passe de (x,y) *à* (y,x), *change* D *en* D^c *et réciproquement. On retrouve alors le premier terme de* $\int_A P(x,B)\, f(x)\, dx$, *le second terme ne changeant pas.* ■

Ce théorème nous donne une mesure de l'universalité de ce type d'algorithmes. Afin de vérifier que celui présenté entre bien dans le cadre des algorithmes MCMC, il faut établir l'ergodicité de $\left(x^{(t)}\right)$. En se référant à Tierney [1994] et Robert [1996a], on apprend qu'il suffit de montrer la $f-$irréductibilité et l'apériodicité de $\left(x^{(t)}\right)$.

L'irréductibilité de $\left(x^{(t)}\right)$ découle de conditions suffisantes comme la positivité : $q(y \mid x) > 0$ pour tout $(x,y) \in E^2$.

L'apériodicité semble évidente puisque l'algorithme autorise formellement les événements $\left\{x^{(t+1)} = x^{(t)}\right\}$. Il faut cependant que la probabilité de ces événements soit non nulle, c'est-à-dire :

$$\mathbb{P}\left[f\left(x^{(t)}\right) q\left(y_t \mid x^{(t)}\right) = f(y_t)\, q\left(x^{(t)} \mid y_t\right)\right] < 1.$$

Cette condition signifie que q ne doit pas être le noyau de transition d'une chaîne de Markov réversible admettant f comme loi stationnaire. L'échantillonnage de Gibbs que l'on abordera dans la section suivante ne vérifie pas cette dernière condition. En effet, nous verrons que celui-ci accepte automatiquement toute nouvelle génération d'état.

Afin de mieux comprendre le fonctionnement des méthodes MCMC à sauts réversibles introduites dans le chapitre suivant, il est nécessaire d'aborder ici une version hybride de l'algorithme de Metropolis-Hastings.

3.3.2 Algorithme de Metropolis-Hastings composant par composant

Les algorithmes habituels mettent à jour tout les composants de $x^{(t)}$ à la fois. Il peut parfois être utile de mettre à jour les différents composants de $x^{(t)} = \left(x_1^{(t)}, ..., x_d^{(t)}\right)$ séparément, c'est ce que fait l'algorithme présenté ici. Les composants seront mis à jour selon leur ordre d'indiçage, c'est-à-dire que l'on mettra successivement à jour $x_1^{(t)}, x_2^{(t)}, ...,$ et $x_d^{(t)}$. On notera les lois conditionnelles $f_i\left(x_i^{(t)}\right) = f\left(x_i^{(t)} \mid x_{-i}^{(t)}\right)$ avec $x_{-i}^{(t)} = \left(x_1^{(t)}, ..., x_{i-1}^{(t)}, x_{i+1}^{(t)}, ..., x_d^{(t)}\right)$.

La valeur y_i utilisée pour la mise à jour de $x_i^{(t)}$ est générée selon la densité notée $g_i\left(y_i \mid x_i^{(t)}\right)$ même si elle peut dépendre de $x_{-i}^{(t)}$. L'algorithme de mise à jour du $i^{ème}$ composant à l'étape t est alors :

1- `Générer` $\quad y_i \sim g_i\left(y_i \mid x_i^{(t)}\right)$

2- `Prendre` $\quad x_i^{(t+1)} = \begin{cases} y_i \text{ `avec probabilité`} \min\left[1, \frac{f_i(y_i) g_i\left(y_i \mid x_i^{(t)}\right)}{f_i\left(x_i^{(t)}\right) g_i\left(x_i^{(t)} \mid y_i\right)}\right] \\ x_i^{(t)} \text{ `sinon`} \end{cases}$

En fait, c'est sous cette forme qu'était présenté l'algorithme dans l'article original de Metropolis et al (1953, Metropolis et al. [1953]). C'est un compromis entre l'algorithme de Metropolis-Hastings présenté plus haut et celui de l'échantillonnage de Gibbs que nous allons voir. Tous sont en fait des cas particuliers des algorithmes à sauts réversibles du chapitre suivant.

3.3.3 L'échantillonnage de Gibbs

Formellement, celui-ci peut se décrire comme un cas particulier de l'algorithme de Metropolis-Hastings, cependant il s'en distingue par plusieurs caractéristiques.

1. Le taux d'acceptation est uniformément égal à 1 (Toutes les valeurs simulées sont acceptées).
2. Son utilisation entraîne des limitations fortes sur le choix des paramètres des lois instrumentales et suppose la connaissance de certaines propriétés de f.
3. L'échantillonnage de Gibbs ne fonctionne pas lorsque le nombre de paramètres est variable (cf méthodes MCMC à sauts réversibles du chapitre suivant).

Définissons maintenant la structure de cet algorithme.

Définition 7 : Etant donné une densité f, une densité g satisfaisant

$$\int_E g(x, z)\, dz = f(x),$$

est appelée une complétion de f.

Soit f la densité ciblée par l'algorithme. On choisit alors une complétion de f de manière à ce que ses lois conditionnelles soient faciles a simuler. On échantillonnera g plutôt que f. Pour $p > 1$, en notant $y = (x, z)$, les lois conditionnelles de $g(y) = g(y_1, ..., y_p)$ s'écrivent

$$g_1(y_1 \mid y_2, ..., y_p), g_2(y_2 \mid y_1, y_3, ..., y_p), ..., g_p(y_p \mid y_1, ..., y_{p-1}).$$

L'algorithme d'échantillonnage de Gibbs associé à cette décomposition est alors fourni par la transition de $y^{(t)}$ à $y^{(t+1)}$ suivante :

```
Simuler
1 -  y_1^(t+1) ~ g_1 ( y_1 | y_2^(t), ..., y_p^(t) )
2 -  y_2^(t+1) ~ g_2 ( y_2 | y_1^(t+1), y_3^(t), ..., y_p^(t) )
⋮
p -  y_p^(t+1) ~ g_p ( y_p | y_1^(t+1), ..., y_{p-1}^(t+1) )
```

Cette technique de complétion d'une densité f en une densité g telle que f apparaisse comme densité marginale de g est souvent naturelle, comme dans le cas des modèles à données manquantes que nous aborderons par la suite. On peut noter la similarité de cette approche avec l'algorithme EM de maximisation de la vraisemblance. De même, il faut remarquer que la méthode d'échantillonnage de Gibbs ne requiert aucunement que la complétion de f en g et de x en $y = (x, z)$ soit reliée au problème inférentiel. Ceci signifie que cette façon d'aborder les choses est justifiée techniquement sans qu'il soit nécessaire de la relier au problème réel sous-jacent.

Robert et Casella [1999] montrent que la plupart des algorithmes de Gibbs vérifient les conditions minimales nécessaire à l'ergodicité de la chaîne. Pour ce qui est des propriétés de convergence de cet algorithme, il faut montrer que la chaîne $\left(y^{(t)}\right)$ converge vers la distribution g et que $\left(x^{(t)}\right)$ est une chaîne qui converge vers f. En effet, même si $\left(y^{(t)}\right)$ est par construction une chaîne de Markov, $\left(x^{(t)}\right)$ n'en est pas forcément une.

Théorème 6 Si la chaîne $\left(y^{(t)}\right)$ produite par l'algorithme de Gibbs est ergodique, alors la distribution g est stationnaire pour la chaîne $\left(y^{(t)}\right)$ et f est la distribution limite de la chaîne $\left(x^{(t)}\right)$.

On pourra trouver la preuve de ce théorème dans Robert et Casella [1999]. Afin de compléter le parallèle entre l'échantillonnage de Gibbs et les méthodes de Metropolis-Hastings, on peut citer le théorème suivant. Cf Robert et Casella [1999] p. 296.

Théorème 7 : La méthode d'échantillonnage de Gibbs décrite plus haut correspond à la composition de p algorithmes de Metropolis-Hastings de probabilités d'acceptation égales à 1.

Preuve *On peut considérer l'échantillonnage de Gibbs comme la composition de p algorithmes "élémentaires" correspondant chacun à la simulation suivant une des lois conditionnelles de g. Il suffit donc de montrer que chacun de ces algorithmes a une probabilité d'acceptation égale à 1.*

Considérons l'étape i. La loi instrumentale s'écrit

$$q_i(y' \mid y) = \delta_{(y_1,...,y_{i-1},y_{i+1},...,y_p)}\left(y'_1, ..., y'_{i-1}, y'_{i+1}, ..., y'_p\right) g_i(y'_i \mid y_1, ..., y_{i-1}, y_{i+1}, ..., y_p).$$

Le rapport définissant la probabilité $\rho(y, y')$ s'écrit alors

$$\begin{aligned}
\frac{g(y')\, q_i(y \mid y')}{g(y)\, q_i(y' \mid y)} &= \frac{g(y')\, g_i\left(y_i \mid y'_1, ..., y'_{i-1}, y'_{i+1}, ..., y'_p\right)}{g(y)\, g_i(y'_i \mid y_1, ..., y_{i-1}, y_{i+1}, ..., y_p)} \\
&= \frac{g_i(y'_i \mid y_1, ..., y_{i-1}, y_{i+1}, ..., y_p)\, g_i\left(y_i \mid y'_1, ..., y'_{i-1}, y'_{i+1}, ..., y'_p\right)}{g_i\left(y_i \mid y'_1, ..., y'_{i-1}, y'_{i+1}, ..., y'_p\right) g_i(y'_i \mid y_1, ..., y_{i-1}, y_{i+1}, ..., y_p)} \\
&= 1.
\end{aligned}$$

■

3.4 Considérations pratiques : initialisation, temps de chauffe et convergence

En tant qu'algorithmes MCMC, les diverses méthodes présentées ici nécessitent toutes le choix d'une initialisation $x^{(0)}$. De manière idéale, on devrait choisir $x^{(0)}$ selon la loi invariante f ciblée par l'algorithme. Cela est rarement possible, c'est pourquoi nous utiliserons généralement une valeur au hasard tirée selon la loi a priori. L'influence du point de départ de l'algorithme devient vite négligeable. Toutefois, pour réduire au maximum l'influence de $x^{(0)}$ sur l'estimation de la densité ciblée, il est d'usage de ne pas tenir compte d'un certain nombre m d'itérations, ceci signifiant que les m premières valeurs générées par l'algorithme ne seront tout simplement pas prises en compte. Les auteurs anglophones font alors référence à la période de "burn in" ou "temps de chauffe".

Si la chaîne de Markov générée est capable de mouvements rapides et amples permettant de se déplacer rapidement dans tout l'espace d'état, alors la valeur générée $x^{(m)}$ pourra être considérée comme étant indépendante de $x^{(0)}$ pour m relativement grand. Une chaîne ayant de bonnes propriétés de mélange ne nécessitera qu'un temps de chauffe relativement court, inversement, si la chaîne peut se retrouver piégée dans certaines petites zones de l'espace d'état pour un temps trop long, la période de chauffe sera d'autant plus longue.

Les bonnes ou mauvaises propriétés de mélange de la chaîne déterminent aussi la taille totale N de la chaîne à générer afin d'obtenir une estimation fiable de la densité f ciblée. Plus la chaîne se déplace vite, et donc se "mélange" bien, plus la convergence sera rapide. Plusieurs méthodes existent permettant de déterminer de manière théorique les ordres de grandeur pour m et N. On pourra par exemple se référer à Cowles et Carlin [1996] pour un passage en revue des diverses solutions existantes. On remarquera cependant qu'on ne peut connaître a priori le nombre d'itérations nécessaires à la convergence.

Stephens [2000a] génére des chaînes de taille $N = 20000$ et élimine les $m = 10000$ premières itérations. En effet, il a comparé les chaînes obtenues avec diverses initialisations pour avoir une idée du temps de chauffe, et considère alors que $m = 10000$ est suffisant. Richardson et Green [1997] avaient quant à eux généré $N = 200000$ valeurs pour en éliminer $m = 100000$.

De manière similaire à Stephens [1997], nous chercherons généralement un compromis entre le temps de calcul (beaucoup plus long en multivarié) et une taille de chaîne raisonnable de l'ordre de 20000 ou 30000. On pourra noter que l'algorithme de Gibbs pour des cas relativement simples comme ici, converge plus rapidement et ne nécessite que 10000 à 15000 itérations.

4 Echantillonnage de Gibbs pour les mélanges Gaussiens univariés.

Dans le cadre de cette thèse, nous avons choisi de reprendre la modélisation des mélanges gaussiens utilisée par Richardson et Green [1997] et Stephens [1997]. Bien d'autres modèles hiérarchiques sont utilisables, mais dans un souci d'efficacité, nous avons préféré ne présenter que celui utilisé par la suite.

Nous explicitons l'algorithme de Gibbs en donnant les lois a posteriori et en l'illustrant sur les données réelles étudiées au premier chapitre. On suppose que les données étudiées sont issues d'observations indépendantes d'un mélange gaussien univarié a k composantes de densité :

$$p\left(y \mid \pi, \mu, \sigma^2\right) = \pi_1 \mathcal{N}\left(y; \mu_1, \sigma_1^2\right) + ... + \pi_k \mathcal{N}\left(y; \mu_k, \sigma_k^2\right) \tag{II.4}$$

avec

$$\pi = (\pi_1, ..., \pi_k) \qquad \mu = (\mu_1, ..., \mu_k) \qquad \sigma^2 = \left(\sigma_1^2, ..., \sigma_k^2\right).$$

On utilise le modèle à données manquantes abordé au chapitre introductif. C'est-à-dire que l'on introduit la variable z_i pour tout $i = 1, ..., n$ ayant pour valeur l'indice du composant auquel appartient y_i. Rappelons avant de poursuivre que l'annexe A contient quelques précisions concernant les lois de probabilité utilisées dans la suite.

Le modèle hiérarchique utilisé par Richardson et Green [1997] introduit un hyperparamètre β permettant de modérer l'influence des paramètres fixés par l'utilisateur. Pour le mélange gaussien univarié à k composantes de densité (II.4), on définit les lois a priori de la manière suivante.

$$\begin{array}{rcl} p(\pi_1, ..., \pi_k \mid \delta) & \sim & \mathcal{D}(\delta, ..., \delta) : \text{ loi de Dirichlet de dimension } k \\ p(z_i \mid k, \pi) & \sim & \sum_{j=1}^{k} \pi_j \delta_j \text{ cad } \mathbb{P}(z_i = j) = \pi_{ij} \text{ pour } i = 1, ..., n \text{ et } j = 1, ..., k \\ p\left(\sigma_j^{-2} \mid \alpha, \beta\right) & \sim & \Gamma(\alpha, \beta) \quad \text{pour } j = 1, ..., k \text{ loi gamma} \\ p\left(\mu_j \mid \kappa, \xi\right) & \sim & \mathcal{N}\left(\xi, \kappa^{-1}\right) \quad \text{pour } j = 1, ..., k \\ p(\beta \mid g; h) & \sim & \Gamma(g, h). \end{array} \tag{II.5}$$

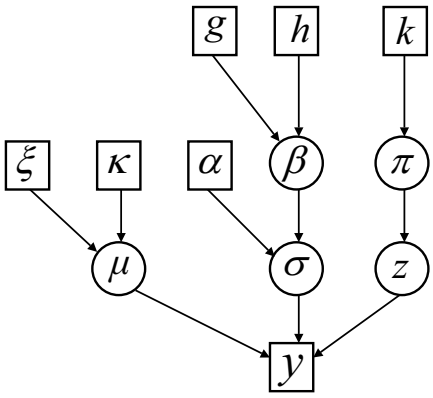


FIG. II.1 – Graphe acyclique ordonné pour le second modèle bayésien.

L'hyperparamètre δ de la loi de Dirichlet vaut généralement 1. Le graphe acyclique ordonné correspondant (Directed Acyclic Graph dans la littérature anglosaxonne) est présenté en figure II.1. Les hyperparamètres sont définis de la manière suivante :

ξ = milieu de l'intervalle de variation des données	R = longueur de cet intervalle de variation	$\kappa = \frac{1}{R^2}$
$\alpha = 2$	$g = 0.2$	$h = \frac{100g}{\alpha R^2}$

Ainsi définis, ξ et κ fournissent un a priori peu informatif sur les μ_j dans la mesure où la variance très grande (κ très petit) conduit à une loi de densité assez "plate". Le choix de β comme hyperparamètre permet de se situer dans un cadre peu informatif pour la taille des σ_j^2 en gardant toutefois une préférence pour des σ_j^2 similaires (selon Stephens [1997] page 22, c'est moins restrictif que de les contraindre à être égales). L'approche par des lois a priori peu informatives sera reprise dans le cas multivarié. Cette approche, choisie par Richardson et Green [1997], avait déjà été utilisée par Tierney [1996] dans le cadre des mélanges gaussiens.

Nous noterons la loi a posteriori (ou loi conditionnelles par rapport à toutes les autre variables), pour le paramètre β par exemple, de la manière suivante :

$$p\left(\beta \mid k, z, \pi, \mu, \sigma^2\right) = p(\beta \mid ...).$$

Les lois a posteriori conditionnelles permettant de mettre les paramètres à jour dans ce modèle sont :

$$
\begin{array}{rcl}
p\left(\pi_1,...,\pi_k \mid ...\right) & \sim & \mathcal{D}\left(\delta+n_1,...,\delta+n_k\right) : \text{ loi de Dirichlet de dimension } k \\
\mathbb{P}\left(z_i=j \mid ...\right) & \propto & \pi_j \mathcal{N}\left(y_i;\mu_j,\sigma_j\right) \text{ pour } i=1,...,n \text{ et } j=1,...,k \\
p\left(\beta \mid ...\right) & \sim & \Gamma\left(g+k\alpha, h+\sum_{j=1}^{k}\sigma_j^{-2}\right) \\
p\left(\mu_j \mid ...\right) & \sim & \mathcal{N}\left(\frac{\sigma_j^{-2} n_j \bar{y}_j + \kappa\xi}{\sigma_j^{-2} n_j + \kappa}, \left(\sigma_j^{-2} n_j + \kappa\right)^{-1}\right) \quad \text{pour j=1,...,k} \\
p\left(\sigma_j^{-2} \mid ...\right) & \sim & \Gamma\left(\alpha+\frac{n_j}{2}, \beta+\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\mu_j\right)^2 \mathbb{1}_{[z_i=j]}\right) \quad \text{pour j=1,...,k}
\end{array}
$$

avec $n_j = \sum_{i=1}^{n} \mathbb{1}_{[z_i=j]}$ $\quad n_j\bar{y}_j = \sum_{i=1}^{n} y_i \mathbb{1}_{[z_i=j]}$ $\quad S_j^2 = \sum_{i=1}^{n}\left(y_i-\bar{y}_j\right)^2 \mathbb{1}_{[z_i=j]}$

(cf Stephens [1997] page 23).

4.1 Echantillonnage de Gibbs

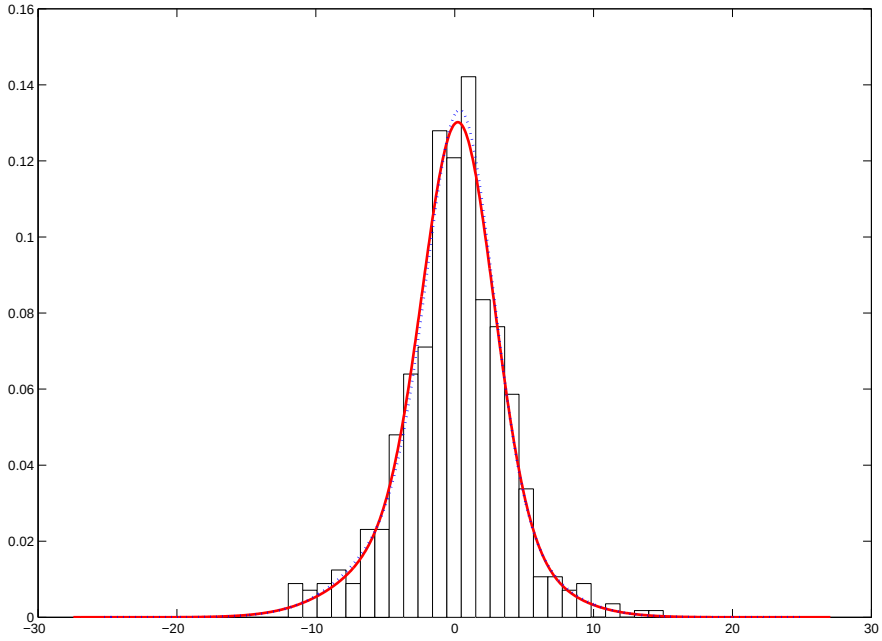


FIG. II.2 – Histogramme des données épidémiologiques étudiées au premier chapitre, et densité obtenue avec l'algorithme de Gibbs (traits pleins) et par l'algorithme EM (traits pointillés).

L'algorithme de Gibbs est complètement déterminé par les lois a posteriori conditionnelles. Dans le cas d'un mélange gaussien univarié à k composantes et pour le modèle bayesien présenté

plus haut, nous avons :

Algorithme de Gibbs pour les mélanges gaussiens à k composants

1- Simuler	z_i tq $\mathbb{P}(z_i = j \mid ...) \propto \pi_j \mathcal{N}\left(y_i; \mu_j, \sigma_j^2\right)$.
Simuler	$\beta \sim \Gamma\left(g + k\alpha, h + \sum_{j=1}^{k} \sigma_j^{-2}\right)$.
Calculer	n_j , $n_j \bar{y}_j$ et S_j^2.
2- Générer	$\mu_j \sim \mathcal{N}\left(\frac{\sigma_j^{-2} n_j \bar{y}_j + \kappa\xi}{\sigma_j^{-2} n_j + \kappa}, \left(\sigma_j^{-2} n_j + \kappa\right)^{-1}\right)$
	$\sigma_j^{-2} \sim \Gamma\left(\alpha + \frac{n_j}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \mu_j\right)^2 \mathbb{1}_{[z_i=j]}\right)$
	$\pi_1, ..., \pi_k \sim \mathcal{D}\left(\delta + n_1, ..., \delta + n_k\right)$

Les sorties de ce type d'algorithmes sont des suites de valeurs obtenues à chaque itération et pour chaque paramètre. L'évolution de ces valeurs au fil des itérations donne des indications sur les capacités d'exploration de l'algorithme. Nous avons choisi d'évaluer la distribution obtenue en l'approchant par son estimation par noyaux gaussiens. Nous avons ensuite tracé en pointillés la moyenne a posteriori, mais nous utiliserons plutôt le mode a posteriori dans toute la thèse. L'estimation finale du mélange est donc obtenue en réinjectant ces dernières valeurs dans l'expression de la densité du mélange gaussien à k composants. Certains auteurs préfèrent cependant utiliser une moyenne des distributions approximées à chaque itération.

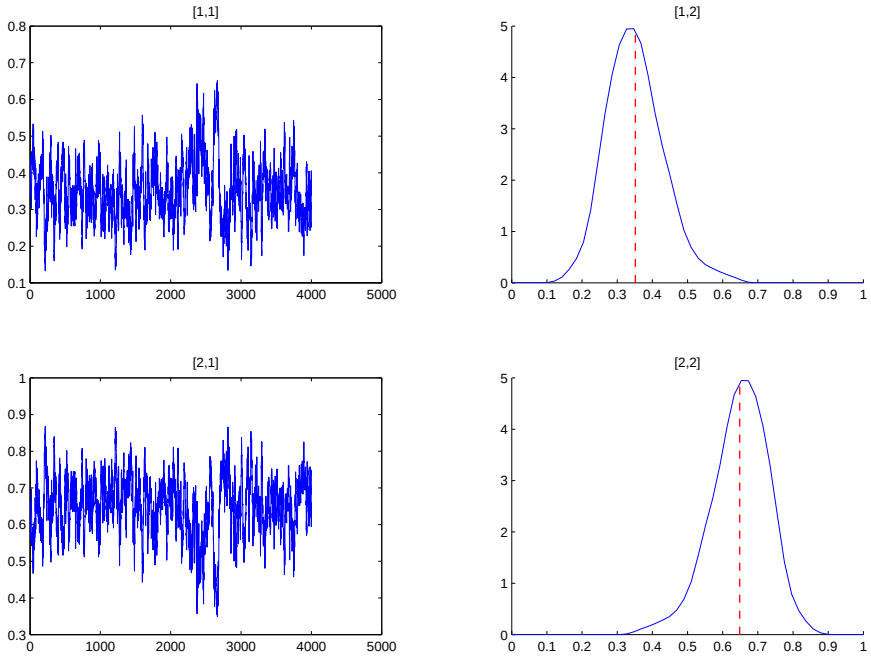


FIG. II.3 – [1,1] : évolution au cours du temps des valeurs produites par la chaîne MCMC pour la proportion du premier composant ; [1,2] : densité a posteriori de la proportion du premier composant estimée par la méthode des noyaux ; [2,1] et [2,2] : mêmes graphiques mais pour le second composant.

4.2 Illustrations et résultats

Afin d'illustrer les performances de l'algorithme de Gibbs dans le cas univarié, nous l'avons testé sur les données épidémiologiques étudiées au premier chapitre. Nous avons préalablement centré les observations pour des raisons pratiques et initialisé l'algorithme de Gibbs avec deux composants. 5000 itérations on été effectuées en éliminant les 1000 premières considérées comme le "temps de chauffe". Le graphique (II.2) et le tableau (II.1) présentent les densités de mélange obtenues par l'algorithme EM (initialisé par les k-means), et par l'algorithme de Gibbs en utilisant la procédure décrite plus haut. La partie gauche du graphique (II.3) montre l'évolution des proportions durant en fonction des itérations. La partie située à droite fournit quant à elle la densité a posteriori estimée par la méthode des noyaux. La figure (II.4) représente les mêmes éléments graphiques pour les moyennes et les variances. La première ligne de graphiques concerne la moyenne du premier composant, et la troisième sa variance. On peut en profiter pour constater que la moyenne et la variance du second composant (seconde et dernière ligne des graphiques) sont très stables alors que les variations sont plus importantes sur le premier composant. Ce phénomène est dû à l'influence de la variance dont la valeur est plus élevée pour le premier composant que pour le second.

Dans le cas présent, 6000 itérations (dont 1000 considérées comme "temps de chauffe") ont suffi à atteindre la convergence vers un des modes de la loi a posteriori. Le cas multivarié exploré à la section suivante montre déjà de plus grands signes d'instabilité.

GIBBS	$0.35\mathcal{N}(-0.69; 26.7) + 0.65\mathcal{N}(0.28; 6.26)$
EM	$0.42\mathcal{N}(-0.63; 24.6) + 0.58\mathcal{N}(0.42; 5.32)$

Tab. II.1 – Estimations d'un mélange à deux composants.

5 Échantillonnage de Gibbs pour les mélanges Gaussiens multivariés.

Cette section a pour objet de préciser le modèle hiérarchique adapté aux mélanges gaussiens multivariés. Nous suivrons ici l'approche utilisée par Stephens [2000a] qui généralise au cas multidimensionnel celle adoptée par Richardson et Green [1997]. Nous détaillons cette approche dans le cas multivarié, avec β pris comme paramètre et élément de la chaîne MCMC.

La densité d'un mélange gaussien multivarié à k composants est

$$p\left(y \mid \pi, \mu, \Sigma\right) = \pi_1 \mathcal{N}\left(y; \mu_1, \Sigma_1\right) + ... + \pi_k \mathcal{N}\left(y; \mu_k, \Sigma_k\right).$$

avec

$$\pi = (\pi_1, ..., \pi_k) \qquad \mu = (\mu_1, ..., \mu_k) \qquad \Sigma = (\Sigma_1, ..., \Sigma_k).$$

Dans le cas des mélanges de lois gaussiennes multivariées, la structure hiérarchique des lois a priori est identique. Les lois normales qui servent d'a priori pour les moyennes sont alors écrites en dimension r quelconque, et les lois gamma sont remplacées par des lois de Wishart qui en sont la généralisation. On pourra se reporter à l'Annexe A pour quelques précisions concernant ces lois.

Les lois a priori utilisées ayant été modifiées sont les suivantes :

$$\begin{array}{rcl} p\left(\beta \mid g, h\right) & \sim & W_r\left(2g, (2h)^{-1}\right) \\ p\left(\mu_j \mid k, \eta\right) & \sim & \mathcal{N}_r\left(\xi, \kappa^{-1}\right) \quad \text{pour } j = 1, ..., k \ \text{ loi Normale en dimension } r \\ p\left(\Sigma_j^{-1} \mid k, \eta\right) & \sim & W_r\left(2\alpha, (2\beta)^{-1}\right) \quad \text{pour } j = 1, ..., k \ \text{ loi de Wishart en dimension } r. \end{array}$$

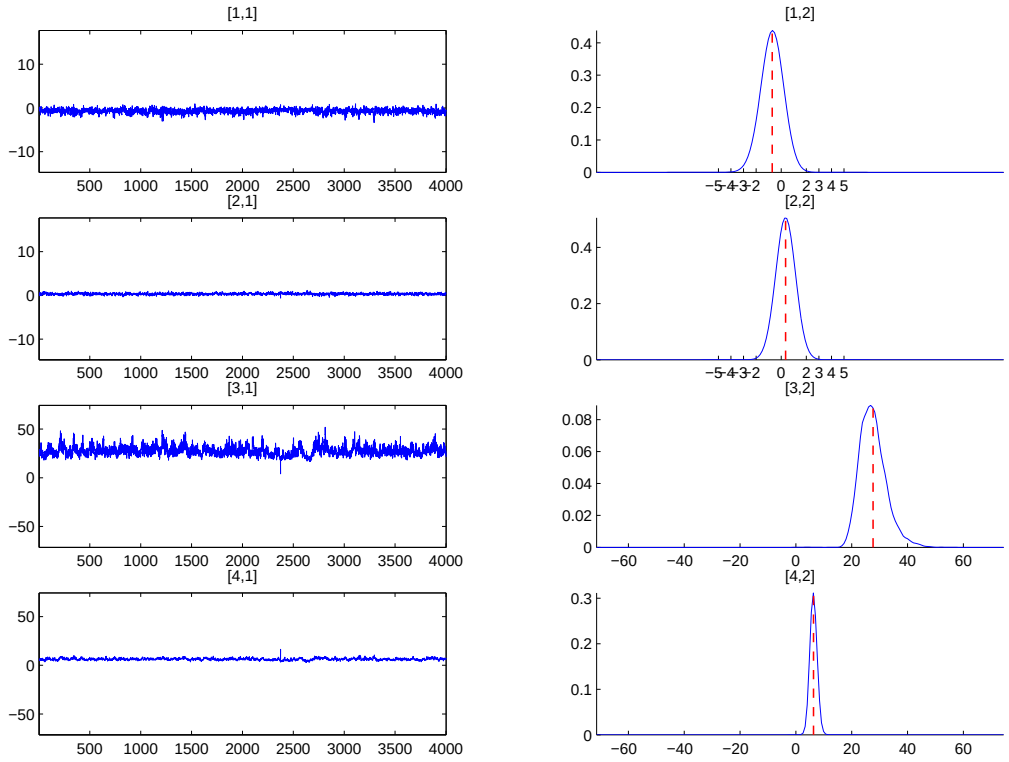


FIG. II.4 – Evolution et densité a posteriori pour les moyennes et les variances ; [1,1] et [1,2] : évolution et densité a posteriori estimée par la méthode des noyaux pour la première moyenne ; [2,1] et [2,2] : mêmes graphiques mais pour la seconde moyenne ; [3,1] et [3,2] : évolution et densité a posteriori estimée par la méthode des noyaux pour la première variance ; [4,1] et [4,2] : mêmes graphiques mais pour la seconde variance.

Les lois a posteriori, conditionnellement à toutes les autres variables de la chaîne sont les suivantes :

$$
\begin{aligned}
p\left(\beta \mid ...\right) &\sim W_r\left(2g + 2k\alpha, \left(2h + \textstyle\sum_{j=1}^{k} \Sigma_j^{-1}\right)^{-1}\right), \\
&\text{pour } j = 1, ..., k \ : \\
p\left(\mu_j \mid ...\right) &\sim \mathcal{N}_r\left(\left(\kappa + n_j\Sigma_j^{-1}\right)^{-1}\left(\kappa\xi + \Sigma_j^{-1}\textstyle\sum_{i \text{ tq } z_i=j}^{n} y_i\right), \left(\kappa + n_j\Sigma_j^{-1}\right)^{-1}\right), \\
&\text{pour } j = 1, ..., k : \\
p\left(\Sigma_j^{-1} \mid ...\right) &\sim W_r\left(2\alpha + n_j, \left(2\beta + \textstyle\sum_{i \text{ tq } z_i=j}^{n}\left(y_i - \mu_j\right)\left(y_i - \mu_j\right)'\right)^{-1}\right), \\
&\text{avec } n_j = \#\left\{i : z_i = j\right\}.
\end{aligned}
$$

Le choix des hyperparamètres provient là aussi de Stephens [1997]. Pour le cas où $r = 3$, nous les avons fixées de la manière suivante :

$$
\xi = (\xi_1, \xi_2, \xi_3) \quad \kappa = \begin{bmatrix} \frac{1}{R_1^2} & 0 & 0 \\ 0 & \frac{1}{R_2^2} & 0 \\ 0 & 0 & \frac{1}{R_3^2} \end{bmatrix} \quad h = \begin{bmatrix} \frac{100g}{\alpha R_1^2} & 0 & 0 \\ 0 & \frac{100g}{\alpha R_2^2} & 0 \\ 0 & 0 & \frac{100g}{\alpha R_3^2} \end{bmatrix}.
$$

$$
\alpha = 3 \qquad g = 0.3 \qquad \delta = 1
$$

On prend pour (ξ_1, ξ_2, ξ_3) les milieux des intervalles de variation des données sur chaque dimension, et R_1, R_2 et R_3 correspondant aux longueurs respectives de ces intervalles de variation. On suit là-aussi Stephens en prenant $\alpha = 3$ et $g = 0.3$; ceci représentant le fait que l'on introduit une contrainte sur les variances un peu plus grande que dans le cas univarié présenté dans Richardson et Green [1997], où α valait 2.

Ces hyperparamètres ainsi que le choix de la structure hiérarchique permettent de disposer d'un modèle où les a priori sont peu informatifs. En effet, le choix de ces valeurs est argumenté dans Richardson et Green [1997] par le fait qu'elles sont choisies afin de conduire à des a priori les plus "plats" possibles (et donc le moins informatifs possibles) tout en gardant la possibilité d'utiliser des lois conjuguées, bien pratiques pour les simulations.

On peut remarquer ici que la loi a posteriori proposée pour β est impropre lorsque $g = 0.3$. Ceci n'est cependant pas gênant pour l'analyse bayesienne que nous effectuons ici car la distribution a posteriori est bien une loi propre. En effet, comme l'écrit Stephens [1997] : "*We note that the prior on β is an improper distribution, but careful checking of the necessary integrals shows that the posterior distributions are proper*".

Remarque 5 ***:*** *Il sera utile de développer plus profondément les lois a priori sur les variances des composants en utilisant la paramétrisation suivante introduite par Banfield et Raftery [1993] et modifiée par Celeux et Govaert [1995] :*

$$\Sigma_i = \lambda_i D_i A_i D_i'$$

où $D_i = (d_{i1}, ..., d_{ir})$ est la matrice orthogonale des vecteurs propres de Σ_i. Si l'on ordonne les valeurs propres de Σ_i de la manière suivante :

$$\lambda_{i1} \geqslant \lambda_{i2} \geqslant ...\lambda_{ir} > 0,$$

la convention utilisée est alors de prendre $\lambda_i = \lambda_{i1}$ et $A_i = diag(1, \frac{\lambda_{i2}}{\lambda_{i1}}, ..., \frac{\lambda_{ir}}{\lambda_{i1}})$. Remarquons qu'il est aussi possible de considérer que $|A_i| = 1$, ce qui consiste à prendre $\lambda_i = |\Sigma_i|^{1/r}$ et $A_i = diag(\frac{\lambda_{i1}}{\lambda_i}, \frac{\lambda_{i2}}{\lambda_i}, ..., \frac{\lambda_{ir}}{\lambda_i})$. Le paramètre λ_i contrôle le volume du ième composant du mélange, A_i sa forme, et D_i son orientation. En introduisant un modèle hiérarchique sur tous ces paramètres afin d'estimer de manière dynamique la forme et l'orientation des composants, on peut espérer créer une chaîne de Markov dont la convergence serait plus facile à obtenir.

5.1 Illustrations et résultats

Pour les besoins des simulations, nous avons généré un grand nombre d'échantillons bivariés à 3 composants. Nous en avons retenu un en particulier que nous avons appelé "échantillon 1" et qui nous servira d'échantillon de référence dans toute la suite. Cet échantillon a été généré par Matlab avec les paramètres suivants :

$$0.4\mathcal{N}\left(\begin{bmatrix} -5 \\ -3 \end{bmatrix}; \begin{bmatrix} 7 & 1.5 \\ 1.5 & 2.8 \end{bmatrix}\right) + 0.5\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}; \begin{bmatrix} 0.2 & 0.6 \\ 0.6 & 7.3 \end{bmatrix}\right)$$
$$+0.1\mathcal{N}\left(\begin{bmatrix} 4 \\ 4 \end{bmatrix}; \begin{bmatrix} 0.1 & 0.08 \\ 0.08 & 0.2 \end{bmatrix}\right).$$

Sur les figures II.5 et II.6 on peut visualiser les points générés ainsi que l'estimation de la densité bivariée calculée par la méthode des noyaux. Nous utiliserons de plus quelques routines Matlab tirées de Martinez et Martinez [2002] nous permettant de représenter les mélanges. Nous utiliserons notamment ce qui est appelé le"csdfplot", permettant de représenter un mélange gaussien multivarié. Le "csdfplot" des données "échantillon 1" est visible dans la figure II.7. Ce graphique

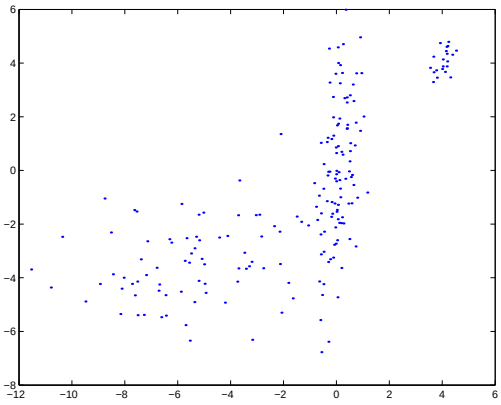


FIG. II.5 – Nuage de points des données générées, appelées "échantillon1".

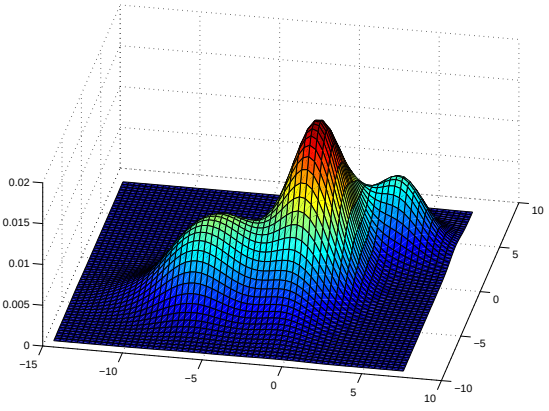


FIG. II.6 – Estimation par la méthode des noyaux de la densité des données de l'"échantillon1".

permet de visualiser de manière agréable et significative les paramètres d'un mélange multivarié dont le nombre de composants est important. En dimension deux, chaque composant est représenté par une ellipse centrée sur la moyenne de la densité associée, l'excentricité de l'ellipse représentant la structure de covariance associée. La proportion de chaque composant est affiché au centre de l'ellipse de covariance.

Les proportions des composants, lorsque ceux-ci sont nombreux, sont assez souvent mal estimées. Tracer directement la densité bivariée du mélange obtenu peut avoir des conséquences directes sur l'appréciation de la qualité d'une estimation. Le "csdfplot" permet de limiter ces erreurs d'appréciations, en donnant un "poids graphique" égal selon les paramètres et permet donc de mieux cerner sur quel paramètre l'estimation a échoué. Cela permet de détecter des algorithmes ayant convergé vers des jeux de paramètres donnant des mélanges apparemment très différents, alors qu'une seule des coordonnées d'une moyenne (par exemple) a été mal estimée. Ceci est particulièrement utile lorsque survient le phénomène de "label switching" que nous étudierons en détail un peu plus loin.

La figure II.9 concerne les densités a posteriori des proportions. La figure II.10 détaille les mêmes éléments pour chaque coordonnée des moyennes des composants, et la figure II.11 fait de même pour les variances. Nous utilisons comme estimateurs des paramètres le mode a posteriori, qui coïncide ici fortement avec la moyenne a posteriori (en pointillés). Remarquons aussi qu'une procédure de recuit simulé programmée pour maximiser la loi a posteriori globale sur tous les paramètres aurait permis d'obtenir directement une estimation proche du maximum a

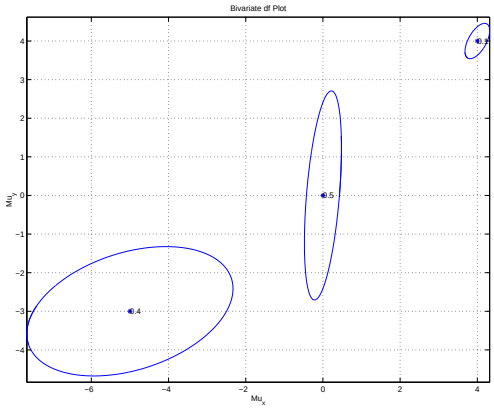


FIG. II.7 – Csdfplot du mélange original.

posteriori pour chaque paramètre. Une analyse selon chaque coordonnée nous permet cependant d'évaluer précisément des comportements comme la multimodalité, qui peut être caractéristique du phénomène de label switching.

On peut notamment constater que les 6000 itérations effectuées ici (dont 1000 de "temps de chauffe") ont permis d'atteindre la convergence vers un des modes de la loi a posteriori sans la présence du phénomène de label switching (chapitre IV). Le csdfplot de la figure II.8 permet de juger de la qualité de l'estimation obtenue. Le mélange apparaît comme étant très proche de celui d'origine.

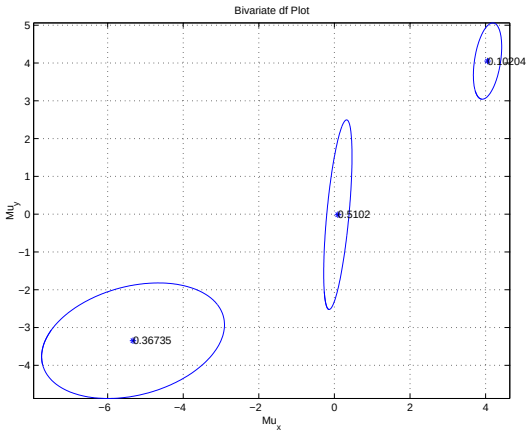


FIG. II.8 – Csdfplot du mélange estimé via l'algorithme de Gibbs.

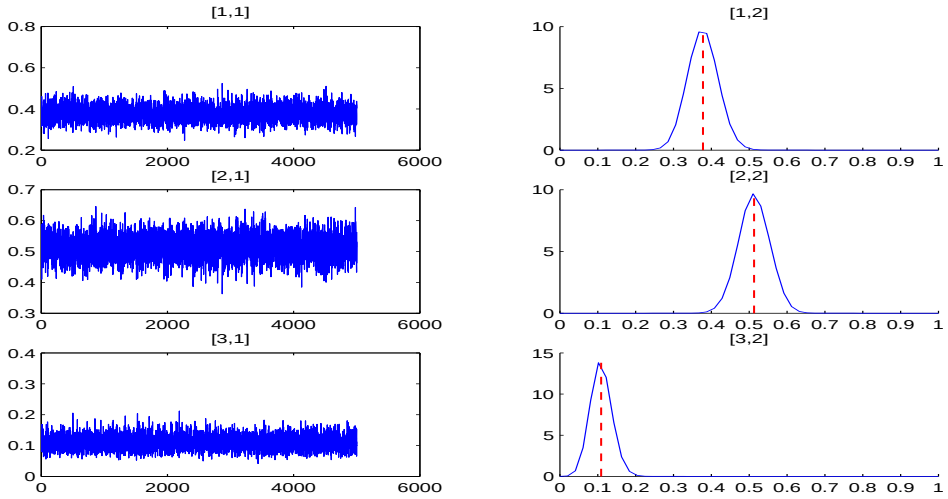


FIG. II.9 – Evolution des proportions pour l'algorithme de Gibbs appliqué aux mélanges multivariés. [1,1] : évolution de la proportion du premier composant ; [2,2] : densité a posteriori de la proportion du premier composant estimée par la méthode des noyaux ; [2,1] et [2,2] : mêmes graphiques mais pour le second composant ; [3,1] et [3,2] : mêmes graphiques mais pour le troisième composant.

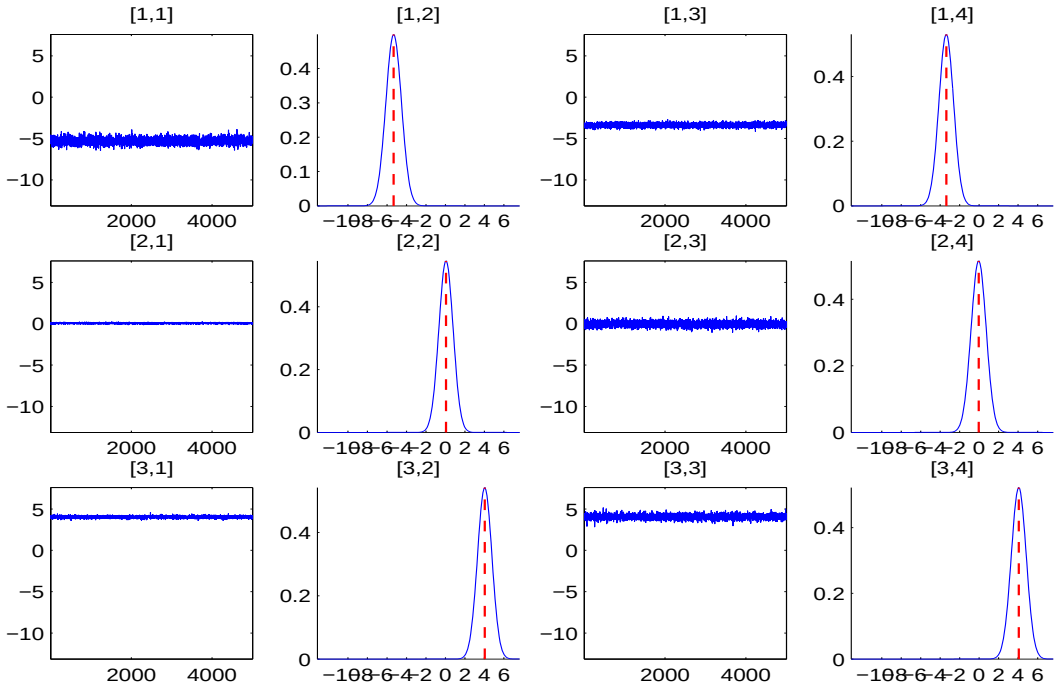


FIG. II.10 – Evolution et densité a posteriori pour les moyennes lors de l'algorithme de Gibbs appliqué aux mélanges multivariés. Chaque ligne de graphiques correspond aux deux coordonnées de la moyenne associée.

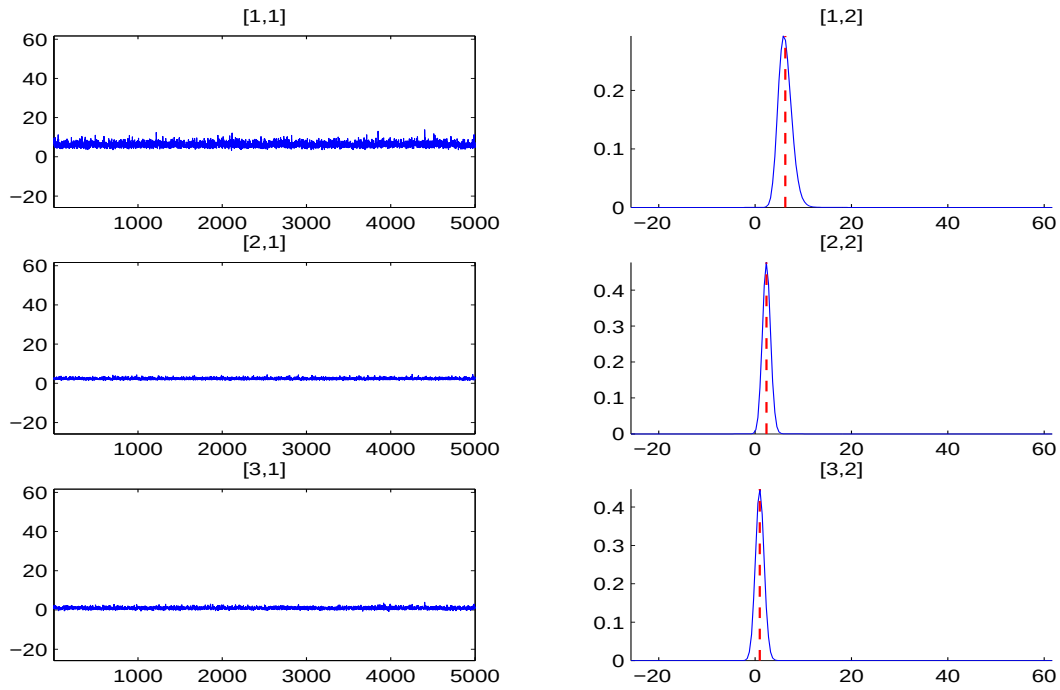


FIG. II.11 – Evolution et densité a posteriori pour les variances du second composant lors de l'algorithme de Gibbs appliqué aux mélanges multivariés. .

Chapitre III

Méthode MCMC à sauts réversibles

1 Introduction

Nous allons aborder dans ce chapitre les problèmes liés à la simulation d'une loi a posteriori sous un angle différent de celui des techniques MCMC. En effet, les algorithmes présentés au chapitre précédent permettent de simuler des distributions sur des espaces de dimension fixe, mais ils ne peuvent appréhender les situations où la dimension de l'espace est un des paramètres à simuler. Peter Green a développé dans Green [1995] une version hybride de l'Algorithme de Metropolis-Hastings, permettant de simuler une chaîne de Markov sur des espaces de dimension variable, cette méthode pouvant alors être appliquée à la problématique des "choix de modèles" très fréquente en Statistique. Pour l'exposé théorique de la méthode, nous nous sommes inspirés entre autre de Waagepetersen et Sorensen [2000].

Si l'on étudie un mélange de lois de densités f,

$$g\left(y \mid \pi, \theta\right) = \pi_1 f\left(y \mid \theta_1\right) + ... + \pi_k f\left(y \mid \theta_k\right),$$

on peut considérer qu'à chaque $k \in \{1, 2, ..., k_{\max}\} = \mathcal{K}$, correspond un modèle noté $\mathcal{M}_k$ de dimension différente. On considère qu'à chaque modèle $\mathcal{M}_k$ correspond un vecteur de paramètres $\theta^{(k)} = (\pi_1, ..., \pi_k, \theta_1, ..., \theta_k) \in \mathbb{R}^{n_k}$ $(n_k \in \mathbb{N})$. Dans une optique bayésienne, on utilisera le modèle naturel suivant :

$$p\left(k, \theta^{(k)}, y\right) = p\left(k\right) p\left(\theta^{(k)} \mid k\right) p\left(y \mid k, \theta^{(k)}\right), \tag{III.1}$$

où

- $p\left(k\right)$ est la loi a priori sur la "dimension" du modèle,
- $p\left(\theta^{(k)} \mid k\right)$ est la loi a priori sur le vecteur des paramètres conditionnellement à k,
- $p\left(y \mid k, \theta^{(k)}\right)$ est la densité des observations y (vraisemblance).

L'inférence sur les paramètres d'intérêt k et $\theta^{(k)}$ est alors effectuée en étudiant la loi a posteriori

$$p\left(k, \theta^{(k)} \mid y\right) \propto p\left(k\right) p\left(\theta^{(k)} \mid k\right) p\left(y \mid k, \theta^{(k)}\right). \tag{III.2}$$

En effet, on a

$$p\left(k, \theta^{(k)} \mid y\right) = \frac{p\left(k, \theta^{(k)}, y\right)}{\int\int p\left(k, \theta^{(k)}, y\right) dk d\theta^{(k)}},$$

quantité qu'il est plus commode de factoriser sous la forme

$$p\left(k, \theta^{(k)} \mid y\right) = p\left(k \mid y\right) p\left(\theta^{(k)} \mid k, y\right). \tag{III.3}$$

Cette loi est portée par l'ensemble $\mathcal{C} = \cup_{k\in\mathcal{K}} \{k\} \times \mathbb{R}^{n_k} = \cup_{k\in\mathcal{K}} \mathcal{C}_k$. Comme on l'a vu, pour les mélanges gaussiens on s'intéresse beaucoup au choix du nombre de composants et donc à $p\left(k \mid \theta^{(k)}, y\right)$. Cependant, il semble plus judicieux d'étudier $p\left(k, \theta^{(k)} \mid y\right)$ qui nous permet d'estimer en même temps les paramètres du mélange. Le but de la méthode est donc de simuler la loi a posteriori $p\left(k, \theta^{(k)} \mid y\right)$ en utilisant une chaîne de Markov se déplaçant sur $\mathcal{C} = \cup_{k\in\mathcal{K}} \mathcal{C}_k$.

On définit $X = (K, \Theta)$ les deux variables permettant de générer $\left(k, \theta^{(k)}\right)$, considéré comme l'élément courant de la chaîne de Markov. La loi a posteriori de (K, Θ), notée $\mathbb{P}(K = k, \Theta \in A_k \mid y)$ avec $A_k \in \mathbb{R}^{n_k}$, s'écrit de manière similaire à (III.3), c'est-à-dire :

$$\begin{aligned} \mathbb{P}(K = k \mid y)\, \mathbb{P}(\Theta \in A_k \mid K = k, y) &= p_k \int_{A_k} f(x \mid K = k, y)\, dx \\ &= p_k \int_{A_k} f_k(x)\, dx. \end{aligned}$$

On considère en effet que la mesure ciblée sur $\{k\} \times \mathbb{R}^{n_k}$ admet la densité $p_k f_k(x)$ par rapport à la mesure de Lebesgue.

Nous allons maintenant formaliser la méthode pour le cas général.

2 La méthode à sauts réversibles dans le cas général

On désire générer une chaîne de Markov réversible sur l'espace $\mathcal{C} = \cup_{k\in\mathcal{K}} \{k\} \times \mathbb{R}^{n_k} = \cup_{k\in\mathcal{K}} \mathcal{C}_k$. On note l'état courant de la chaîne $X_n = (K_n, \Theta_n)$, et $\left(k, \theta^{(k)}\right)$ une de ses réalisations. On a donc $\left(k, \theta^{(k)}\right) \in \{k\} \times A_k$ avec $A_k \subset \mathbb{R}^{n_k}$, et on se propose de formaliser le mécanisme permettant de générer une proposition pour l'état X_{n+1}, $\left(k', \theta^{(k')}\right) \in \{k'\} \times B_{k'}$, avec $B_{k'} \subset \mathbb{R}^{n_{k'}}$. On s'intéresse surtout à l'expression adéquate de la probabilité d'acceptation de ce nouvel état.

Le cadre est donc celui des transitions entre les espaces $\mathcal{C}_k$ et $\mathcal{C}_{k'}$ sans considérations sur leurs dimensions respectives. C'est-à-dire que nous formaliserons la méthode en suivant l'approche de Waagepetersen et Sorensen [2000], qui consiste à compléter les deux espaces pour se ramener à un espace global de dimension toujours identique. Il est donc nécessaire de compléter les deux espaces $\mathbb{R}^{n_k}$ et $\mathbb{R}^{n_{k'}}$ pour se placer dans un espace de même dimension. Il existe alors deux entiers positifs tels que

$$n_k + n_{kk'} = n_{k'} + n_{k'k}.$$

Cette condition est appelée "dimension matching condition" par Peter Green. Elle est nécessaire dans toute la suite. On est alors en mesure de définir deux applications correspondant aux sauts de la chaîne de Markov, permettant de sauter d'un espace à un autre.

$$g_{kk'1} \left\{ \begin{array}{ccc} \mathbb{R}^{n_k + n_{kk'}} & \longrightarrow & \mathbb{R}^{n_{k'}} \\ \left(\theta^{(k)}, u\right) & \longmapsto & \theta^{(k')} \end{array} \right.$$

et

$$g_{k'k1} \left\{ \begin{array}{ccc} \mathbb{R}^{n_{k'} + n_{k'k}} & \longrightarrow & \mathbb{R}^{n_k} \\ \left(\theta^{(k')}, u'\right) & \longmapsto & \theta^{(k)} \end{array} \right. .$$

On suppose de plus qu'il y a injectivité de ces applications par rapport à la deuxième composante, c'est-à-dire :

$$g_{i1}(x, \alpha) = g_{i1}(x, \beta) \Longrightarrow \alpha = \beta \quad \text{pour } i = kk', k'k.$$

On suppose aussi qu'il existe une condition d'inversion qui nous permet de revenir en arrière, c'est-à-dire que l'on a :

$$\forall \theta^{(k)} \in \mathbb{R}^{n_k}, \forall u \in \mathbb{R}^{n_{kk'}}, \exists u' \in \mathbb{R}^{n_{k'k}} \quad \text{tel que} \quad g_{k'k1}\left(g_{kk'1}\left(\theta^{(k)}, u\right), u'\right) = \theta^{(k)}.$$

Cette condition d'inversion nous permet alors de définir une nouvelle application permettant de déterminer u' de manière unique connaissant $\left(\theta^{(k)}, u\right)$. On obtient alors la seconde application :

$$g_{kk'2} \left\{ \begin{array}{c} \mathbb{R}^{n_k+n_{kk'}} \longrightarrow \mathbb{R}^{n_{k'k}} \\ \left(\theta^{(k)}, u\right) \longmapsto u' \end{array} \right. .$$

De façon symétrique, on a

$$\forall \theta^{(k')} \in \mathbb{R}^{n_{k'}}, \forall u' \in \mathbb{R}^{n_{k'k}}, \exists u \in \mathbb{R}^{n_{kk'}} \quad \text{tel que} \quad g_{kk'1}\left(g_{k'k1}\left(\theta^{(k')}, u'\right), u\right) = \theta^{(k')},$$

qui nous permet de définir

$$g_{k'k2} \left\{ \begin{array}{c} \mathbb{R}^{n_{k'}+n_{k'k}} \longrightarrow \mathbb{R}^{n_{kk'}} \\ \left(\theta^{(k')}, u'\right) \longmapsto u \end{array} \right. .$$

On obtient finalement deux applications inverses l'une de l'autre :

– Transition du modèle $\mathcal{M}_k$ au modèle $\mathcal{M}_{k'}$

$$g_{kk'} \left\{ \begin{array}{l} g_{kk'1} \left\{ \begin{array}{c} \mathbb{R}^{n_k+n_{kk'}} \longrightarrow \mathbb{R}^{n_{k'}} \\ \left(\theta^{(k)}, u\right) \longmapsto \theta^{(k')} \end{array} \right. \\ g_{kk'2} \left\{ \begin{array}{c} \mathbb{R}^{n_k+n_{kk'}} \longrightarrow \mathbb{R}^{n_{kk'}} \\ \left(\theta^{(k)}, u\right) \longmapsto u' \end{array} \right. \end{array} \right. .$$

– Transition du modèle $\mathcal{M}_{k'}$ au modèle $\mathcal{M}_k$

$$g_{k'k} \left\{ \begin{array}{l} g_{k'k1} \left\{ \begin{array}{c} \mathbb{R}^{n_{k'}+n_{k'k}} \longrightarrow \mathbb{R}^{n_k} \\ \left(\theta^{(k')}, u'\right) \longmapsto \theta^{(k)} \end{array} \right. \\ g_{k'k2} \left\{ \begin{array}{c} \mathbb{R}^{n_{k'}+n_{k'k}} \longrightarrow \mathbb{R}^{n_{kk'}} \\ \left(\theta^{(k')}, u'\right) \longmapsto u \end{array} \right. \end{array} \right. .$$

Avec $g_{kk'} = g_{k'k}^{-1}$. Toutes ces définitions sont résumées par la figure III.1.

Afin de bien définir la chaîne de Markov, il nous faut définir $P(A, B)$, la probabilité de se déplacer dans un ensemble $B \subset \mathcal{C}$ en provenant d'un ensemble $A \subset \mathcal{C}$. On considère $A = \{k\} \times A_k$ et $B = \{k'\} \times B_{k'}$. On définit $P(A, B)$ à l'aide du noyau de transition de la chaîne : $P(x, B)$ avec $X_n = x = (k, \theta^{(k)}) \in A$. On écrit :

$$\begin{aligned} P(x, B) &= \mathbb{P}\left[X_{n+1} \subset B \mid X_n = x\right] \\ &= \mathbb{P}\left[K_{n+1} = k', \Theta_{n+1} = \theta^{(k')} \mid K_n = k, \Theta_n = \theta^{(k)}\right]. \end{aligned} \tag{III.4}$$

On obtient alors :

$$\begin{aligned} P(A, B) &= \mathbb{P}\left[K_{n+1} = k', \Theta_{n+1} \subset B_{k'} \mid K_n = k, \Theta_n \subset A_k\right] \\ &= p_k \int_{A_k} f_k\left(\theta^{(k)}\right) \mathbb{P}\left[K_{n+1} = k', \Theta_{n+1} \subset B_{k'} \mid K_n = k, \Theta_n = \theta^{(k)}\right] d\theta^{(k)} \end{aligned}$$

où

$$\begin{aligned} p_k &= \mathbb{P}\left[K = k \mid y\right], \\ f_k\left(\theta^{(k)}\right) &= f\left(\theta^{(k)} \mid K = k, y\right). \end{aligned} \tag{III.5}$$

Avant de détailler l'expression de (III.4), il faut tout d'abord définir certains termes.

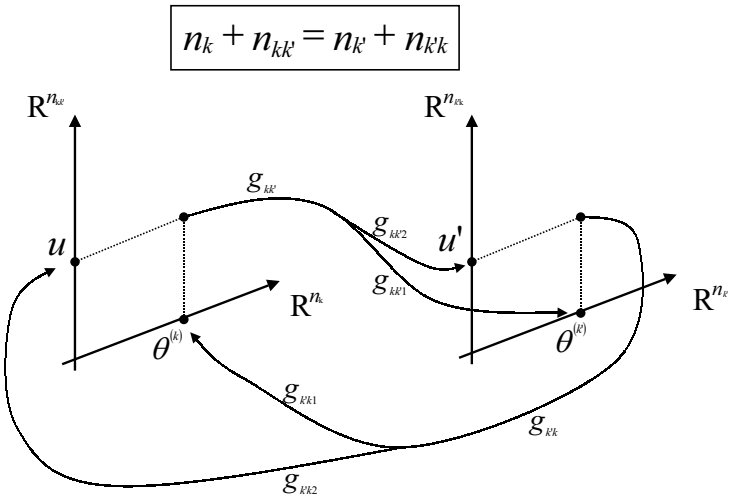


FIG. III.1 – Applications permettant de définir les transitions entre $\mathbb{R}^{n_k+n_{kk'}}$ et $\mathbb{R}^{n_{k'}+n_{k'k}}$.

- On note $p_{kk'}$, la probabilité a priori de passer d'un modèle de dimension k à un modèle de dimension k'.
- On note $a_{kk'}\left(\theta^{(k)}, g_{kk'1}\left(\theta^{(k)}, u\right)\right)$ (ou bien $a_{kk'}\left(\theta^{(k)}, \theta^{(k')}\right)$), la probabilité d'acceptation du nouvel état proposé $\theta^{(k')} = g_{kk'1}\left(\theta^{(k)}, u\right)$, resp $a_{k'k}\left(\theta^{(k')}, \theta^{(k)}\right)$.
- On génère $u \in \mathbb{R}^{n_{kk'}}$ selon la loi de densité $q_{kk'}\left(\theta^{(k)}, u\right)$, (resp $q_{k'k}\left(\theta^{(k')}, u'\right)$).

Dans le cadre général qui nous intéresse, (III.4) s'écrit comme la somme de deux termes : la probabilité de proposer $(k', \theta^{(k')})$ comme nouvel état et de l'accepter, et la probabilité d'être déjà en $(k', \theta^{(k')})$ et de refuser toute nouvelle proposition. On écrira sous forme mathématique :

$$\mathbb{P}\left[K_{n+1} = k', \Theta_{n+1} = \theta^{(k')} \mid K_n = k, \Theta_n \subset B_{k'}\right] = Q_{kk'}\left(\theta^{(k)}, B_{k'}\right) + s_k\left(\theta^{(k)}\right) \mathbb{1}_{\left[k=k', \theta^{(k)} \in B_{k'}\right]},$$

avec

$$Q_{kk'}\left(\theta^{(k)}, B_{k'}\right) = p_{kk'} \int \mathbb{1}_{\left[g_{kk'1}\left(\theta^{(k)}, u\right) \in B_{k'}\right]} a_{kk'}\left(\theta^{(k)}, g_{kk'1}\left(\theta^{(k)}, u\right)\right) q_{kk'}\left(\theta^{(k)}, u\right) du,$$

et

$$\begin{aligned} s_k\left(\theta^{(k)}\right) &= \mathbb{P}\left[\text{la proposition est rejetée} \mid X_n = (k, \theta^{(k)})\right] \\ &= \sum_{k'=1}^{k_{\max}} p_{kk'} \int q_{kk'}\left(\theta^{(k)}, u\right)\left[1 - a_{kk'}\left(\theta^{(k)}, g_{kk'1}\left(\theta^{(k)}, u\right)\right)\right] q_{kk'}\left(\theta^{(k)}, u\right) du. \end{aligned}$$

On rappelle que l'équation de balance représentant la condition d'inversibilité de la chaîne s'écrit :

$$P(A, B) = \int_A P(x, B)\, \pi(x)\, dx = \int_B P(x, A)\, \pi(x)\, dx = P(B, A), \tag{III.6}$$

où π est la mesure ciblée par la chaîne de Markov, $P(x, B)$ son noyau de transition, et A et B deux ensembles inclus dans le support de la loi π. Dans le cas présent, lorsque :

$$\begin{aligned} (Bal1) &= \mathbb{P}\left[K_{n+1} = k', \Theta_{n+1} \subset B_{k'} \mid K_n = k, \Theta_n \subset A_k\right], \\ (Bal2) &= \mathbb{P}\left[K_{n+1} = k, \Theta_{n+1} \subset A_k \mid K_n = k', \Theta_n \subset B_{k'}\right], \end{aligned}$$

la condition de balance s'écrit alors :

$$(Bal1) = (Bal2)\,.$$

On doit déterminer la probabilité d'acceptation $a_{kk'}\left(\theta^{(k)}, g_{kk'1}\left(\theta^{(k)}, u\right)\right)$ pour avoir l'égalité entre $(Bal1)$ et $(Bal2)$. Ecrivons explicitement $(Bal1)$:

$$\begin{aligned}
& (Bal1) \\
= \; & p_k \int_{A_k} f_k\left(\theta^{(k)}\right) \mathbb{P}\left[K_{n+1} = k', \Theta_{n+1} = \theta^{(k')} \mid K_n = k, \Theta_n = \theta^{(k)}\right] d\theta^{(k)} \\
= \; & p_k \int_{A_k} f_k\left(\theta^{(k)}\right) Q_{kk'}\left(\theta^{(k)}, B_{k'}\right) d\theta^{(k)} \\
& + p_k \int_{A_k} f_k\left(\theta^{(k)}\right) s_k\left(\theta^{(k)}\right) \mathbb{1}_{\left[k=k', \theta^{(k)} \in B_{k'}\right]} d\theta^{(k)} \\
= \; & p_k \int_{A_k} f_k\left(\theta^{(k)}\right) Q_{kk'}\left(\theta^{(k)}, B_{k'}\right) d\theta^{(k)} \\
& + p_k \int f_k\left(\theta^{(k)}\right) s_k\left(\theta^{(k)}\right) \mathbb{1}_{\left[k=k', \theta^{(k)} \in B_{k'} \cap A_k\right]} d\theta^{(k)}.
\end{aligned}$$

Ecrivons $(Bal2)$:

$$\begin{aligned}
(Bal2) \; = \; & p_{k'} \int_{B_{k'}} f_{k'}\left(\theta^{(k')}\right) \\
& \times \mathbb{P}\left[K_{n+1} = k, \Theta_{n+1} = \theta^{(k)} \mid K_n = k', \Theta_n = \theta^{(k')}\right] d\theta^{(k')} \\
= \; & p_{k'} \int_{B_{k'}} f_{k'}\left(\theta^{(k')}\right) Q_{k'k}\left(\theta^{(k')}, A_k\right) d\theta^{(k')} \\
& + p_{k'} \int f_{k'}\left(\theta^{(k')}\right) s_k\left(\theta^{(k')}\right) \mathbb{1}_{\left[k=k', \theta^{(k')} \in B_{k'} \cap A_k\right]} d\theta^{(k')}.
\end{aligned}$$

Les deux seconds membres sont identiques. La condition de balance est donc vérifiée si

$$(Bal1') = (Bal2')\,,$$

où :

$$\begin{aligned}
(Bal1') \; &= \; p_k \int_{A_k} f_k\left(\theta^{(k)}\right) Q_{kk'}\left(\theta^{(k)}, B_{k'}\right) d\theta^{(k)}, \\
(Bal2') \; &= \; p_{k'} \int_{B_{k'}} f_{k'}\left(\theta^{(k')}\right) Q_{k'k}\left(\theta^{(k')}, A_k\right) d\theta^{(k')}.
\end{aligned}$$

On a :

$$\begin{aligned}
& (Bal1') \\
= \; & p_k \int_{A_k} f_k\left(\theta^{(k)}\right) p_{kk'} \int \mathbb{1}_{\left[g_{kk'1}\left(\theta^{(k)}, u\right) \in B_{k'}\right]} a_{kk'}\left(\theta^{(k)}, g_{kk'1}\left(\theta^{(k)}, u\right)\right) q_{kk'}\left(\theta^{(k)}, u\right) du d\theta^{(k)} \\
= \; & \int \int \mathbb{1}_{\left[\theta^{(k)} \in A_k\right]} \mathbb{1}_{\left[g_{kk'1}\left(\theta^{(k)}, u\right) \in B_{k'}\right]} p_k f_k\left(\theta^{(k)}\right) p_{kk'} a_{kk'}\left(\theta^{(k)}, g_{kk'1}\left(\theta^{(k)}, u\right)\right) q_{kk'}\left(\theta^{(k)}, u\right) du d\theta^{(k)}.
\end{aligned}$$

De même :

$$\begin{aligned}
& (Bal2') \\
= \; & p_{k'} \int_{B_{k'}} f_{k'}\left(\theta^{(k')}\right) Q_{k'k}\left(\theta^{(k')}, A_k\right) d\theta^{(k')} \\
= \; & p_{k'} \int_{B_{k'}} f_{k'}\left(\theta^{(k')}\right) p_{k'k} \\
& \times \int \mathbb{1}_{\left[g_{k'k1}\left(\theta^{(k')},u'\right)\in A_k\right]} a_{k'k}\left(\theta^{(k')}, g_{k'k1}\left(\theta^{(k')},u'\right)\right) q_{k'k}\left(\theta^{(k')},u'\right) du' d\theta^{(k')} \\
= \; & \int\int \mathbb{1}_{\left[\theta^{(k')}\in B_{k'}\right]} \mathbb{1}_{\left[g_{k'k1}\left(\theta^{(k')},u'\right)\in A_k\right]} \\
& \times p_{k'} f_{k'}\left(\theta^{(k')}\right) p_{k'k} a_{k'k}\left(\theta^{(k')}, g_{k'k1}\left(\theta^{(k')},u'\right)\right) q_{k'k}\left(\theta^{(k')},u'\right) du' d\theta^{(k')}.
\end{aligned}$$

On doit alors faire le changement de variables suivant :

$$\begin{aligned}
\left(\theta^{(k')},u'\right) & \longmapsto & \left(\theta^{(k)},u\right) &= \left(g_{k'k1}\left(\theta^{(k')},u'\right), g_{k'k2}\left(\theta^{(k')},u'\right)\right) \\
du' d\theta^{(k')} & = & & \left|\frac{\partial g_{kk'}}{\partial\theta^{(k)}\partial u}\right| du d\theta^{(k)}.
\end{aligned}$$

Le Jacobien peut aussi s'écrire de la manière suivante :

$$\begin{vmatrix} \frac{\partial g_{kk'1}\left(\theta^{(k)},u\right)}{\partial\theta^{(k)}} & \frac{\partial g_{kk'2}\left(\theta^{(k)},u\right)}{\partial\theta^{(k)}} \\ \frac{\partial g_{kk'1}\left(\theta^{(k)},u\right)}{\partial u} & \frac{\partial g_{kk'2}\left(\theta^{(k)},u\right)}{\partial u} \end{vmatrix}.$$

Finalement :

$$\begin{aligned}
& (Bal2') = \\
& \int\int \mathbb{1}_{\left[\theta^{(k)}\in A_k\right]} \mathbb{1}_{\left[g_{kk'1}\left(\theta^{(k)},u\right)\in B_{k'}\right]} p_{k'} f_{k'}\left(g_{kk'1}\left(\theta^{(k)},u\right)\right) p_{k'k} a_{k'k}\left(g_{kk'1}\left(\theta^{(k)},u\right),\theta^{(k)}\right) \\
& \times q_{k'k}\left(g_{kk'1}\left(\theta^{(k)},u\right), g_{kk'2}\left(\theta^{(k)},u\right)\right) \left|\frac{\partial g_{kk'}}{\partial\theta^{(k)}\partial u}\right| du d\theta^{(k)}.
\end{aligned}$$

L'égalité entre $(Bal1')$ et $(Bal2')$ est donc vérifiée avec

$$\begin{aligned}
& p_k f_k\left(\theta^{(k)}\right) p_{kk'} a_{kk'}\left(\theta^{(k)}, g_{kk'1}\left(\theta^{(k)},u\right)\right) q_{kk'}\left(\theta^{(k)},u\right) \\
= \; & p_{k'} f_{k'}\left(g_{kk'1}\left(\theta^{(k)},u\right)\right) p_{k'k} a_{k'k}\left(g_{kk'1}\left(\theta^{(k)},u\right),\theta^{(k)}\right) \\
& \times q_{k'k}\left(g_{kk'1}\left(\theta^{(k)},u\right), g_{kk'2}\left(\theta^{(k)},u\right)\right) \left|\frac{\partial g_{kk'}}{\partial\theta^{(k)}\partial u}\right|.
\end{aligned}$$

En suivant Richardson et Green [1997] et en considérant que l'on doit choisir $a_{kk'}\left(\theta^{(k)},\theta^{(k')}\right)$ aussi grand que possible, on a immédiatement :

$$\boxed{a_{kk'}\left(\theta^{(k)},\theta^{(k')}\right) = \min\left[1, \frac{p_{k'} f_{k'}\left(\theta^{(k')}\right) p_{k'k} q_{k'k}\left(\theta^{(k')},u'\right)}{p_k f_k\left(\theta^{(k)}\right) p_{kk'} q_{kk'}\left(\theta^{(k)},u\right)} \left|\frac{\partial g_{kk'}}{\partial\theta^{(k)}\partial u}\right|\right].}$$

Le jacobien intervenant dans la probabilité d'acceptation d'un nouvel élément rend les calculs nécessaires souvent très complexes. Dans certaines applications, le jacobien est cependant simple à calculer, mais dans le cas des mélanges gaussiens multivariés, il constitue un éceuil majeur.

3 Un cas particulier

Dans certains cas, il est fort utile de proposer le nouvel état de manière déterministe (c'est-à-dire sans générer un u de manière aléatoire), tout en gardant l'aspect aléatoire pour le mouvement inverse. Ceci survient en particulier lorsqu'on désire passer d'un modèle $\mathcal{M}_{k'}$ à un modèle $\mathcal{M}_k$ de dimension inférieure ($k < k'$). Le mouvement est alors effectué de manière déterministe et destructive en générant directement le nouvel état $\theta^{(k)}$ comme $\theta^{(k)} = g_{k'k1}\left(\theta^{(k')}\right)$.

La condition sur les dimensions peut alors s'exprimer par

$$n_k + n_{kk'} = n_{k'}.$$

On définit à peu de choses près les mêmes applications que dans le cas général, qui sont résumées dans la figure III.2.

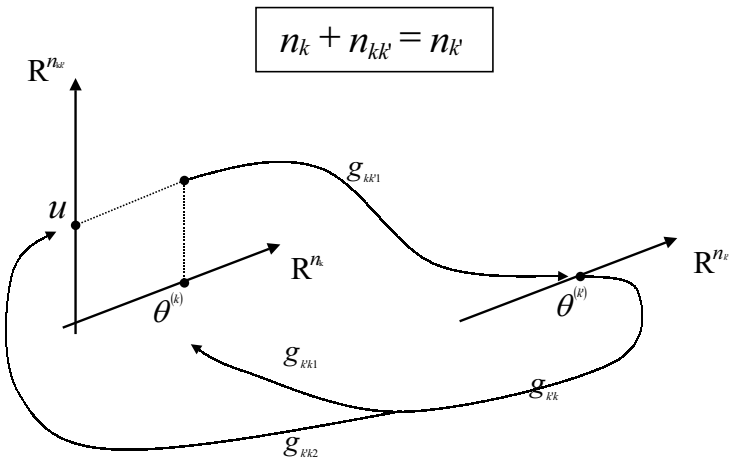


FIG. III.2 – fonctions permettant de définir les transitions entre $\mathbb{R}^{n_k+n_{kk'}}$ et $\mathbb{R}^{n_{k'}}$.

La principale différence est qu'alors on a

$$Q_{k'k}\left(\theta^{(k')}, A_k\right) = p_{k'k} \mathbb{1}_{\left[g_{k'k1}\left(\theta^{(k')},u'\right)\in A_k\right]} a_{k'k}\left(\theta^{(k')}, g_{k'k1}\left(\theta^{(k')}, u'\right)\right),$$

c'est-à-dire qu'on n'a plus besoin d'intégrer par rapport à u'. La probabilité d'acceptation s'écrit alors :

$$\boxed{a_{k'k}\left(\theta^{(k')}, \theta^{(k)}\right) = \min\left[1, \frac{p_k f_k\left(\theta^{(k)}\right) p_{kk'} q_{kk'}\left(\theta^{(k)}, u\right)}{p_{k'} f_{k'}\left(\theta^{(k')}\right) p_{k'k}} \left|\frac{\partial g_{k'k}\left(\theta^{(k')}\right)}{\partial \theta^{(k')}}\right|\right].}$$

Lorsqu'on effectue le mouvement inverse, c'est-à-dire lorsqu'on désire passer d'un modèle $\mathcal{M}_k$ à un modèle $\mathcal{M}_{k'}$ de dimension supérieure ($k < k'$), les fonctions à définir sont exactement les mêmes, et la probabilité d'acceptation est en fait l'inverse de la précédente, c'est-à-dire

$$\boxed{a_{kk'}\left(\theta^{(k)}, \theta^{(k')}\right) = \min\left[1, \frac{p_{k'} f_{k'}\left(\theta^{(k')}\right) p_{k'k}}{p_k f_k\left(\theta^{(k)}\right) p_{kk'} q_{kk'}\left(\theta^{(k)}, u\right)} \left|\frac{\partial g_{kk'}\left(\theta^{(k)}, u\right)}{\partial \theta^{(k)} \partial u}\right|\right].}$$

Précisons que dans l'application envisagée au chapitre V qui concerne les mélanges gaussiens multivariés, le passage d'un mélange à k composants à partir d'un mélange à $k-1$ composants, s'opère de cette façon.

Chapitre IV

Le problème du label switching

1 Introduction

Au chapitre introductif, nous avons abordé les problèmes posés par la symétrie de la loi a posteriori pour les modèles de mélanges. En effet, la non-identifiabilité de ces modèles induit la présence de $k!$ modes dans la vraisemblance, et donc aussi dans la loi a posteriori.

Le problème particulier du label switching, que l'on pourrait traduire par "changement d'indice", est lié à la multimodalité de la vraisemblance et intervient lorsqu'on simule une chaîne de Markov destinée à estimer les paramètres d'un mélange de lois.

En effet, lorsque les lois a priori utilisées sont symétriques, la multimodalité de la vraisemblance se transmet à la loi a posterori du fait de la relation de proportionnalité existant entre ces deux lois. Lorsqu'on génère une chaîne de Markov sans contraintes afin d'explorer la loi a posteriori, on peut observer deux comportements différents.

Dans certains cas, la chaîne reste dans le voisinage d'un unique mode de la loi a posteriori. Les lois a posteriori obtenues ont alors un aspect se prêtant agréablement à l'inférence puisqu'elles sont unimodales. Ce comportement peut sembler satisfaisant, mais on a alors la certitude que la chaîne n'a pas exploré tout l'espace des paramètres. Cela peut être gênant pour affirmer la validité de l'inférence.

Dans d'autres cas où la chaîne explore mieux l'espace d'état, celle ci se déplace d'un mode à l'autre. On est alors en présence de lois a posteriori multimodales et identiques pour tous les paramètres. Cela est interprété par certains auteurs (cf Celeux et al. [2000] par exemple) comme un signe de convergence de la chaîne vers la vraie distribution a posteriori. Cependant, les lois a posteriori obtenues sont difficilement utilisables pour l'inférence. En effet, les estimateurs classiques comme la moyenne a posteriori peuvent se situer dans des "vallées" séparant les différents modes de la loi a posteriori. Les difficultés posées par cette multimodalité portent le nom de problème du "label switching". Il y a là un paradoxe apparent, car le problème du "label switching" est justement le signe d'une bonne convergence et de bonnes propriétés d'exploration de l'algorithme.

La solution adoptée jusqu'ici, notamment par Richardson et Green [1997], consiste à utiliser une contrainte d'identifiabilité (sur les moyennes ou les variances par exemple). Cependant, l'imposition d'une contrainte d'identifiabilité tend à déformer l'évolution des paramètres pendant l'algorithme et contraint alors ceux-ci à rester dans certaines régions de l'espace d'état.

L'idée sous jacente à l'imposition d'une telle contrainte est de restreindre l'exploration de l'espace d'état à une région correspondant à un seul des modes de la loi a posteriori (ceux ci étant équivalents par permutation). Le problème posé par ce type de contraintes a notamment été illustré par Richardson et Green [1997] pour un problème d'estimation d'un mélange gaussien à deux composants. Les diverses contraintes d'identifiabilité imposées ne mènent pas aux mêmes lois a posteriori et il y a quand même la preuve d'un changement d'indice à travers la multimodalité.

En effet, comme il a été récemment souligné par Celeux et al. [2000], l'imposition de contraintes d'identifiabilité ne fonctionne pas très bien. La condition imposée ne respecte pas forcément la forme et la géométrie de la loi a posteriori. En voulant restreindre l'exploration à un seul mode, il arrive souvent que l'on explore plutôt des petites parties de régions correspondant en fait à plusieurs modes différents. Le choix d'une contrainte est donc très difficile à faire, et il n'existe pas de contraintes fonctionnant mieux que d'autres.

A l'inverse, ne pas imposer de contraintes laisse la chaîne totalement libre pour l'exploration, mais ne résout pas le problème de la multimodalité, c'est-à-dire le problème du "label switching". Ce constat incite à rechercher des moyens d'effectuer l'inférence à partir d'une chaîne non contrainte. Certains auteurs comme Celeux et al. [2000] et Stephens [2000b] préconisent un post-traitement des données obtenues afin de les réorganiser selon un critère prédéterminé.

C'est cette dernière approche que nous allons essayer de mettre en oeuvre en n'imposant pas de condition d'identifiabilité particulière. Nous allons étudier ici deux méthodes récemment proposées pour solutionner le phénomène de label switching sans imposer de contraintes particulières sur la chaîne simulée. Nous examinerons leur capacité à réordonner les chaînes de Markov obtenues afin d'obtenir des lois a posteriori unimodales et concentrées autour d'un mode unique.

Nous allons maintenant essayer d'illustrer le phénomène de "label switching". Supposons que l'on dispose d'un petit échantillon tiré d'un mélange gaussien à deux composants

$$\frac{1}{2}\mathcal{N}\left(\mu_1, \sigma^2\right) + \frac{1}{2}\mathcal{N}\left(\mu_2, \sigma^2\right),$$

où la variance commune est considérée comme connue. On essaie d'estimer les moyennes via un algorithme de Gibbs classique. Lorsque la variance commune est petite, les deux composants du mélange sont bien séparés et la chaîne permettant d'estimer μ_1 reste aux alentours de la vraie valeur sans être influencée par les données issues de l'autre composant. C'est à dire que les deux modes de la loi a posteriori sont bien séparés et que la chaîne reste autour d'un seul des deux. Les lois a posteriori obtenues sont alors unimodales et centrées autour de la vraie valeur.

On dispose maintenant d'un autre échantillon tiré d'un mélange gaussien identique au précédent, excepté le fait que la variance commune est plus grande. Les composants sont alors plus imbriqués. La chaîne MCMC générée pour μ_1 par exemple, peut très bien atteindre une zone où l'influence des données issues du second composant est prépondérante. La chaîne peut donc se stabiliser autour de la vraie valeur de μ_2. Elle est alors en train d'estimer μ_2 au lieu de μ_1, tandis que l'autre chaîne fait l'inverse. Dans ce cas là, les modes de la loi a posteriori sont suffisamment proches pour que la chaîne arrive à passer de l'un à l'autre. Tout se passe comme si les deux chaînes avaient été permutées.

Ces permutations apparentes peuvent survenir à plusieurs reprises lors de l'évolution de la chaîne. Lorsque ce phénomène est fréquent (signe de déplacements de la chaîne d'un mode à l'autre), les chaînes ont toutes exploré les mêmes valeurs et présentent toutes le phénomène de multimodalité, caractéristique du "label switching".

Pour illustrer le phénomène, nous avons généré un échantillon de taille 250 issu d'une densité de mélange de la forme

$$\frac{1}{2}\mathcal{N}\left(0, 6\right) + \frac{1}{2}\mathcal{N}\left(6, 6\right).$$

La variance n'est pas considérée comme connue. On cherche donc à estimer les paramètres de ce mélange à l'aide de l'algorithme de Gibbs présenté au chapitre II. Les mélanges de densités bimodales comme en figure IV.1 sont favorables à l'apparition du phénomène de label switching. En effet, l'échantillon obtenu après 5000 itérations (dont 1000 de temps de chauffe) de l'échantillonnage de Gibbs (figure IV.2) présente bien les symptômes d'une "inversion" des chaînes estimant les deux moyennes. Les traits pointillés représentant la moyenne a posteriori donnent une idée de l'erreur d'estimation commise. Nous allons maintenant détailler les deux solutions que nous nous proposons d'étudier et de comparer.

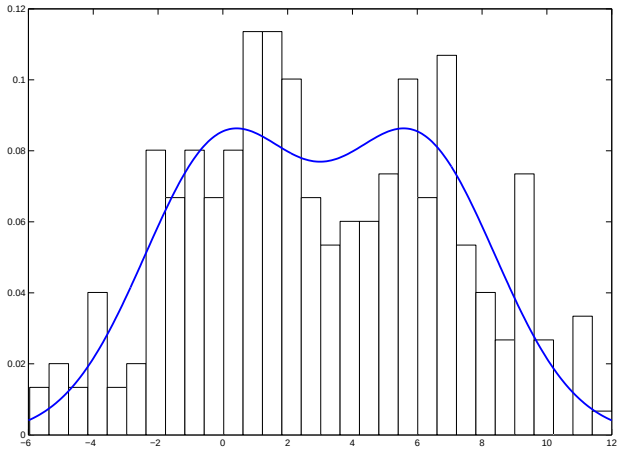


FIG. IV.1 – Histogramme de l'échantillon généré superposé à sa densité

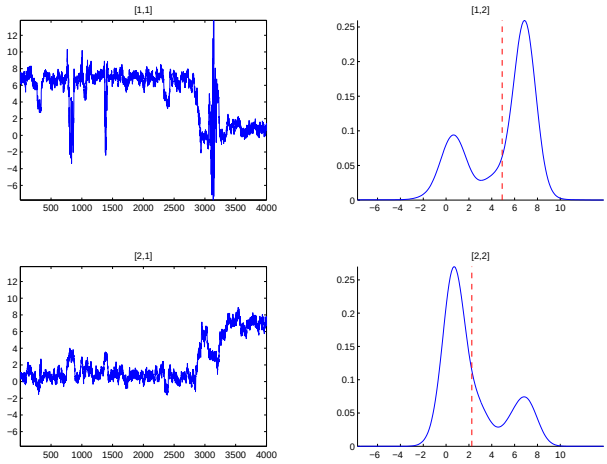


FIG. IV.2 – Loi a posteriori pour les moyennes produite par l'algorithme de Gibbs avec présence de label switching

2 La solution de Stephens

L'idée apportée par Stephens [2000b] est d'éliminer cet effet de changement d'indice en recherchant des permutations $\nu_1, \nu_2, ..., \nu_N$ de manière à ce que les paramètres permutés à chaque étape de l'algorithme, c.a.d. $\nu_1\left(\theta^{(1)}\right), \nu_2\left(\theta^{(2)}\right), ..., \nu_N\left(\theta^{(N)}\right)$, soient tous indicés de manière équivalente.

Le principe est de s'assurer que le $i^{ème}$ élément du jeu de paramètres à l'étape N ($i^{ème}$ composant de $\nu_N\left(\theta^{(N)}\right)$) correspond toujours à l'estimation courante des paramètres du $i^{ème}$ composant. Rappellons que $\theta^{(j)}$ désigne le vecteur des paramètres $\left(\pi_j, \mu_j, \Sigma_j\right)$.

On dit que $\theta^{(1)}$ et $\theta^{(2)}$ sont indicés de manière équivalente ou "correctement indicés" lorsque :

$$D\left[\theta^{(1)} \parallel \theta^{(2)}\right] \stackrel{def}{=} \sum_{i=1}^{k} \Delta\left[\pi_i^{(1)} \mathcal{N}\left(.; \mu_i^{(1)}, \Sigma_i^{(1)}\right) \parallel \pi_i^{(2)} \mathcal{N}\left(.; \mu_i^{(2)}, \Sigma_i^{(2)}\right)\right]$$

est la plus petite possible, avec $\Delta\left[\theta \parallel \theta'\right]$ une mesure de divergence entre densités paramétrées

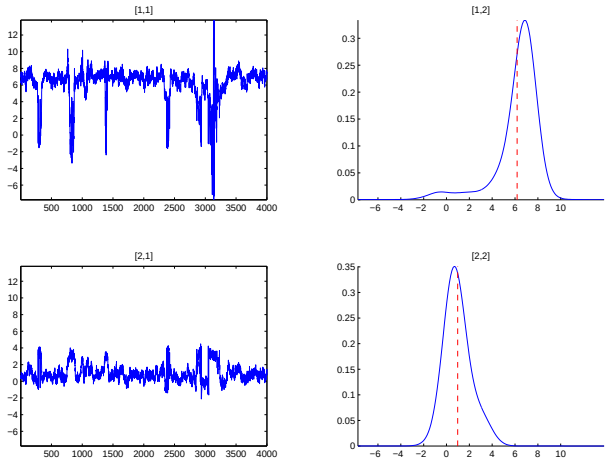


FIG. IV.3 – Loi a posteriori pour les moyennes après la gestion du label switching par la méthode de Stephens

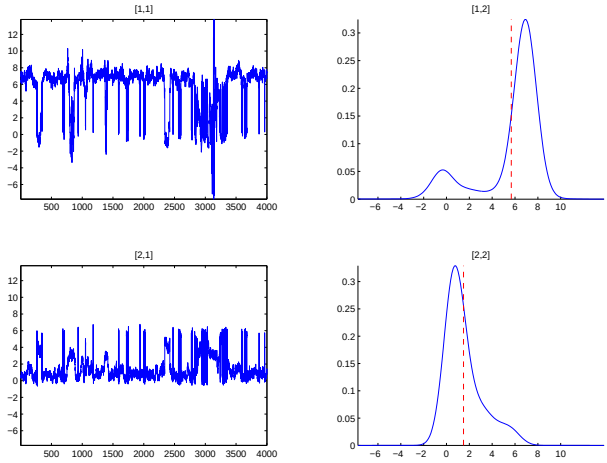


FIG. IV.4 – Loi a posteriori pour les moyennes après la gestion du label switching par la méthode de Celeux

par θ et θ'.

La procédure de post-traitement des données $\left(\theta^{(1)}, ..., \theta^{(N)}\right)$ consiste donc à trouver les permutations $\nu_1, ..., \nu_N$ et une valeur $\hat{\theta}$ du paramètre, qui permettent de minimiser

$$\mathcal{D} \stackrel{def}{=} \sum_{t=1}^{N} D\left[\nu_t\left(\theta^{(t)}\right) \parallel \hat{\theta}\right].$$

En suivant Stephens [2000b], nous utiliserons une mesure de divergence entre une densité pon-

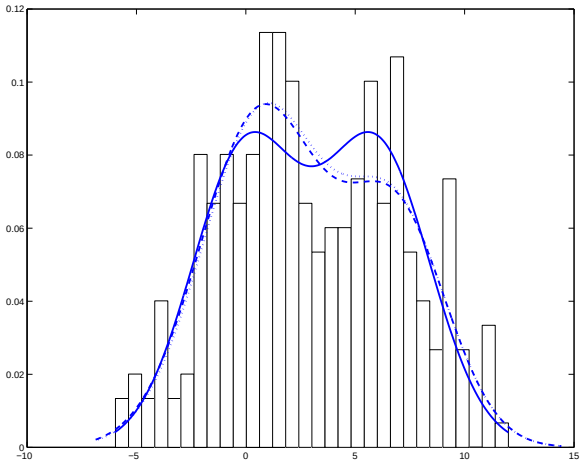


FIG. IV.5 – Histogramme de l'échantillon généré superposé à sa densité réelle (trait plein). En pointillés, la densité estimée après gestion du label switching par la méthode de Celeux et en tirets la densité estimée après gestion du label switching par la méthode de Stephens.

dérée $pf(.)$ et une autre notée $qg(.)$. Cette mesure est définie par (Stephens [2000b] page 49) :

$$\begin{aligned}
&\Delta\left[\pi_i^{(1)}\mathcal{N}\left(.;\mu_i^{(1)},\Sigma_i^{(1)}\right) \parallel \pi_i^{(2)}\mathcal{N}\left(.;\mu_i^{(2)},\Sigma_i^{(2)}\right)\right] \\
= \;& \pi_i^{(1)}\log\left(\frac{\pi_i^{(1)}}{\pi_i^{(2)}}\right) + \left(1-\pi_i^{(1)}\right)\log\left(\frac{\left(1-\pi_i^{(1)}\right)}{\left(1-\pi_i^{(2)}\right)}\right) \\
& +\pi_i^{(1)}\int \mathcal{N}\left(x;\mu_i^{(1)},\Sigma_i^{(1)}\right)\log\frac{\mathcal{N}\left(x;\mu_i^{(1)},\Sigma_i^{(1)}\right)}{\mathcal{N}\left(x;\mu_i^{(2)},\Sigma_i^{(2)}\right)}dx,
\end{aligned}$$

qui est une adaptation de la divergence de Kullback-Leibler aux densités pondérées.

L'algorithme est le suivant :

Algorithme 3.1 : On initialise les permutations $\nu_1,...,\nu_N$ (égales à l'identité par exemple), pour ensuite réitérer les étapes suivantes jusqu'à l'obtention d'un point fixe.

`Etape 1 :` choisir $\hat{\theta}$ qui minimise $\sum_{t=1}^{N} D\left[\nu_t\left(\theta^{(t)}\right) \parallel \hat{\theta}\right]$

`Etape 2 :` Pour $t=1,...,N$ choisir ν_t qui minimise :
$D\left[\nu_t\left(\theta^{(t)}\right) \parallel \hat{\theta}\right] = \sum_{i=1}^{k}\Delta\left[\pi_{\nu_t(i)}^{(t)}\mathcal{N}\left(.;\mu_{\nu_t(i)}^{(t)},\Sigma_{\nu_t(i)}^{(t)}\right) \parallel \hat{\pi}_i\mathcal{N}\left(.;\hat{\mu}_i,\hat{\Sigma}_i\right)\right]$.

Selon Stephens, cet algorithme peut être vu comme un algorithme de type k-means avec $k!$ classes correspondants aux $k!$ manières d'ordonner les paramètres des composants. L'idée est de regrouper les points de l'échantillon $(t=1,...,N)$ qui fournissent des densités estimées identiques pour chaque composant.

Plus précisément, si on appelle $\{\pi_1\mathcal{N}(.;\mu_1,\Sigma_1),...,\pi_k\mathcal{N}(.;\mu_k,\Sigma_k)\}$ les densités pondérées des composants, on rassemblera dans une même classe les éléments de la chaîne $(t=1,...,N)$ qui fournissent des densités pondérées équivalentes pour chaque composant. En effet, en présence de label swiching, la densité estimée pour le composant 1 $\left(\text{cad } \pi_1^{(t)}\mathcal{N}\left(.;\mu_1^{(t)},\Sigma_1^{(t)}\right)\right)$ à l'étape t peut être différente de celle estimée à l'étape $(t+1)$ de la chaîne (qui correspond alors à la densité

estimée d'un autre composant). L'étape 1 correspond à l'estimation des $k!$ centres de classes, tandis que l'étape 2 alloue chaque point à la classe la plus proche.

On trouve dans Stephens [2000b] la preuve que cet algorithme atteint un point fixe, il est signalé que celui-ci peut dépendre de l'initialisation. Il est donc suggéré d'exécuter l'algorithme plusieurs fois avec des initialisations différentes afin d'évaluer la qualité des résultats.

Les calculs nécessaires à l'étape 1 tiennent compte du choix de $\sum_{i=1}^{k} \Delta\left[\ .\ \|\ .\right]$. L'étape 2 peut être vue comme N fois le problème de minimisation suivant :

$$\begin{aligned} &\text{choisir } \nu \text{ qui minimise } \sum_{i=1}^{k} c\left(i,\nu\left(i\right)\right) \qquad \text{(IV.1)}\\ &\text{où } c\left(i,l\right) \text{ est de la forme suivante}\\ &c\left(i,l\right)=\Delta\left[\pi_l^{(1)}\mathcal{N}\left(.;\mu_l^{(1)},\Sigma_l^{(1)}\right)\ \|\ \pi_i^{(2)}\mathcal{N}\left(.;\mu_i^{(2)},\Sigma_i^{(2)}\right)\right]. \end{aligned}$$

Le problème précédent (IV.1) est alors équivalent au problème de programmation entière suivant :

$$\begin{aligned} &\text{choisir } \left\{y_{il}\right\}\left(i=1,...,k,l=1,...,k\right) \text{ qui minimise } \sum_{i=1}^{k}\sum_{l=1}^{k} y_{il}c\left(i,l\right) \qquad \text{(IV.2)}\\ &\text{sous les contraintes } y_{il}=0 \text{ ou } 1 \text{ et } \sum_{i=1}^{k} y_{il}=\sum_{l=1}^{k} y_{il}=1. \end{aligned}$$

Si $\left\{\hat{y}_{il}\right\}$ est une solution optimale pour le problème (IV.2) alors la solution optimale du problème (IV.1) est $\hat{\nu}\left(i\right)=l$ si et seulement si $\hat{y}_{il}=1$.

Le problème (IV.2) est connu comme étant une version spéciale du problème de transport, nommée "assignment problem". De nombreux algorithmes permettant de résoudre ce problème existent. Nous avons utilisé la boîte à outils "optimisation" de Matlab.

Lorsqu'on utilise la distance définie plus haut, il est possible d'expliciter les calculs de l'algorithme 3.1. On obtient la forme suivante :

Algorithme 3.2 : On initialise les permutations $\nu_1,...,\nu_N$ (égales à l'identité par exemple), pour ensuite réitérer les étapes suivantes jusqu'à l'obtention d'un point fixe.

`Etape 1 :` Soit $\hat{\theta}$ donné par :

$$\begin{aligned} &\hat{\pi}_i=\tfrac{1}{N}\textstyle\sum_{t=1}^{N}\pi_{\nu_t(i)}^{(t)}\\ &\hat{\mu}_i=\left(\textstyle\sum_{t=1}^{N}\pi_{\nu_t(i)}^{(t)}\mu_{\nu_t(i)}^{(t)}\right)\big/\textstyle\sum_{t=1}^{N}\pi_{\nu_t(i)}^{(t)}\\ &\hat{\Sigma}_i=\textstyle\sum_{t=1}^{N}\pi_{\nu_t(i)}^{(t)}\left[\Sigma_{\nu_t(i)}^{(t)}+\left(\mu_{\nu_t(i)}^{(t)}-\hat{\mu}_i\right)\left(\mu_{\nu_t(i)}^{(t)}-\hat{\mu}_i\right)'\right]\big/\textstyle\sum_{t=1}^{N}\pi_{\nu_t(i)}^{(t)}. \end{aligned}$$

`Etape 2 :` Pour $t=1,...,N$ choisir ν_t qui minimise :

$$\begin{aligned} &\textstyle\sum_{i=1}^{k}\Big\{\pi_{\nu_t(i)}^{(t)}\frac{1}{2}\log\left|\hat{\Sigma}_i\right|+\pi_{\nu_t(i)}^{(t)}\frac{1}{2}tr\left[\hat{\Sigma}_i^{-1}\left(\Sigma_{\nu_t(i)}^{(t)}+\left(\mu_{\nu_t(i)}^{(t)}-\hat{\mu}_i\right)\left(\mu_{\nu_t(i)}^{(t)}-\hat{\mu}_i\right)'\right)\right]\\ &-\pi_{\nu_t(i)}^{(t)}\log\hat{\pi}_i-\left(1-\pi_{\nu_t(i)}^{(t)}\right)\log\left(1-\hat{\pi}_i\right)\Big\}. \end{aligned}$$

qui est celle que nous utilisons. L'étape 1 est uniquement calculatoire, tandis que l'étape 2 est résolue via le problème de transport (IV.2) avec :

$$\begin{aligned} c\left(i,l\right)=\sum_{i=1}^{k}\Big\{&\pi_l^{(t)}\frac{1}{2}\log\left|\hat{\Sigma}_i\right|+\pi_l^{(t)}\frac{1}{2}tr\left[\hat{\Sigma}_i^{-1}\left(\Sigma_l^{(t)}+\left(\mu_l^{(t)}-\hat{\mu}_i\right)\left(\mu_l^{(t)}-\hat{\mu}_i\right)'\right)\right]\\ &-\pi_l^{(t)}\log\hat{\pi}_i-\left(1-\pi_l^{(t)}\right)\log\left(1-\hat{\pi}_i\right)\Big\}. \end{aligned}$$

3 La solution de Celeux

La solution proposée dans Celeux [1998] est dans le même esprit que celle vue précédemment. Il s'agit d'une version séquentielle de l'algorithme des k-means appliqué à la séquence des données normalisées. Les $k!$ centres des k-means sont déterminés à l'aide des m premières itérations (que l'on suppose sans présence de label switching, en prenant généralement $m = 100$), et chacun des éléments suivants est permuté de manière à être associé au centre de référence (c'est-à-dire celui obtenu sans aucune permutation des m premières données). On détermine ainsi la suite des permutations nécessaires pour réordonner les données.

La procédure se détaille comme suit :

Soit $\theta^{(1)}, ..., \theta^{(N)}$ la séquence des vecteurs des paramètres de dimension p obtenue par un algorithme MCMC (ou tout autre algorithme stochastique). L'algorithme est initialisé avec $k!$ classes dont les centres sont déterminés de la façon suivante.

Pour $i = 1, ..., p$ la variance de chaque coordonnée θ_i est calculée :

$$\left(s_i^{[0]}\right)^2 = \frac{1}{m}\sum_{j=1}^{m}\left(\theta_i^{(j)} - \bar{\theta}_i\right)^2 \qquad \text{avec } \bar{\theta}_i = \frac{1}{m}\sum_{j=1}^{m}\theta_i^{(j)}.$$

Le centre initial de référence est défini par $\bar{\theta}_1^{[0]} = \bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\theta^{(j)}$. Les $k! - 1$ autres centres sont déduits en permutant les indices des paramètres des composants du mélange dans $\bar{\theta}_1^{[0]}$. A partir de cette position initiale, la r-ième itération de l'algorithme séquentiel se détaille en deux étapes :

Algorithme 3.3 : Pour r variant de 1 à $N - m$, on répète les étapes suivantes

`Etape 1 :` Le vecteur $\theta^{(m+r)}$ est assigné à la classe j^* qui minimise la distance normalisée :
$\left\|\theta^{(m+r)} - \bar{\theta}_j^{[r-1]}\right\| = \sum_{i=1}^{p}\frac{\left(\theta^{(m+r)} - \bar{\theta}_{ij}^{[r-1]}\right)^2}{\left(s_i^{[r-1]}\right)^2}, \quad j = 1, ..., k!$
où $\bar{\theta}_{ij}^{[r-1]}$ représente la $i-$ème coordonnée. Si $j^* \neq 1$, les coordonnées du vecteur $\theta^{(m+r)}$ sont permutées de façon à avoir $j^* = 1$, c'est-à-dire de manière à retrouver l'indiçage initial.

`Etape 2 :` Les $k!$ centres et les p coefficients s_i^2 sont mis à jour. Le centre de référence devient
$\bar{\theta}_1^{[r]} = \frac{m+r-1}{m+r}\bar{\theta}_1^{[r-1]} + \frac{1}{m+r}\theta^{(m+r)}$
et les $k! - 1$ centres restants sont à nouveau déduits de $\bar{\theta}_1^{[r]}$ en permutant les indices des composants. Les variances sont mises à jour de la façon suivante : pour $i = 1, ..., p$
$\left(s_i^{[r]}\right)^2 = \frac{m+r-1}{m+r}\left(s_i^{[r-1]}\right)^2 + \frac{m+r-1}{m+r}\left(\bar{\theta}_i^{[r-1]} - \bar{\theta}_i^{[r]}\right)^2 + \frac{1}{m+r}\left(\bar{\theta}_i^{(m+r)} - \bar{\theta}_i^{[r]}\right)^2.$

4 Comparaison des méthodes

La méthode présentée dans Stephens [2000b] est un algorithme nécessitant la chaîne générée dans son ensemble. La procédure nécessite donc un espace mémoire important ainsi que la résolution d'un problème de transport en utilisant des techniques de programmation entière.

L'algorithme 3.3 ne nécessite quant à lui qu'une manipulation de permutations et peut s'utiliser de manière séquentielle. Il est en effet beaucoup plus simple à programmer et conduit à un temps de calcul très réduit. A l'inverse, l'algorithme 3.2 est très coûteux en temps de calcul.

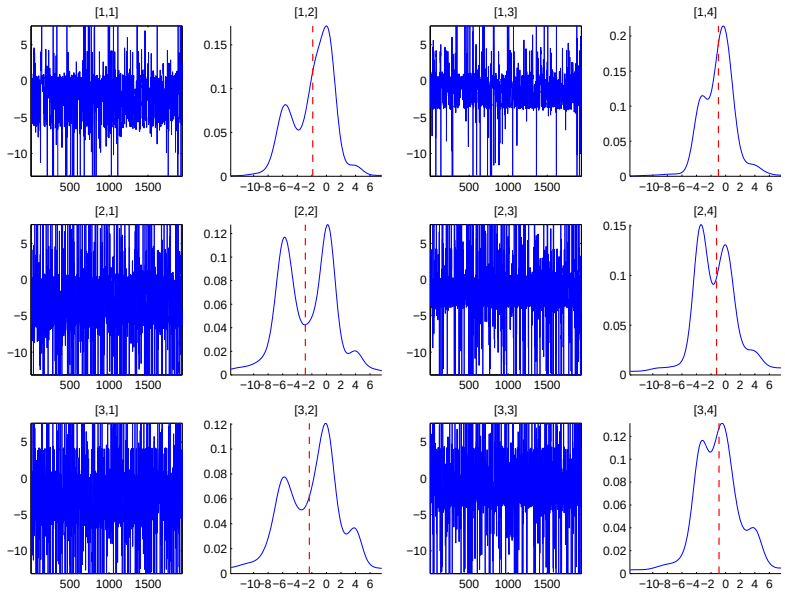


FIG. IV.6 – Lois a posteriori pour les moyennes produite par l'algorithme de Gibbs dans le cas multivarié, avec présence de label switching

Dans un cadre univarié, il est difficile d'apprécier graphiquement la qualité de l'estimation d'une méthode à l'autre. En effet, celles-ci éliminent toutes les deux le phénomène de label switching présent, et ceci dans la même proportion. En effet, en observant les figures IV.3 et IV.4, même si la méthode de Celeux n'a pas bien réordonné l'intégralité des points, comme l'a presque fait la méthode de Stephens, l'impact est relativement faible sur la distribution a posteriori. La figure IV.5 nous confirme que l'écart d'estimation des densités de mélange est très faible.

L'algorithme 3.3 peut donc sembler pertinent dans un cadre univarié où la présence de label switching est relativement réduite et aisée à repérer. L'algorithme 3.2 est plus complexe et plus long pour des résultats légèrement meilleurs.

Le cas multivarié présente des difficultés supplémentaires. En effet nos simulations effectuées avec l'algorithme à sauts réversibles présenté au chapitre suivant ont mises en évidence un phénomène de label switching quasiment systématique. Le nombre important de paramètres de la chaîne semble favoriser son apparition, pour la simple raison qu'une permutation d'indice peut être provoquée par n'importe quelle coordonnée, entraînant ainsi toutes les autres. Sur la figure IV.6 on peut apprécier l'effet du label switching sur l'estimation des moyennes du mélange à 3 composants présenté au chapitre II. Le composant ayant la plus faible proportion (0.1) est celui qui est le plus mal estimé et ayant la multimodalité la plus prononcée. La multimodalité (présente sur chaque moyenne) est à peu près éliminée par les 2 algorithmes pour les 2 premiers composants, ayant la plus forte proportion (0.4 et 0.5).

Le troisième composant est assez mal estimé, étant donné que le nombre d'observations lui appartenant est faible. C'est donc sur lui que se concentrent les faiblesses des algorithmes 3.2 et 3.3. Les figures IV.7 et IV.8 montrent les résultats des 2 algorithmes appliqués à la même chaîne que celle de la figure IV.6. Les faiblesses de l'algorithme 3.3 apparaissent ici plus nettes. La multimodalité pour le second composant n'a pas complètement disparue, alors que l'algorithme 3.2 donne une estimation très correcte des 2 premiers composants.

Le choix entre ces deux algorithmes semble donc se situer au niveau suivant : quel temps de calcul est on prêt à sacrifier pour une bonne estimation ?

L'algorithme 3.3 est très simple à programmer et le temps de calcul est très réduit. Les solutions au label switching apportées par ce dernier semblent légèrement moins pertinentes

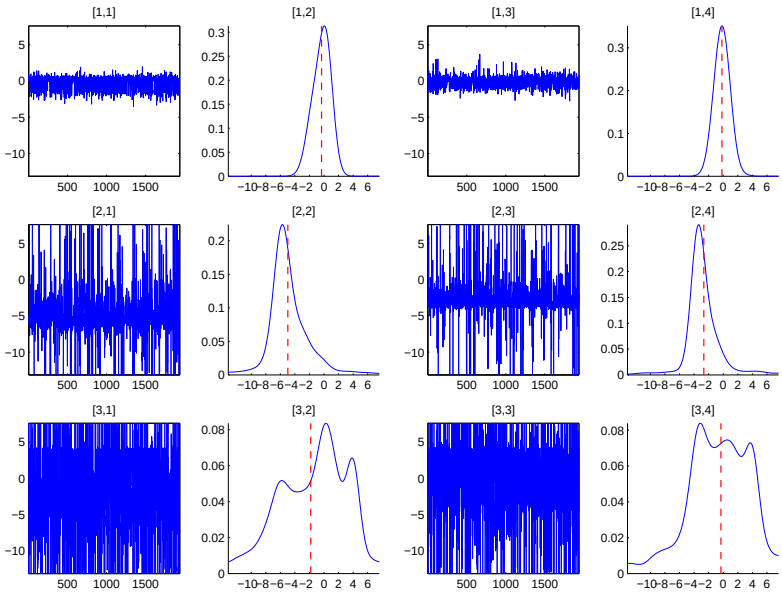


FIG. IV.7 – Loi a posteriori pour les moyennes après la gestion du label switching par la méthode de Stephens

que celles de l'algorithme 3.2, mais mènent néanmoins à des solutions comparables. Pour s'en convaincre, on peut considérer les figures IV.9, IV.10 et IV.11 qui nous donnent les estimations finales de la densité du mélange obtenues par ces 2 algorithmes. Les faibles différences permettent de relativiser la supériorité de l'algorithme 3.2.

Pour nos simulations, nous effectuons systématiquement les 2 algorithmes, même si nous ne présentons souvent que les résultats obtenus par l'algorithme 3.2.

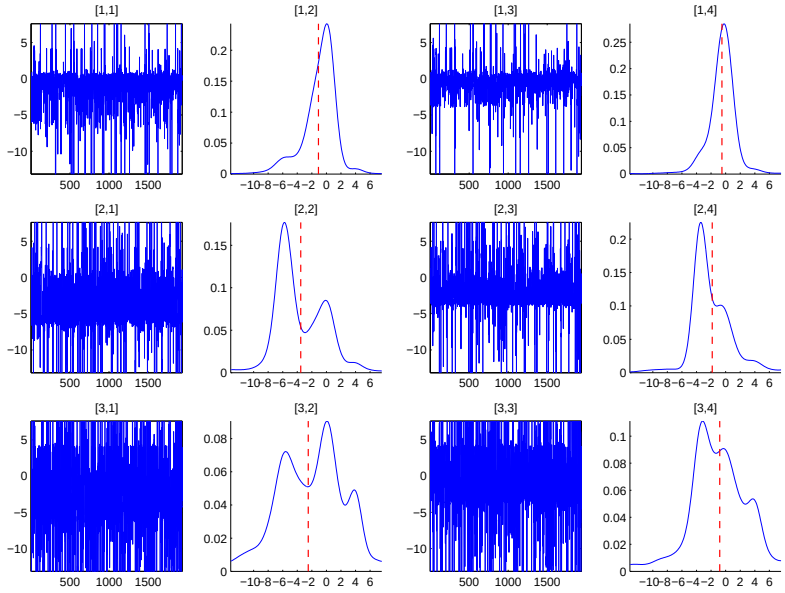


FIG. IV.8 – Loi a posteriori pour les moyennes après la gestion du label switching par la méthode de Celeux

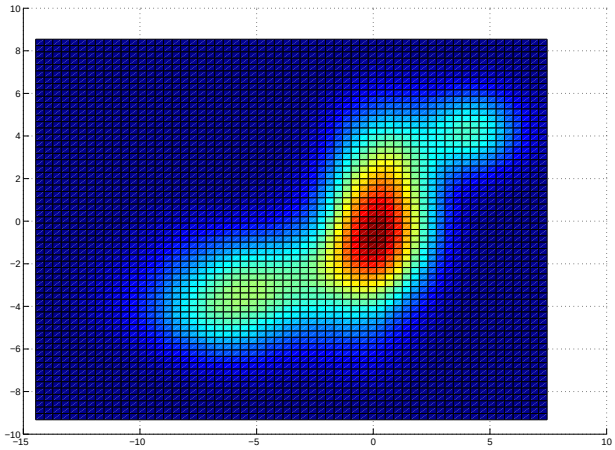


FIG. IV.9 – Estimation par la méthode des noyaux de la densité du mélange multivarié original.

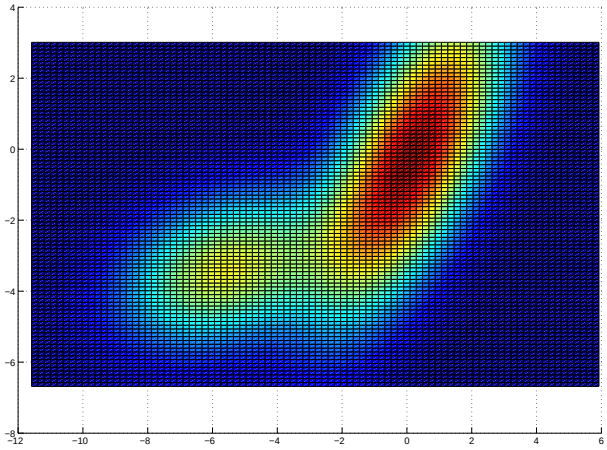


FIG. IV.10 – Estimation par la méthode des noyaux de la densité estimée après la gestion du label switching par la méthode de Celeux.

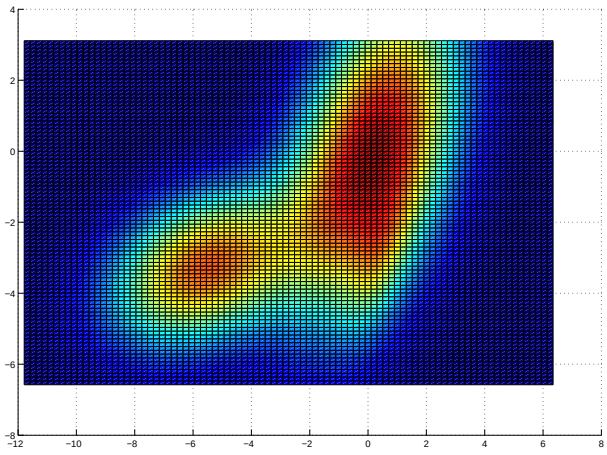


FIG. IV.11 – Estimation par la méthode des noyaux de la densité estimée après la gestion du label switching par la méthode de Stephens.

Chapitre V

Méthode MCMC à sauts réversibles appliquée aux mélanges gaussiens multivariés

1 Introduction

Dans ce chapitre, nous allons donner les détails de l'extension de l'algorithme à sauts réversibles présenté dans Richardson et Green [1997] au cas des mélanges gaussiens multivariés. Nous détaillons les diverses lois multidimensionnelles utilisées, ainsi que les divers mouvements nécessitant un changement de dimension du modèle. En effet, la différence entre l'algorithme à sauts réversibles et l'algorithme de Gibbs réside justement dans ces nouveaux types de mouvements qui sont difficiles à écrire. La nécessité du calcul d'un jacobien de dimension élevée rend la tâche particulièrement ardue dans le cas multivarié.

L'approche adoptée dans ce chapitre, est une généralisation fidèle à l'esprit de Richardson et Green [1997]. Nous avons utilisé le même modèle hiérarchique de lois a priori, avec notamment l'adjonction d'un hyperparamètre β à la chaîne de Markov générée. De la même façon, nous reprenons les mouvements de naissance/mort et de séparation/combinaison. Il est suggéré dans la discussion faisant suite à Richardson et Green [1997] d'étoffer ces mouvements afin d'améliorer l'exploration de l'espace d'état. Les très bonnes capacitées d'explorations de l'algorithme rencontrées dans nos simulations, nous ont cependant incités à ne pas autoriser d'autres mouvements changeant la dimension du modèle.

Afin de conserver une probabilité élevée de changer de dimension, nous avons aussi conservé la condition sur les moments appelée "dimension matching condition" par Richardson et Green [1997]. Cette condition est cependant très contraignante dans le cas multivarié, et un des travaux d'extension envisagé est de spécifier des mouvements permettants de s'en affranchir.

L'exposé détaillé de la méthode MCMC à sauts réversibles appliquée au cas des mélanges gaussiens multivariés suit donc la présentation adoptée par les auteurs originaux. Nous traitons en premier le mouvement de naissance/mort, celui-ci étant le plus simple, pour traiter ensuite le mouvement de séparation/combinaison contenant les difficultés les plus importantes. Nous explicitons le calcul de la probabilité d'acceptation de ce dernier mouvement ainsi que le calcul du jacobien.

2 Algorithme et notations

On rappelle la densité d'un mélange gaussien multivarié à k composants :

$$p\left(y \mid \pi, \mu, \Sigma\right) = \pi_1 \mathcal{N}\left(y; \mu_1, \Sigma_1\right) + ... + \pi_k \mathcal{N}\left(y; \mu_k, \Sigma_k\right).$$

La vraisemblance associée à l'échantillon $y^{(n)} = (y_1, ..., y_n)$ est alors

$$l\left(y^{(n)} \mid \pi, \mu, \Sigma\right) = \prod_{i=1}^{n} \left[\pi_1 \mathcal{N}\left(y_i; \mu_1, \Sigma_1\right) + ... + \pi_k \mathcal{N}\left(y_i; \mu_k, \Sigma_k\right)\right].$$

On introduit les modèles à données manquantes comme au chapitre introductif (page 3), c'est-à-dire que l'on introduit les variables z_i telles que

$$\mathbb{P}\left[z_i = j\right] = \pi_j, \quad \text{pour } i = 1, ..., n \text{ et } j = 1, ..., k$$

et on note $z^{(n)} = (z_1, ..., z_n)$. Suivant Richardson et Green [1997], on rajoute une couche d'hyperparamètres en autorisant les variances à dépendre d'un paramètre β faisant donc partie intégrante de la chaîne générée. Ceci a pour but d'améliorer la flexibilité de l'algorithme.

On met en place une mise à jour des paramètres de manière successive en utilisant différents types de mouvements. Rappelons que la méthode à sauts réversibles ne s'applique que lorsque le composant mis à jour induit un changement de dimension du modèle, ou autrement dit, lorsque le mouvement change de modèle. Dans le cas contraire on utilise toujours l'échantillonnage de Gibbs, ceci signifiant que l'on génère la nouvelle valeur du paramètre selon sa loi conditionnelle a posteriori. On combine en fait l'algorithme à sauts réversibles et l'échantillonnage de Gibbs. Ceci permet de mieux comprendre le terme d'hybride utilisé pour qualifier ce type d'algorithme.

L'état courant de la chaîne est $\left(k, \pi, z^{(n)}, \beta, \mu, \Sigma\right)$. On met successivement à jour les différents composants selon l'ordre suivant :

a- Séparer ou combiner deux composants.
b- Mise à jour de π.
c- Mise à jour de (μ, Σ).
d- Mise à jour des allocations $z^{(n)}$.
e- Mise à jour de l'hyperparamètre β.
f- Naissance ou mort d'un composant.

Les mouvements b,c,d et e nécessitent une mise à jour automatique selon les lois conditionnelles car il n'y a pas de changement de modèle. On pourra se référer dans ce cas à Diebolt et Robert [1994] ou aux rappels du chapitre II. Les mouvements a et f nécessitent quant à eux l'emploi des sauts réversibles. L'ensemble de ces mouvements (de a à f) est répété 25 000 fois. A chaque itération on garde en mémoire les paramètres courants. Lorsqu'on stoppe l'algorithme, les 5000 premières itérations sont enlevées et on effectue une analyse a posteriori sur les 20 000 éléments de la chaîne restants. Le choix de ce petit nombre d'itérations en regard de la taille élevée des espaces à explorer, proviens surtout de limitations liées aux moyens informatiques. La programmation de l'algorithme en Matlab est simple, mais ce langage possède l'inconvénient de ralentir la procédure par rapport à des langages compilés comme le C++ ou le Fortran. Il est envisagé de transcrire ce programme en Fortran afin de pouvoir evaluer les capacités de l'algorithme sur un nombre d'itérations comparable à Richardson et Green [1997], ce qui nous est impossible pour le moment.

Comme on le verra au chapitre suivant, cet algorithme se déplace énormément sur l'espace d'état, en particulier, il est très rare d'effectuer les mouvements a à f plus de 2 ou 3 fois sans observer de changement de dimension. Nous avons donc légèrement modifié le schéma hybride proposé dans Richardson et Green [1997] de manière à rester un temps légèrement plus long dans chaque modèle traversé. En respectant la théorie des algorithmes hybrides (cf Robert et Casella [1999] par exemple) ainsi que le schéma qui vient d'être présenté, nous effectuons donc cinq fois

les mouvements b à e après chaque changement de dimension accepté par les mouvements a ou f. Afin de ne pas perturber la loi a posteriori sur le nombre de composants, nous ne retenons que la dernière valeur générée. La nouvelle valeur du paramètre généré par les mouvements a ou f, sera donc celle obtenue après la procédure des sauts réversibles, suivie de ces 5 itérations de Gibbs, en ne retenant au final que la dernière valeur des paramètres obtenue. Cet élément de la chaîne de Markov suit la même loi que celle obtenue après avoir effectué les mouvements a à e.

Afin d'éclaircir les notations utilisées, nous faisons ici un parallèle avec la présentation utilisée au chapitre III. Dans un premier temps on ne tient pas compte de β, et pour faire l'analogie avec (III.1) on note $\theta^{(k)} = \left(\pi, z^{(n)}, \mu, \Sigma\right)$ avec

$$\begin{aligned} \pi &= (\pi_1, ..., \pi_k), \\ \mu &= (\mu_1, ..., \mu_k), \\ \Sigma &= (\Sigma_1, ..., \Sigma_k). \end{aligned}$$

On peut alors écrire d'après (III.1),

$$p\left(k, \theta^{(k)}, y^{(n)}\right) = p(k)\, p\left(\theta^{(k)} \mid k\right) p\left(y^{(n)} \mid k, \theta^{(k)}\right)$$

c'est-à-dire

$$p\left(k, \pi, z^{(n)}, \mu, \Sigma, y^{(n)}\right) = p(k)\, p(\pi \mid k)\, p\left(z^{(n)} \mid k, \pi\right) p\left(\mu, \Sigma \mid k, \pi, z^{(n)}\right) p\left(y^{(n)} \mid k, \theta^{(k)}\right).$$

Or, on peut simplifier certains de ces termes

$$\begin{aligned} p\left(\mu, \Sigma \mid k, \pi, z^{(n)}\right) &= p(\mu, \Sigma \mid k), \\ p\left(y^{(n)} \mid k, \theta^{(k)}\right) &= p\left(y^{(n)} \mid z^{(n)}, \mu, \Sigma\right). \end{aligned}$$

On a alors

$$p\left(k, \pi, z^{(n)}, \mu, \Sigma, y^{(n)}\right) = p(k)\, p(\pi \mid k)\, p\left(z^{(n)} \mid k, \pi\right) p(\mu, \Sigma \mid k)\, p\left(y^{(n)} \mid z^{(n)}, \mu, \Sigma\right). \quad \text{(V.1)}$$

3 Modèle hiérarchique complet

On va maintenant introduire le modèle hiérarchique bayésien. Suivant Richardson et Green [1997], le paramètre β est partie intégrante de la chaîne, et dépend lui aussi de certains hyperparamètres. Cela permet d'améliorer la flexibilité du modèle. On autorise les lois a priori sur les paramètres μ, Σ, k, π, à dépendre respectivement des hyperparamètres $(\xi, \kappa, \alpha, \beta, \lambda, \delta)$ avec $\eta = (\xi, \kappa, \alpha, \beta)$. Le modèle hiérarchique bayésien s'écrit donc

$$\begin{aligned} & p\left(k, \pi, z^{(n)}, \mu, \Sigma, y^{(n)}\right) \\ = \;& p(\lambda)\, p(\delta)\, p(\eta)\, p(k \mid \lambda)\, p(\pi \mid k, \delta)\, p\left(z^{(n)} \mid k, \pi\right) p(\mu, \Sigma \mid k, \eta)\, p\left(y^{(n)} \mid z^{(n)}, \mu, \Sigma\right). \end{aligned}$$

3.1 Lois a priori

Les lois a priori utilisées sont les suivantes :

$$\begin{aligned} p(k \mid \lambda) &\sim \mathcal{P}(\lambda) : \text{ loi de Poisson tronquée à } k_{\max} \quad \text{(V.2)} \\ p(\pi \mid k, \delta) &\sim \mathcal{D}(\delta, ..., \delta) : \text{ loi de Dirichlet de dimension } k \\ p(z_i \mid k, \pi) &\sim \sum_{j=1}^{k} \pi_j \delta_j \text{ pour } i = 1, ..., n \text{ et } j = 1, ..., k \\ p\left(\mu_j \mid k, \eta\right) &\sim \mathcal{N}_r\left(\xi, \kappa^{-1}\right) \quad \text{pour } j = 1, ..., k \text{ loi Normale en dimension } r \\ p\left(\Sigma_j^{-1} \mid k, \eta\right) &\sim W_r\left(2\alpha, (2\beta)^{-1}\right) \quad \text{pour } j = 1, ..., k \text{ loi de Wishart en dimension } r. \end{aligned}$$

Afin de rajouter un niveau de flexibilité pour les variances des composants, on autorise β à suivre une Loi de Wishart en dimension r, c'est-à-dire que l'on a

$$p\left(\beta \mid g, h\right) \sim W_r\left(2g, (2h)^{-1}\right),$$

où g et h sont deux nouveaux hyperparamètres. Stephens [1997] fait justement remarquer que cette loi est impropre lorsque $2g < r$. Comme on l'a vu précédemment au chapitre II, une étude approfondie effectuée par Stephens montre cependant que la loi conditionnelle associée est propre. On peut donc tout de même utiliser cette loi. Le modèle hiérarchique complet peut se résumer par le schéma (V.1), appelé "directed acyclic graph" dans la littérature.

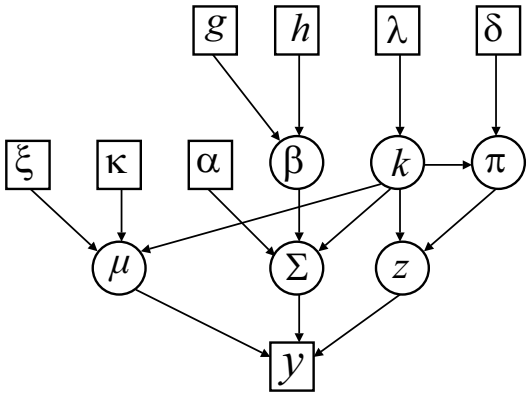


FIG. V.1 – Graphe acyclique ordonné correspondant au modèle bayésien pour les lois de mélanges multivariées

La loi conjointe des paramètres pour le modèle complet peut donc finalement s'écrire

$$\begin{aligned} p\left(\beta, k, \pi, z^{(n)}, \mu, \Sigma, y^{(n)}\right) &= p\left(\beta \mid g, h\right) p\left(k \mid \lambda\right) p\left(\pi \mid k, \delta\right) \\ &\quad \times p\left(z^{(n)} \mid k, \pi\right) p\left(\mu, \Sigma \mid k, \eta\right) p\left(y^{(n)} \mid z^{(n)}, \mu, \Sigma\right). \end{aligned} \tag{V.3}$$

La vraisemblance $p\left(y^{(n)} \mid z^{(n)}, \mu, \Sigma\right)$ du modèle s'écrit de manière explicite :

$$p\left(y^{(n)} \mid z^{(n)}, \mu, \Sigma\right) = \prod_{i=1}^{n} \sum_{j=1}^{k} \mathcal{N}_r\left(y_i; \mu_j, \Sigma_j\right) \mathbb{1}_{[z_i=j]}.$$

L'expression complète de la loi conjointe des paramètres est alors :

$$\begin{aligned} p\left(\beta, k, \pi, z^{(n)}, \mu, \Sigma, y^{(n)}\right) &= W_r\left(\beta; 2g, (2h)^{-1}\right) \sum_{l=0}^{k_{\max}} e^{-\lambda} \frac{\lambda^l}{l!} \mathbb{1}_{[k=l]} \\ &\quad \times C_\delta \times \frac{\Gamma\left(k\delta\right)}{\Gamma\left(\delta\right)^k} . \pi_1^{\delta-1} \ldots \pi_{k-1}^{\delta-1} \left(1 - \pi_1 - \ldots - \pi_{k-1}\right)^{\delta-1} \times \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j \mathbb{1}_{[z_i=j]} \\ &\quad \times \prod_{j=1}^{k} \mathcal{N}_r\left(\mu_j; \xi, \kappa^{-1}\right) W_r\left(\Sigma_j^{-1}; 2\alpha, (2\beta)^{-1}\right) \\ &\quad \times \prod_{i=1}^{n} \sum_{j=1}^{k} \mathcal{N}_r\left(y_i; \mu_j, \Sigma_j\right) \mathbb{1}_{[z_i=j]}. \end{aligned} \tag{V.4}$$

Les hyperparamètres $\xi, \kappa, \alpha, g, h, \lambda$ et δ dépendent des données étudiées. Certains sont fixés par l'utilisateur, notamment α, g et δ. On pourra se référer au chapitre II pour voir comment ces hyperparamètres sont spécifiés dans le cas de données bidimensionnelles. La dépendance aux données de ces paramètres permet notamment d'obtenir des lois a priori peu informatives afin d'éviter des biais sur l'inférence.

3.2 Lois conditionnelles a posteriori

On notera, pour le paramètre β par exemple, la loi conditionnelle par rapport à toutes les autres variables de la façon suivante :

$$p\left(\beta \mid k, \pi, z^{(n)}, \mu, \Sigma, y^{(n)}\right) = p\left(\beta \mid ...\right).$$

Les lois a posteriori pour tous les paramètres en jeu sont donc les suivantes :

$$\begin{aligned}
p\left(\beta \mid ...\right) &\sim W_r\left(2g + 2k\alpha, \left(2h + \sum_{j=1}^{k} \Sigma_j^{-1}\right)^{-1}\right), \\
p\left(\pi \mid ...\right) &\sim \mathcal{D}\left(\delta + n_1, \delta + n_2, ..., \delta + n_k\right), \\
\text{pour } j = 1, ..., k \ &: \\
p\left(\mu_j \mid ...\right) &\sim \mathcal{N}_r\left(\left(\kappa + n_j\Sigma_j^{-1}\right)^{-1}\left(\kappa\xi + \Sigma_j^{-1}\sum_{i \text{ tq } z_i = j}^{n} y_i\right), \left(\kappa + n_j\Sigma_j^{-1}\right)^{-1}\right), \\
\text{pour } j = 1, ..., k \ &: \\
p\left(\Sigma_j^{-1} \mid ...\right) &\sim W_r\left(2\alpha + n_j, \left(2\beta + \sum_{i \text{ tq } z_i = j}^{n} \left(y_i - \mu_j\right)\left(y_i - \mu_j\right)'\right)^{-1}\right), \\
\mathbb{P}\left(z_i = j \mid ...\right) &\propto \pi_j \mathcal{N}_r\left(y_i; \mu_j, \Sigma_j\right) \text{ pour } i = 1, ..., n \text{ et } j = 1, ..., k \\
\text{avec } n_j &= \#\left\{i : z_i = j\right\}.
\end{aligned}$$

On ne peut effectuer d'inférence statistique sur le paramètre k représentant l'ordre du modèle car la loi a posteriori conditionnelle est trop complexe. Nous devons alors étudier la loi a posteriori de tous les paramètres, puis extraire l'information concernant k. Cela est résumé par l'écriture suivante :

$$\begin{aligned}
p\left(k, \theta^{(k)} \mid y^{(n)}\right) &= p\left(k, \pi, z^{(n)}, \beta, \mu, \Sigma \mid y^{(n)}\right) = p\left(k \mid y^{(n)}\right) p\left(\pi, z^{(n)}, \beta, \mu, \Sigma \mid y^{(n)}, k\right) \\
&= p_k f_k\left(\theta^{(k)}\right).
\end{aligned}$$

La dernière ligne de cette égalité reprend les notations de (III.5) en page 57.

4 Etude du mouvement de naissance/mort

Le mouvement f permet de faire naître un composant ainsi que d'en éliminer un. Ces deux mouvements doivent être l'inverse l'un de l'autre. On choisit de manière aléatoire de tuer un

composant (réduire la taille du modèle) ou d'en créer un (augmenter la taille du modèle). Pour cela on posera

$$\begin{aligned} b_k &= \text{probabilité de passer à un modèle de dimension supérieure,} \\ d_k &= 1 - b_k = \text{probabilité de passer à un modèle de dimension inférieure.} \end{aligned} \tag{V.5}$$

Remarque 6 *On utilise $d_1 = 0$ et $b_{k_{\max}} = 0$, où $k_{\max}$ est la plus grande valeur autorisée pour le nombre de gaussiennes du mélange. De plus on prend $b_k = d_k = \frac{1}{2}$ pour $k = 2, ..., k_{\max} - 1$. Il est possible de prendre $b_k = d_k = \frac{1}{3}$, et de donner ainsi la probabilité de $\frac{1}{3}$ au fait de ne pas changer de dimension.*

On fait naître un nouveau composant avec la probabilité b_k qui aura pour paramètres les valeurs générées de la façon suivante :

$$\pi_{j*} \sim \beta(1, k) \qquad \mu_{j^*} \sim \mathcal{N}_r\left(\xi, \kappa^{-1}\right) \qquad \Sigma_{j^*}^{-1} \sim W_r\left(2\alpha, (2\beta)^{-1}\right).$$

Il est alors nécessaire de mettre à jour les proportions des composants restants :

$$\pi'_j = \pi_j (1 - \pi_{j*}).$$

On notera que le nouveau composant généré est vide, c'est-à-dire que l'on ne modifie pas les allocations (variables manquantes) des observations. Voici le détail des opérations à effectuer pour le mouvement de naissance :

1. Augmenter le nombre de composants de 1 et créer un composant vide indicé par $k + 1$.
2. Générer π_{j*}, μ_{j^*} et $\Sigma_{j^*}^{-1}$ et les assigner au composant vide que l'on vient de créer.
3. Mettre à jour les poids des composants.
4. Vérifier que le mouvement est accepté avec les nouveaux paramètres, sinon on ne change rien.

Le mouvement de mort d'un composant s'effectue en le sélectionnant au hasard parmi les composants vides. Les poids des composants restants sont alors redimensionnés pour avoir une somme égale à un. Dans le calcul de la probabilité d'acceptation, π_{j*}, μ_{j^*} et Σ_{j^*} jouent le rôle de u, ce qui permet une simplification de l'expression du taux d'acceptation. Voici le détail des opérations à effectuer pour le mouvement de mort :

1. Choisir un composant j^* parmi les composants vides.
2. Le poids π_{j^*} du composant tué est réparti entre les composants restants.
3. Diminuer le nombre de composants de 1 et assigner de nouveaux indices à ceux restants.
4. Vérifier que le mouvement est accepté avec les nouveaux paramètres, sinon on ne change rien.

Proposition 1 Le mouvement de naissance d'un composant est accepté avec la probabilité $\min(1, A)$, avec :

$$A = \frac{\Gamma(\delta k + \delta)}{\Gamma(\delta)\Gamma(\delta k)} (1 - \pi_{j*})^{n + k(\delta - 1)} \pi_{j*}^{\delta - 1} \frac{1}{\beta_{1k}(\pi_{j*})} \frac{d_{k+1}}{(k_0 + 1) b_k} (1 - \pi_{j*})^{k-1}.$$

Pour le mouvement de mort, la probabilité d'acceptation est $\min\left(1, A^{-1}\right)$, avec :

$$A = \frac{\Gamma(\delta(k-1) + \delta)}{\Gamma(\delta)\Gamma(\delta(k-1))} (1 - \pi_{j*})^{n + (k-1)(\delta - 1)} \pi_{j*}^{\delta - 1} \frac{1}{\beta_{1k-1}(\pi_{j*})} \frac{d_k}{k_0 b_{k-1}} (1 - \pi_{j*})^{k-2}.$$

On remarquera qu'alors k_0 est le nombre de composants vides après avoir effectué le mouvement de mort.

Preuve *: Nous ne donnons ici que les grandes lignes de la démonstration qui est présentée en détails dans le cas du mouvement de séparation/combinaison. On a :*

$$a_{kk+1}\left(\theta^{(k)},\theta^{(k+1)}\right)=\min\left(1,A\right),$$

avec :

$$A=\frac{p\left(k+1\right)}{p\left(k\right)}\frac{p\left(\theta^{(k+1)}\mid k+1\right)}{p\left(\theta^{(k)}\mid k\right)}\frac{p\left(y^{(n)}\mid k+1,\theta^{(k+1)}\right)}{p\left(y^{(n)}\mid k,\theta^{(k)}\right)}\frac{p_{k+1k}}{q_{kk+1}\left(\theta^{(k)},u\right)p_{kk+1}}\left|\frac{\partial g_{kk+1}\left(\theta^{(k)}\right)}{\partial\theta^{(k)}}\right|,$$

où $p\left(\theta^{(k)}\mid k\right)$ *représente la loi a priori de* $\theta^{(k)}$ *conditionnée par le nombre de composants du mélange. On rappelle que*

$$\begin{aligned}
\theta^{(k)} &= \left(\pi,z^{(n)},\beta,\mu,\Sigma\right)\quad,\quad\theta^{(k+1)}=\left(\pi',z'^{(n)},\beta',\mu',\Sigma'\right),\\
\pi' &= \left(\pi_1\left(1-\pi_{j*}\right),...,\pi_k\left(1-\pi_{j*}\right),\pi_{j*}\right),\\
z'^{(n)} &= z^{(n)},\\
\mu' &= \left(\mu_1,...,\mu_k,\mu_{j^*}\right),\\
\Sigma' &= \left(\Sigma_1,...,\Sigma_k,\Sigma_{j^*}\right),\\
\beta' &= \beta.
\end{aligned}$$

Explicitons maintenant les differents termes :

- $\frac{p(k+1)}{p(k)}$ *représente le rapport des probabilités a priori pour le nombre de composants* (*cf page 77*) ;
- $\frac{p\left(y^{(n)}|k+1,\theta^{(k+1)}\right)}{p\left(y^{(n)}|k,\theta^{(k)}\right)}$ *représente le rapport de la vraisemblance du modèle augmenté, sur celle du modèle précédent le mouvement ; il vaut* 1 *ici ;*
- $\frac{p\left(\theta^{(k+1)}|k+1\right)}{p\left(\theta^{(k)}|k\right)}=\frac{p(\beta|g,h)p\left(\pi'|k+1,\delta\right)p\left(z'^{(n)}|k+1,\pi'\right)p\left(\mu',\Sigma'|k+1,\eta\right)}{p(\beta|g,h)p(\pi|k,\delta)p\left(z^{(n)}|k,\pi\right)p(\mu,\Sigma|k,\eta)}$ *représente le rapport des probabilités a priori.*

Détaillons maintenant le dernier terme. Une simplification peut déjà se faire avec la loi a priori de l'hyperparamètre β*. On a ensuite :*

$$\begin{aligned}
\frac{p\left(\theta^{(k+1)}\mid k+1\right)}{p\left(\theta^{(k)}\mid k\right)} &= \frac{p\left(\pi'\mid k+1,\delta\right)}{p\left(\pi\mid k,\delta\right)}\frac{\prod_{i=1}^{n}P\left(Z'_i=z_i\mid k+1,\pi'\right)}{\prod_{i=1}^{n}P\left(Z_i=z_i\mid k,\pi\right)}\\
&\quad\times\frac{\prod_{j=1}^{k}\mathcal{N}_r\left(\mu_j;\xi,\kappa^{-1}\right)W_r\left(\Sigma_j^{-1};2\alpha,\left(2\beta\right)^{-1}\right)}{\prod_{j=1}^{k}\mathcal{N}_r\left(\mu_j;\xi,\kappa^{-1}\right)W_r\left(\Sigma_j^{-1};2\alpha,\left(2\beta\right)^{-1}\right)}\times\mathcal{N}_r\left(\xi,\kappa^{-1}\right)W_r\left(2\alpha,\left(2\beta\right)^{-1}\right)\\
&= \frac{\Gamma\left(\delta k+\delta\right)}{\Gamma\left(\delta\right)\Gamma\left(\delta k\right)}\left(1-\pi_{j*}\right)^{k(\delta-1)}\pi_{j*}^{\delta-1}\frac{\prod_{i=1}^{n}\pi'_{z_i}}{\prod_{i=1}^{n}\pi_{z_i}}\times\mathcal{N}_r\left(\xi,\kappa^{-1}\right)W_r\left(2\alpha,\left(2\beta\right)^{-1}\right)\\
&= \frac{\Gamma\left(\delta k+\delta\right)}{\Gamma\left(\delta\right)\Gamma\left(\delta k\right)}\left(1-\pi_{j*}\right)^{n+k(\delta-1)}\pi_{j*}^{\delta-1}\times\mathcal{N}_r\left(\xi,\kappa^{-1}\right)W_r\left(2\alpha,\left(2\beta\right)^{-1}\right).
\end{aligned}$$

Il reste à détailler $\frac{1}{q_{kk+1}\left(\theta^{(k)},u\right)}\frac{p_{k+1k}}{p_{kk+1}}\left|\frac{\partial g_{kk+1}\left(\theta^{(k)}\right)}{\partial\theta^{(k)}}\right|$. *Ici,* π_{j*},μ_{j^*} *et* Σ_{j^*} *jouent le rôle de* u. *Cette dernière expression se détaille donc sous la forme :*

$$\begin{aligned}\frac{1}{q_{kk+1}\left(\theta^{(k)},u\right)} &= \frac{1}{\beta_{1k}\left(\pi_{j*}\right)\mathcal{N}_r\left(\xi,\kappa^{-1}\right)W_r\left(2\alpha,(2\beta)^{-1}\right)},\\ \frac{p_{k+1k}}{p_{kk+1}} &= \frac{d_{k+1}}{\left(k_0+1\right)b_k},\\ \left|\frac{\partial g_{kk+1}\left(\theta^{(k)}\right)}{\partial\theta^{(k)}}\right| &= \left(1-\pi_{j*}\right)^{k-1}.\end{aligned}$$

L'expression du jacobien provient de Richardson et Green [1997] , d_{k+1} *et* b_k *sont explicités à la page 80 et* k_0 *représente le nombre de composants vides avant la naissance.*
L'expression finale de la probabilité d'acceptation reprend celle de Richardson et Green [1997]. Ceux çi obtiennent cependant un terme $(k+1)=\frac{(k+1)!}{k!}$ *qui provient de la condition d'ordre sur les moyennes que ceux çi imposent afin de conserver l'identifiabilité. Dans notre cas, cette condition n'est plus utilisée, et le terme* $K+1$ *a donc disparu. On a donc :*

$$A=\frac{\Gamma\left(\delta k+\delta\right)}{\Gamma\left(\delta\right)\Gamma\left(\delta k\right)}\left(1-\pi_{j*}\right)^{n+k(\delta-1)}\pi_{j*}^{\delta-1}\frac{1}{\beta_{1k}\left(\pi_{j*}\right)}\frac{d_{k+1}}{\left(k_0+1\right)b_k}\left(1-\pi_{j*}\right)^{k-1}.$$

■

5 Etude du mouvement de séparation / combinaison

Les deux mouvements qui le composent sont apparemment différents mais en fait ils sont l'inverse l'un de l'autre. Selon le mécanisme des sauts réversibles, ils doivent donc être définis simultanément. On choisit de manière aléatoire de combiner deux composants (c'est-à-dire réduire la taille du modèle) ou de séparer deux composants (c'est-à-dire augmenter la taille du modèle). Le choix entre ces deux possibilités se fait de la même manière que précédemment, en utilisant les probabilités b_k et d_k.

Dans cette section, on notera j_1 et j_2 les indices des composants à combiner ou obtenus par séparation, et j^* l'indice du composant à séparer ou obtenu par combinaison.

Lors de ce type de mouvement, la dimension du modèle augmente ou diminue. Il devient alors nécessaire, afin de se conformer à la théorie détaillée par Richardson et Green [1997] et explicitée au chapitre III, de définir les nouveaux paramètres du mélange par un difféomorphisme. L'optique que nous avons choisis pour ce travail, est de rester conforme à l'esprit de Richardson et Green [1997] en conservant la condition de conservation des trois premiers moments. Cette condition n'est pas obligatoire, mais il est tout de même nécessaire de conserver quelques "garde-fous" dans la génération des nouveaux paramètres. Le choix de la condition sur les moments est délibéré, mais nous devons signaler que d'autres choix semblent possibles, et font partie des extensions envisageables à ce travail.

Dans le cas multivarié, les conditions sur les moments s'écrivent de la façon suivante :

$$\begin{aligned}&\pi_{j^*}=\pi_{j_1}+\pi_{j_2}\\ &\pi_{j^*}\mu_{j^*}=\pi_{j_1}\mu_{j_1}+\pi_{j_2}\mu_{j_2}\\ &\pi_{j^*}\left(\mu_{j^*}\mu'_{j^*}+\Sigma_{j^*}\right)=\pi_{j_1}\left(\mu_{j_1}\mu'_{j_1}+\Sigma_{j_1}\right)+\pi_{j_2}\left(\mu_{j_2}\mu'_{j_2}+\Sigma_{j_2}\right).\end{aligned}\tag{V.6}$$

Comme on l'a vu précédemment, il devient nécessaire d'introduire des nouvelles variables, générées aléatoirement, correspondant au nombre de degrés de liberté manquants lorsqu'on passe

Difféomorphisme	Jacobien
$\begin{array}{c}(\pi_{j_1},\mu_{j_1},\Sigma_{j_1},\pi_{j_2},\mu_{j_2},\Sigma_{j_2})\\ \Downarrow \\ (\pi_{j^*},\mu_{j^*},\Sigma_{j^*},u1,u2,u3)\end{array}$	$J=\left\lvert\frac{\partial(\pi_{j_1},\mu_{j_1},\Sigma_{j_1},\pi_{j_2},\mu_{j_2},\Sigma_{j_2})}{\partial(\pi_{j^*},\mu_{j^*},\Sigma_{j^*},u1,u2,u3)}\right\rvert$

TAB. V.1 – Formulation explicite du jacobien

d'un espace à k composants à un espace à $k+1$ composants. En dimension deux par exemple, nous devons créer six nouvelles variables correspondant aux paramètres associés à un composant du mélange (proportion, moyenne et variance). En dimension r, cela signifie que nous devons générer $1+r+\frac{r(r+1)}{2}$ nouvelles variables car la matrice $u3$ est symétrique :

$$
\begin{aligned}
&u1 \sim \beta\,(2,2) \text{ la loi Bêta de paramètres } (2,2)\\
&u2 = \begin{bmatrix} u2_1 \\ u2_2 \\ \vdots \\ u2_r \end{bmatrix}\\
&u3 = \begin{bmatrix} u3_{1,1} & u3_{1,2} & \cdots & u3_{1,r} \\ u3_{1,2} & u3_{2,2} & & u3_{2,r} \\ \vdots & & \ddots & \vdots \\ u3_{1,r} & u3_{2,r} & \cdots & u3_{r,r} \end{bmatrix}.
\end{aligned}
$$

Le réel $u1$ est généré de manière à être plus proche de $\frac{1}{2}$ que de 0 ou 1. Nous verrons plus tard comment sont générées les valeurs contenues dans $u2$ et $u3$.

Une fois défini, ce changement de variables intervient dans la probabilité d'acceptation du mouvement par l'intermédiaire de son jacobien. Dans le tableau (V.1) nous avons introduit le jacobien théorique pour le mouvement de combinaison. Pour le mouvement de séparation, on utilise l'inverse de ce jacobien.

En multivarié, des difficultés majeures surviennent lorsqu'on essaie de généraliser le changement de variables introduit par Richardson et Green [1997]. En particulier, il s'avère très difficile de définir correctement les correspondances entre matrices de variance-covariance tout en restant dans l'espace des matrices symétriques définies positives.

5.1 Vers une première généralisation

La généralisation du changement de variables de Richardson et Green [1997] pour les proportions et les moyennes vient assez naturellement. En effet, la première équation des moments nous contraint à poser

$$
\begin{aligned}
\pi_{j_1} &= u1\pi_{j^*}\\
\pi_{j_2} &= (1-u1)\,\pi_{j^*}
\end{aligned}
\tag{V.7}
$$

où $u1$ est un nombre aléatoire compris entre 0 et 1. La seconde équation des moments peut se voir, quant à elle, comme une somme pondérée des deux vecteurs μ_{j_1} et μ_{j_2}. La seule façon d'écrire les nouvelles moyennes est donc :

$$
\begin{aligned}
\mu_{j_1} &= \mu_{j^*} + \tfrac{A}{u1}\\
\mu_{j_2} &= \mu_{j^*} - \tfrac{A}{1-u1},
\end{aligned}
\tag{V.8}
$$

où A est un vecteur quelconque. Lorsqu'on choisit $A = \sqrt{u1\,(1-u1)} \begin{bmatrix} u2_1\sqrt{\Sigma_{j^*_{1,1}}} \\ u2_2\sqrt{\Sigma_{j^*_{2,2}}} \\ \vdots \\ u2_r\sqrt{\Sigma_{j^*_{r,r}}} \end{bmatrix}$, on est ramené au changement de variables de Richardson et Green [1997]. Cela revient à poser :

$$\begin{aligned} \mu_{j_1} &= \mu_{j^*} + \sqrt{\tfrac{1-u1}{u1}} D\Sigma_{j^*} u2 \\ \mu_{j_2} &= \mu_{j^*} - \sqrt{\tfrac{u1}{1-u1}} D\Sigma_{j^*} u2 \end{aligned}$$

avec

$$D\Sigma_{j^*} = \begin{bmatrix} \sqrt{\Sigma_{j^*_{1,1}}} & 0 & \cdots & 0 \\ 0 & \sqrt{\Sigma_{j^*_{2,2}}} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{\Sigma_{j^*_{r,r}}} \end{bmatrix}.$$

La troisième équation des moments (cf V.6) semble donc propice à nous donner de sérieuses indications sur la façon de générer les nouvelles matrices de variance-covariance. En tenant compte de (V.7) et (V.8) la troisième équation des moments (cf V.6) s'écrit :

$$u1\Sigma_{j_1} + (1-u1)\,\Sigma_{j_2} = \Sigma_{j^*} - \frac{AA'}{u1\,(1-u1)}.$$

Une idée naturelle consiste à définir Σ_{j_1} et Σ_{j_2} comme des perturbations de Σ_{j^*} de la manière suivante :

$$\begin{aligned} \Sigma_{j_1} &= \Sigma_{j^*} + \tfrac{(u3-Id)}{u1^2(1-u1)} AA' \\ \Sigma_{j_2} &= \Sigma_{j^*} - \tfrac{u3}{u1(1-u1)^2} AA'. \end{aligned}$$

La troisième équation des moments est ainsi respectée mais il devient alors très difficile de déterminer les conditions nécessaires pour que Σ_{j_1} et Σ_{j_2} soient symétriques définies-positives. En effet, il semble mal adapté d'imposer des transformations linéaires aux variances plutôt que des modifications quadratiques.

Il apparaît ensuite que l'on peut imposer aux éléments diagonaux une transformation identique à celle décrite dans Richardson et Green [1997]. Elle s'écrit :

$$\begin{aligned} \Sigma_{j_{1_{(i,i)}}} &= \tfrac{1}{u1} u3_{i,i}(1-u2_i^2)\Sigma_{j^*_{(i,i)}} \\ \Sigma_{j_{2_{(i,i)}}} &= \tfrac{1}{1-u1}\,(1-u3_{i,i})\,(1-u2_i^2)\Sigma_{j^*_{(i,i)}}. \end{aligned}$$

Le problème essentiel est alors de trouver une expression des termes hors-diagonaux vérifiant l'équation des moments ainsi que la condition de définie-positivité. Afin de rester le plus simple possible, il semble alors naturel de considérer les termes hors-diagonaux de $\Sigma_{j^*} - \frac{AA'}{u1(1-u1)}$ et de les répartir entre Σ_{j_1} et Σ_{j_2}. En effet, après la somme pondérée de Σ_{j_1} et Σ_{j_2} il ne doit rester que ce terme. Nous proposons alors le changement de variables suivant (pour $i \neq j$) :

$$\begin{aligned} \Sigma_{j_{1_{(i,j)}}} &= \tfrac{u3_{i,j}}{u1}\left(\Sigma_{j^*_{(i,j)}} - u2_i u2_j \sqrt{\Sigma_{j^*_{(i,i)}}}\sqrt{\Sigma_{j^*_{(j,j)}}}\right) \\ \Sigma_{j_{2_{(i,j)}}} &= \tfrac{(1-u3_{i,j})}{1-u1}\left(\Sigma_{j^*_{(i,j)}} - u2_i u2_j \sqrt{\Sigma_{j^*_{(i,i)}}}\sqrt{\Sigma_{j^*_{(j,j)}}}\right). \end{aligned}$$

La condition de définie-positivité reste là-aussi très difficile à vérifier. Il devient donc nécessaire d'imposer des conditions aux variables $u2$ et $u3$ afin de diminuer la valeur des termes hors

diagonaux. En effet, une matrice symétrique à diagonale dominante a plus de chances d'être définie-positive. Nous avons imposé ici :

$$\begin{aligned} u3_{i,j} &\sim \mathbb{1}_{[1/3,2/3]} \\ u2_i &\sim \mathbb{1}_{[1/2,1]}. \end{aligned}$$

Ces conditions permettent de maximiser la probabilité de générer des matrices définies-positives et de n'effectuer que de légères modifications de la matrice initiale. En effet, il est naturel de penser que les nouvelles matrices de variance-covariance ne doivent pas être trop éloignées de la matrice d'origine. On obtient ainsi un changement de variables qui correspond à la fonction $g_{kk'1}$ de la section 3 page 61 :

$$\begin{aligned} \pi_{j_1} &= u1\pi_{j^*} \\ \pi_{j_2} &= (1-u1)\,\pi_{j^*} \\ \mu_{j_1} &= \mu_{j^*} + \sqrt{\tfrac{1-u1}{u1}} D\Sigma_{j^*} u2 \\ \mu_{j_2} &= \mu_{j^*} - \sqrt{\tfrac{u1}{1-u1}} D\Sigma_{j^*} u2 \\ \Sigma_{j_1{}_{(i,j)}} &= \begin{cases} \frac{1}{u1} u3_{i,i}(1-u2_i^2)\Sigma_{j^*_{(i,i)}} & \text{si } i=j \\ \frac{u3_{i,j}}{u1}\left(\Sigma_{j^*_{(i,j)}} - u2_i u2_j \sqrt{\Sigma_{j^*_{(i,i)}}}\sqrt{\Sigma_{j^*_{(j,j)}}}\right) & \text{si } i\neq j \end{cases} \\ \Sigma_{j_2{}_{(i,j)}} &= \begin{cases} \frac{1}{1-u1}\left(1-u3_{i,i}\right)\left(1-u2_i^2\right)\Sigma_{j^*_{(i,i)}} \\ \text{si } i=j, \\ \frac{(1-u3_{i,j})}{1-u1}\left(\Sigma_{j^*_{(i,j)}} - u2_i u2_j \sqrt{\Sigma_{j^*_{(i,i)}}}\sqrt{\Sigma_{j^*_{(j,j)}}}\right) \\ \text{si } i\neq j, \end{cases} . \end{aligned}$$

avec

$$D\Sigma_{j^*} = \begin{bmatrix} \sqrt{\Sigma_{j^*_{1,1}}} & 0 & \cdots & 0 \\ 0 & \sqrt{\Sigma_{j^*_{2,2}}} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{\Sigma_{j^*_{r,r}}} \end{bmatrix}.$$

L'expression du jacobien obtenue dans le cas présent à l'aide du logiciel Maple est la suivante :

$$J = -\pi_{j^*} \prod_{i=1}^{r} \left[\frac{\Sigma_{j^*_{i,i}}^{\frac{3}{2}}\left(u2_i^2-1\right)}{\left[u1\,(u1-1)\right]^{\frac{r+2}{2}}} \prod_{j=i+1}^{r} \left(u2_i u2_j \sqrt{\Sigma_{j^*_{(i,i)}}}\sqrt{\Sigma_{j^*_{(j,j)}}} - \sqrt{\Sigma_{j^*_{i,j}}} \right) \right]. \qquad \text{(V.9)}$$

Malgré les conditions imposées précédemment sur $u3$ et $u2$, il reste possible que les matrices Σ_{j_1} et Σ_{j_2} générées ne soient pas définies positives. Un changement de variable permettant de garantir cette propriété reste donc, sinon nécessaire, du moins souhaitable. Ceci nous a conduit à abandonner cette approche. Les collaborations entretenues avec l'INRIA Rhones-Alpes et notamment avec Gilles Celeux, nous ont permis de proposer une alternative.

5.2 Seconde généralisation

L'idée exploitée ici est d'utiliser la décomposition de Cholesky lors du changement de variable. Celle-ci consiste à décomposer $\Sigma_j = V_j V_j'$ où V_j est une matrice triangulaire inférieure. Une décomposition de ce type peut être trouvée dans Posse [1998], qui l'utilise dans un cadre identique au notre malgré une expression légèrement différente. Nous proposons de générer les paramètres des nouveaux composants issus du mouvement de séparation de la manière suivante :

$$\boxed{\begin{array}{ll} \pi_{j_1} = u_1\pi_{j^*} & \pi_{j_2} = (1-u_1)\,\pi_{j^*} \\ \mu_{j_1} = \mu_{j^*} + \sqrt{\frac{1-u_1}{u1}}V_{j^*}\frac{u_2}{\sqrt{2r}} & \mu_{j_2} = \mu_{j^*} - \sqrt{\frac{u_1}{1-u1}}V_{j^*}\frac{u_2}{\sqrt{2r}} \\ \Sigma_{j_1} = \frac{1}{2u1}V_{j^*}\left(I_r - \frac{u_2u_2'}{2r} + \frac{u_3}{2r}\right)V_{j^*}' & \Sigma_{j_2} = \frac{1}{2(1-u1)}V_{j^*}\left(I_r - \frac{u_2u_2'}{2r} - \frac{u_3}{2r}\right)V_{j^*}', \end{array}} \tag{V.10}$$

avec $\Sigma_{j^*} = V_{j^*}'V_{j^*}$. On peut aisément vérifier que les nouveaux paramètres générés selon ce difféomorphisme vérifient les conditions (V.6) sur les moments. Remarquons d'autre part qu'il aurait été difficile d'écrire ce changement de varibles uniquement en fonction de matrices triangulaires. L'avantage de l'utilisation de la décomposition de Cholesky est indéniable lorsqu'on souhaite vérifier la définie-positivité des matrices de variances-covariances Σ_{j_1} et Σ_{j_2}. En effet, pour tout $v = (v_1, ..., v_r)' \neq 0$, pour tout $u2$ et $u3$, nous avons bien :

$$w = v'\left(I_r - \frac{u2u2'}{2r} - \frac{u3}{2r}\right)v \geqslant 0.$$

Le cas le moins favorable apparait pour $u2_l = u3_{i,j} = 1$ et $v_l > 0$ pour tout $i, j, l = 1, ..., r$. Dans ce cas, $w = \sum_{i=1}^r v_i^2 - \frac{\left(\sum_{i=1}^r v_i\right)^2}{r} \geqslant 0$, et on a l'égalité pour $v_1 = ... = v_r$. Le difféomorphisme (V.10) nous garantit donc l'obtention de matrices Σ_{j_1} et Σ_{j_2} convenables. Les contraintes imposées précédemment sur $u1$, $u2$ et $u3$ n'ont alors plus lieu d'être. Nous choisissons donc de générer ces valeurs de la manière suivante :

$$\begin{array}{cc} u1 \sim \beta(2,2) & u3_{ij} \sim \beta(2,2) \\ \text{pour } i = 1, ..., r & \text{et } i \leqslant j \leqslant r. \end{array}$$

En ce qui concerne la génération des vecteurs $u2$, le fait de générer les coordonnées $u2_i$ selon une $\beta(2,2)$ mène à un vecteur de coordonnées toujours positives. L'ensemble des mouvements de séparation/combinaison subirait donc une contrainte très forte. Pour éviter cela et afin d'augmenter l'ampleur des mouvements autorisés, $u2_i$ est généré selon une $\beta(1.5, 2)$, puis avec une probabilité $\frac{1}{2}$ nous changeons le signe de la valeur obtenue.

Un autre avantage à travailler avec la décomposition de Cholesky est de simplifier le calcul du jacobien. En effet, obtenir des dérivées partielles par rapport à la décomposition de Cholesky est relativement aisé. On peut trouver par exemple dans Gupta et Nagar [2000] des jacobiens de multiples transformations linéaires ou quadratiques utilisant cette décomposition. Afin de calculer le jacobien du difféomorphisme (V.10) il nous faut préciser les différentes étapes du processus. Lorsqu'on effectue un mouvement de séparation on doit :

1. Choisir le composant j^* à séparer.
2. Effectuer la décomposition de Cholesky $\Sigma_{j^*} = V_{j^*}'V_{j^*}$.
3. Effectuer le difféomorphisme (V.10).

Rappelons ici que le mouvement de séparation est l'inverse du mouvement de combinaison dont les étapes sont les suivantes :

1. Choisir j_1 et j_2 les composants à combiner.
2. déterminer π_{j^*},μ_{j^*} et Σ_{j^*} à l'aide des équations des moments (V.6).

En effet, u_1, u_2 et u_3 sont ici inutiles puisque nous ne sommes pas obligés d'effectuer le difféomorphisme (V.10) en sens inverse, les équations des moments étant suffisantes. On peut cependant signaler qu'il faut tout de même déterminer u_1, u_2 et u_3 de proche en proche afin de calculer le jacobien nécessaire à l'expression de la probabilité d'acceptation du mouvement (voir plus loin).

Ces mouvements consistent donc en deux opérations combinées dont nous souhaitons calculer le jacobien. Ceci est résumé dans le tableau V.2.

Difféomorphisme	Jacobien
$(\pi_{j_1}, \mu_{j_1}, \Sigma_{j_1}, \pi_{j_2}, \mu_{j_2}, \Sigma_{j_2})$ $\Downarrow$ $(\pi_{j^*}, \mu_{j^*}, V_{j^*}, u_1, u_2, u_3)$ $\Downarrow$ $(\pi_{j^*}, \mu_{j^*}, \Sigma_{j^*}, u_1, u_2, u_3)$	$J_2 = \left\lvert \frac{\partial(\pi_{j_1}, \mu_{j_1}, \Sigma_{j_1}, \pi_{j_2}, \mu_{j_2}, \Sigma_{j_2})}{\partial(\pi_{j^*}, \mu_{j^*}, V_{j^*}, u_1, u_2, u_3)} \right\rvert$ $J_3 = \left\lvert \frac{\partial V_{j^*}}{\partial \Sigma_{j^*}} \right\rvert$

TAB. V.2 – jacobien pour le difféomorphismeV.10

Le jacobien J_3 se calcule aisément. On obtient :

$$J_3 = \left(2^r \prod_{k=1}^{r} V_{j^*,kk}^{r+1-k} \right)^{-1}.$$

Le jacobien J_2 nécessite l'emploi d'un logiciel de calcul formel. Afin de vérifier et de valider le calcul du jacobien avec certitude, nous l'avons implémenté avec les deux principaux outils de calcul formel : Maple et Mathematica. En dimension $r = 2$, nous devons calculer le déterminant d'une matrice 12×12 et pour $r = 3$ d'une matrice 20×20. C'est une opération très longue et complexe que ces deux logiciels gèrent de manière complètement différente. En effet, comme nous avons pu le constater, Maple nécessite un temps de calcul beaucoup plus long que Mathematica (15 min avec $r = 3$ contre 2 min pour Mathematica). Le code nécessaire au calcul est lui aussi beaucoup plus long et complexe. Le jacobien obtenu (identique pour les deux logiciels) a été calculé pour $r = 2, 3$ et 4 (uniquement Mathematica pour $r = 4$). Les résultats obtenus pour $r = 2, 3$ ou 4 nous permettent de conjecturer la forme suivante du jacobien J_2 pour r entier positif quelconque.

$$J_2 = \frac{-\pi_{j^*} \prod_{k=1}^{r} V_{j^*,kk}^{2r+3-k}}{2^{(r^2+r)/2} \left(r u_1 (1-u_1) \right)^{r(r+2)/2}}.$$

Le jacobien de la transformation complète est donc le produit de ces deux jacobiens. Nous avons :

$$J = J_2 J_3 = \frac{-\pi_{j^*} \prod_{k=1}^{r} V_{j^*,kk}^{r+2}}{2^{r(r-3)/2} \left(r u_1 (1-u_1) \right)^{r(r+2)/2}}. \tag{V.11}$$

Finalement, lorsque $r = 1$, le changement de variable (V.10) est très proche de celui de Green et Richardson. Il se simplifie de la manière suivante :

$$\begin{array}{ll} \mu_{j_1} = \mu_{j^*} + \sqrt{\frac{1-u_1}{u1}} \sigma_{j^*} \frac{u_2}{\sqrt{2}} & \mu_{j_2} = \mu_{j^*} - \sqrt{\frac{u_1}{1-u1}} \sigma_{j^*} \frac{u_2}{\sqrt{2}} \\ \Sigma_{j_1} = \frac{1}{2u1} \sigma_{j^*} \left(1 - \frac{u_2^2}{2} + u_3 \right) & \Sigma_{j_2} = \frac{1}{2(1-u1)} \sigma_{j^*} \left(1 - \frac{u_2^2}{2} - \frac{u_3}{2} \right). \end{array}$$

De plus, dans le cas $r = 1$ nous avons $J_3 = \frac{1}{2\sigma_{j^*}}$ et $J_2 = \frac{-\pi_{j^*} \sigma_{j^*}^4}{2 u_1^{3/2} (1-u_1)^{3/2}}$. Le jacobien final est donc :

$$J = J_2 J_3 = \frac{-\sigma_{j^*}^3 \pi_{j^*}}{4 u_1^{3/2} (1-u_1)^{3/2}},$$

que l'on peut comparer avec le jacobien de Richardson et Green :

$$J = \frac{\sigma_{j^*}^3 \pi_{j^*} \left(1 - u_2^2 \right)}{u_1^{3/2} (1-u_1)^{3/2}}.$$

Nous pouvons aussi remarquer que le signe du jacobien importe peu puisque nous en prenons la valeur absolue.

5.3 Mouvement séparation

Pour ce mouvement, le composant j^* à décomposer en j_1 et j_2 est choisi au hasard. On passe ici d'un modèle de dimension k, à un modèle de dimension $k+1$, on est ensuite ramené au cas présenté dans le paragraphe 3-4. Les équations des moments doivent toujours être vérifiées, cependant la dimension de l'espace d'arrivée étant supérieure à celle de départ, il nous faut générer aléatoirement les variables correspondant aux degrés de liberté supplémentaires.

$$
\begin{aligned}
u1 &\sim \beta(2,2) \text{ la loi Beta de paramètres } (2,2)\\
u2 &= \begin{bmatrix} u2_1 \\ u2_2 \\ \vdots \\ u2_r \end{bmatrix}\\
u3 &= \begin{bmatrix} u3_{1,1} & u3_{1,2} & \cdots & u3_{1,r} \\ u3_{1,2} & u3_{2,2} & & u3_{2,r} \\ \vdots & & \ddots & \vdots \\ u3_{1,r} & u3_{2,r} & \cdots & u3_{r,r} \end{bmatrix}.
\end{aligned}
$$

Les éléments de $u2$ sont générés comme précédemment, et ceux de $u3$ selon une loi $\beta(2,2)$. Le détail des opérations précédentes est résumé par :

1. Choisir au hasard le composant j^* à séparer
2. Effectuer la décomposition de Cholesky $\Sigma_{j^*} = V'_{j^*} V_{j^*}$
3. Déterminer $\pi_{j_1}, \mu_{j_1}, \Sigma_{j_1}, \pi_{j_2}, \mu_{j_2}, \Sigma_{j_2}$ selon (V.10)
4. Mettre à jour les allocations
5. Vérifier que le mouvement est bien accepté avec ces nouveaux paramètres, sinon on ne change rien.

Lorsque le mouvement est accepté, on passe de k à $k+1$ composants. Le composant j_1 remplace le composant j^* et le composant j_2 est indicé par $k+1$.

5.4 Mouvement combinaison

On doit, dans un premier temps, décider quels sont les deux composants à combiner. La paire d'indices (j_1, j_2) est alors choisie avec une probabilité relative à la distance les séparant, c'est-à-dire que l'on utilise :

$$
P_{sélection}(j_1, j_2) = 1 - \frac{d(j_1, j_2)}{\sum\limits_{m,n \in \{1,\ldots,k\}} d(m,n)},
$$

où $d(j_1, j_2)$ représente la distance de Mahalanobis entre les composants j_1 et j_2 définie comme suit :

$$
d(j_1, j_2) = \left(\mu_{j_1} - \mu_{j_2}\right) \Sigma_{j_1}^{-1} \left(\mu_{j_1} - \mu_{j_2}\right) + \left(\mu_{j_2} - \mu_{j_1}\right) \Sigma_{j_2}^{-1} \left(\mu_{j_2} - \mu_{j_1}\right).
$$

Remarque 7 *Richardson et Green [1997] choisissent les composants de manière à ce que l'on ait $\mu_{j_1} < \mu_{j_2}$ sans aucun μ_j entre μ_{j_1} et μ_{j_2}. En fait, cela représente simplement le fait qu'ils imposent aux composants d'êtres adjacents au sens de la condition de séparation qu'ils ont considérée, c'est-à-dire : $\mu_1 < \mu_2 < \ldots < \mu_k$. Cette condition est inutilisable dans le cas multivarié.*

Comme il est suggéré par Stephens [2000b] et Celeux et al. [2000], on n'utilisera pas de condition de séparation, mais plutôt une post-gestion du problème de label switching.

Les deux composants choisis sont donc combinés en un seul que l'on nomme j^*. La complexité du modèle, représentée par le nombre de composants, passe donc de k à $k-1$. Il est alors nécessaire de modifier les paramètres correspondants aux deux composants combinés. Tous les z_i tels que $z_i = j_1$ ou $z_i = j_2$ sont mis à la valeur j^*. Les nouvelles valeurs de la moyenne, de la variance et des proportions sont mises à jour en respectant les équations des trois premiers moments : Ces mêmes équations serviront aussi pour le mouvement inverse de séparation. Tout ceci est résumé de la manière suivante :

1. Choisir les composants à combiner j_1 et j_2.
2. Déterminer les paramètres du composant j^* à l'aide des équations des moments.
3. Mettre à jour les allocations.
4. Vérifier que le mouvement est bien accepté avec ces nouveaux paramètres, sinon on ne change rien.

5.5 Probabilité d'acceptation

Proposition 2 La probabilité d'acceptation du mouvement de séparation passant d'un modèle à k composants, à un modèle à k+1 composants s'écrit alors :

$$a_{kk+1}\left(\theta^{(k)}, \theta^{(k+1)}\right) = \min\left(1, A\right),$$

avec :

$$\begin{aligned}
A \;=\; & \frac{p\left(k+1 \mid \lambda\right)}{p\left(k \mid \lambda\right)} \prod_{i/z_i'=j_1}^{n} \sqrt{\frac{\det \Sigma_{j^*}}{\det \Sigma_{j_1}}} \exp\left\{-\frac{1}{2}\left[\left(y_i-\mu_{j_1}\right)' \Sigma_{j_1}^{-1}\left(y_i-\mu_{j_1}\right) - \left(y_i-\mu_{j^*}\right)' \Sigma_{j^*}^{-1}\left(y_i-\mu_{j^*}\right)\right]\right\} \\
& \times \prod_{i/z_i'=j_2}^{n} \sqrt{\frac{\det \Sigma_{j^*}}{\det \Sigma_{j_2}}} \exp\left\{-\frac{1}{2}\left[\left(y_i-\mu_{j_2}\right)' \Sigma_{j_2}^{-1}\left(y_i-\mu_{j_2}\right) - \left(y_i-\mu_{j^*}\right)' \Sigma_{j^*}^{-1}\left(y_i-\mu_{j^*}\right)\right]\right\} \\
& \times \frac{\Gamma\left(\delta k+\delta\right)}{\Gamma\left(\delta\right)\Gamma\left(\delta k\right)} \frac{\pi_{j_1}^{n_{j_1}+\delta-1} \pi_{j_2}^{n_{j_2}+\delta-1}}{\pi_{j^*}^{n_{j^*}+\delta-1}} \\
& \times \frac{\mathcal{N}_r\left(\mu_{j_1}; \xi, \kappa^{-1}\right) W_r\left(\Sigma_{j_1}^{-1}; 2\alpha, (2\beta)^{-1}\right) \mathcal{N}_r\left(\mu_{j_2}; \xi, \kappa^{-1}\right) W_r\left(\Sigma_{j_2}^{-1}; 2\alpha, (2\beta)^{-1}\right)}{\mathcal{N}_r\left(\mu_{j^*}; \xi, \kappa^{-1}\right) W_r\left(\Sigma_{j^*}^{-1}; 2\alpha, (2\beta)^{-1}\right)} \\
& \times \left(\beta_{22}\left(u1\right) \prod_{j=1}^{r} \beta_{41}\left(u2_j\right) \prod_{i=1}^{r} \prod_{j=i}^{r} \beta_{21}\left(u3_{i,j}\right)\right)^{-1} \frac{d_{k+1}}{b_k P_{alloc}} \\
& \times \frac{\pi_{j^*} \prod_{k=1}^{r} V_{j^*,kk}^{r+2}}{2^{r/2}\left(r u_1\left(1-u_1\right)\right)^{r(r+2)2}}.
\end{aligned}$$

Pour le mouvement inverse de combinaison, la probabilité d'acceptation s'écrit

$$a_{k+1k}\left(\theta^{(k+1)}, \theta^{(k)}\right) = \min\left(1, A^{-1}\right).$$

Preuve *: On a vu au chapitre III que la probabilité d'acceptation pour le mouvement de naissance s'écrivait :*

$$a_{kk+1}\left(\theta^{(k)}, \theta^{(k+1)}\right) = \min\left(1, A\right),$$

avec :

$$A = \frac{p_{k+1} f_{k+1}\left(\theta^{(k+1)}\right)}{p_k f_k\left(\theta^{(k)}\right) q_{kk+1}\left(\theta^{(k)}, u\right)} \frac{p_{k+1k}}{p_{kk+1}} \left|\frac{\partial g_{kk+1}\left(\theta^{(k)}\right)}{\partial \theta^{(k)}}\right|.$$

On rappelle qu'avec les notations de ce même chapitre (III.5) on a :

$$\begin{aligned} p_k &= \mathbb{P}\left[K = k \mid y^{(n)}\right], \\ f_k\left(\theta^{(k)}\right) &= f\left(\theta^{(k)} \mid K = k, y^{(n)}\right). \end{aligned}$$

Or, en reprenant (III.2) et (III.3) on peut écrire :

$$\frac{p_{k+1} f_{k+1}\left(\theta^{(k+1)}\right)}{p_k f_k\left(\theta^{(k)}\right)} = \frac{p\left(k+1, \theta^{(k+1)} \mid y^{(n)}\right)}{p\left(k, \theta^{(k)} \mid y^{(n)}\right)} = \frac{p\left(k+1, \theta^{(k+1)}, y^{(n)}\right)}{p\left(k, \theta^{(k)}, y^{(n)}\right)}.$$

Le détail de cette expression s'obtient en considérant les équations (V.1), (V.3) et (V.4). On a donc :

$$\frac{p_{k+1} f_{k+1}\left(\theta^{(k+1)}\right)}{p_k f_k\left(\theta^{(k)}\right)} = \frac{p\left(k+1\right)}{p\left(k\right)} \frac{p\left(\theta^{(k+1)} \mid k+1\right)}{p\left(\theta^{(k)} \mid k\right)} \frac{p\left(y^{(n)} \mid k+1, \theta^{(k+1)}\right)}{p\left(y^{(n)} \mid k, \theta^{(k)}\right)},$$

où $p\left(\theta^{(k)} \mid k\right)$ représente la loi a priori de $\theta^{(k)}$ conditionnée par le nombre de composants du mélange.

Explicitons maintenant les différents termes.

- $\frac{p(k+1)}{p(k)}$ *représente le rapport des probabilités a priori pour le nombre de composants* (*cf page 77*) ;
- $\frac{p\left(\theta^{(k+1)}|k+1\right)}{p\left(\theta^{(k)}|k\right)} = \frac{p(\beta|g,h)p\left(\pi'|k+1,\delta\right)p\left(z'^{(n)}|k+1,\pi'\right)p\left(\mu',\Sigma'|k+1,\eta\right)}{p(\beta|g,h)p(\pi|k,\delta)p\left(z^{(n)}|k,\pi\right)p(\mu,\Sigma|k,\eta)}$ *représente le rapport des probabilités a priori ;*
- $\frac{p\left(y^{(n)}|k+1,\theta^{(k+1)}\right)}{p\left(y^{(n)}|k,\theta^{(k)}\right)}$ *représente le rapport de la vraisemblance du modèle augmenté, sur celle du modèle précédent le mouvement.*

Détaillons maintenant le dernier terme. On rappelle tout d'abord que

$$\begin{aligned} \theta^{(k)} &= \left(\pi, z^{(n)}, \beta, \mu, \Sigma\right) \quad et \quad \theta^{(k+1)} = \left(\pi', z'^{(n)}, \beta', \mu', \Sigma'\right), \\ \pi &= (\pi_1, ..., \pi_{j*}, ..., \pi_k) \quad et \quad \pi' = (\pi_1, ..., \pi_{j_1}, \pi_{j_2}, ..., \pi_k), \\ z^{(n)} &= (z_1, ..., z_{j*}, ..., z_k) \quad et \quad z'^{(n)} = (z_1, ..., z_{j_1}, z_{j_2}, ..., z_k), \\ \mu &= \left(\mu_1, ..., \mu_{j*}, ..., \mu_k\right) \quad et \quad \mu' = \left(\mu_1, ..., \mu_{j_1}, \mu_{j_2}, ..., \mu_k\right), \\ \Sigma &= (\Sigma_1, ..., \Sigma_{j*}, ..., \Sigma_k) \quad et \quad \Sigma' = (\Sigma_1, ..., \Sigma_{j_1}, \Sigma_{j_2}, ..., \Sigma_k), \\ \beta' &= \beta. \end{aligned}$$

On a :

$$\frac{p\left(y^{(n)} \mid k+1, \theta^{(k+1)}\right)}{p\left(y^{(n)} \mid k, \theta^{(k)}\right)} = \frac{\prod_{i=1}^{n} \sum_{j=1}^{k+1} \mathcal{N}_r\left(y_i; \mu'_j, \Sigma'_j\right) \mathbb{1}_{[z'_i=j]}}{\prod_{i=1}^{n} \sum_{j=1}^{k} \mathcal{N}_r\left(y_i; \mu_j, \Sigma_j\right) \mathbb{1}_{[z_i=j]}}.$$

Or ici, seuls les z_i tels que $z_i = j^*$ ont changé de valeur. De même, seuls les paramètres du composant j^* ont évolué. Le rapport des vraisemblances peut donc s'écrire de manière plus simple :

$$\begin{aligned}
&\frac{p\left(y^{(n)} \mid k+1, \theta^{(k+1)}\right)}{p\left(y^{(n)} \mid k, \theta^{(k)}\right)} \\
= &\frac{\prod\limits_{i/z_i'=j_1}^{n} \mathcal{N}_r\left(y_i; \mu_{j_1}, \Sigma_{j_1}\right) \prod\limits_{i/z_i'=j_2}^{n} \mathcal{N}_r\left(y_i; \mu_{j_2}, \Sigma_{j_2}\right)}{\prod\limits_{i/z_i=j^*}^{n} \mathcal{N}_r\left(y_i; \mu_{j^*}, \Sigma_{j^*}\right)}, \\
= &\prod_{i/z_i'=j_1}^{n} \sqrt{\frac{\det \Sigma_{j^*}}{\det \Sigma_{j_1}}} \exp\left\{-\frac{1}{2}\left[\left(y_i-\mu_{j_1}\right)' \Sigma_{j_1}^{-1}\left(y_i-\mu_{j_1}\right) - \left(y_i-\mu_{j^*}\right)' \Sigma_{j^*}^{-1}\left(y_i-\mu_{j^*}\right)\right]\right\} \\
&\times \prod_{i/z_i'=j_2}^{n} \sqrt{\frac{\det \Sigma_{j^*}}{\det \Sigma_{j_2}}} \exp\left\{-\frac{1}{2}\left[\left(y_i-\mu_{j_2}\right)' \Sigma_{j_2}^{-1}\left(y_i-\mu_{j_2}\right) - \left(y_i-\mu_{j^*}\right)' \Sigma_{j^*}^{-1}\left(y_i-\mu_{j^*}\right)\right]\right\}.
\end{aligned}$$

On obtient ainsi :

$$\begin{aligned}
&Log\left[\frac{p\left(y^{(n)} \mid k+1, \theta^{(k+1)}\right)}{p\left(y^{(n)} \mid k, \theta^{(k)}\right)}\right] \\
= &\frac{n_{j_1}}{2} \log\left(\frac{\det \Sigma_{j^*}}{\det \Sigma_{j_1}}\right) - \frac{1}{2} \sum_{i/z_i'=j_1}^{n} \left[\left(y_i-\mu_{j_1}\right)' \Sigma_{j_1}^{-1}\left(y_i-\mu_{j_1}\right) - \left(y_i-\mu_{j^*}\right)' \Sigma_{j^*}^{-1}\left(y_i-\mu_{j^*}\right)\right] \\
&+\frac{n_{j_2}}{2} \log\left(\frac{\det \Sigma_{j^*}}{\det \Sigma_{j_2}}\right) - \frac{1}{2} \sum_{i/z_i'=j_2}^{n} \left[\left(y_i-\mu_{j_2}\right)' \Sigma_{j_2}^{-1}\left(y_i-\mu_{j_2}\right) - \left(y_i-\mu_{j^*}\right)' \Sigma_{j^*}^{-1}\left(y_i-\mu_{j^*}\right)\right].
\end{aligned}$$

Nous allons maintenant détailler les différents termes intervenant dans le rapport des probabilités a priori. Le premier consiste en un rapport de lois de Dirichlet.

$$\begin{aligned}
\frac{p\left(\pi' \mid k+1, \delta\right)}{p\left(\pi \mid k, \delta\right)} &= \frac{\Gamma\left(\delta k+\delta\right)}{\Gamma\left(\delta\right)^{k+1}} \frac{\pi_1^{\delta-1} ... \pi_{j_1}^{\delta-1} \pi_{j_2}^{\delta-1} ... \pi_k^{\delta-1}}{\pi_1^{\delta-1} ... \pi_{j^*}^{\delta-1} ... \pi_k^{\delta-1}} \times \frac{\Gamma\left(\delta\right)^k}{\Gamma\left(\delta k\right)} \\
&= \frac{\Gamma\left(\delta k+\delta\right)}{\Gamma\left(\delta\right)\Gamma\left(\delta k\right)} \frac{\pi_{j_1}^{\delta-1} \pi_{j_2}^{\delta-1}}{\pi_{j^*}^{\delta-1}}.
\end{aligned}$$

Le second terme correspond à la mise à jour des allocations.

$$\begin{aligned}
\frac{p\left(z'^{(n)} \mid k+1, \pi'\right)}{p\left(z^{(n)} \mid k, \pi\right)} &= \frac{\prod_{i=1}^{n} P\left(Z_i' = z_i \mid k+1, \pi'\right)}{\prod_{i=1}^{n} P\left(Z_i = z_i \mid k, \pi\right)} \\
&= \frac{\prod_{i=1}^{n} \pi'_{z_i}}{\prod_{i=1}^{n} \pi_{z_i}} = \frac{\prod_{i=1}^{n} \pi_{z_{j_1}} \prod_{i=1}^{n} \pi_{z_{j_2}}}{\prod_{i=1}^{n} \pi_{z_{j^*}}} \\
&= \frac{\pi_{j_1}^{n_{j_1}} \pi_{j_2}^{n_{j_2}}}{\pi_{j^*}^{n_{j^*}}}.
\end{aligned}$$

En procédant de manière identique on a :

$$\frac{p\left(\mu', \Sigma' \mid k+1, \eta\right)}{p\left(\mu, \Sigma \mid k, \eta\right)} = \frac{\mathcal{N}_r\left(\mu_{j_1}; \xi, \kappa^{-1}\right) W_r\left(\Sigma_{j_1}^{-1}; 2\alpha, (2\beta)^{-1}\right) \mathcal{N}_r\left(\mu_{j_2}; \xi, \kappa^{-1}\right) W_r\left(\Sigma_{j_2}^{-1}; 2\alpha, (2\beta)^{-1}\right)}{\mathcal{N}_r\left(\mu_{j^*}; \xi, \kappa^{-1}\right) W_r\left(\Sigma_{j^*}^{-1}; 2\alpha, (2\beta)^{-1}\right)}.$$

Il reste à détailler $\frac{1}{q_{kk+1}\left(\theta^{(k)},u\right)}\frac{p_{k+1k}}{p_{kk+1}}\left|\frac{\partial g_{kk+1}\left(\theta^{(k)}\right)}{\partial\theta^{(k)}}\right|$.

$$\frac{1}{q_{kk+1}\left(\theta^{(k)},u\right)}=\left(\beta_{22}\left(u1\right)\prod_{j=1}^{r}\beta_{41}\left(u2_j\right)\prod_{i=1}^{r}\prod_{j=i}^{r}\beta_{21}\left(u3_{i,j}\right)\right)^{-1}.$$

De plus on obtient

$$\frac{p_{k+1k}}{p_{kk+1}}=\frac{d_{k+1}}{b_k P_{alloc}},$$

où d_{k+1} *et* b_k *sont explicités à la page 80 et* P_{alloc} *représente la probabilité d'obtenir les allocations* z'*. En effet, les données du composant* j^* *sont allouées à* j_1 *et* j_2 *de manière aléatoire selon la loi a posteriori. On a donc :*

$$P_{alloc}=\prod_{i/z'_i=j_1}^{n}\left(\frac{P_{1i}}{P_{1i}+P_{2i}}\right)\prod_{i/z'_i=j_2}^{n}\left(\frac{P_{2i}}{P_{1i}+P_{2i}}\right),$$

avec

$$\begin{aligned}P_{1i} &= \frac{\pi_{j_1}}{\sqrt{\det\Sigma_{j_1}}}\exp\left[-\frac{1}{2}\left(y_i-\mu_{j_1}\right)'\Sigma_{j_1}^{-1}\left(y_i-\mu_{j_1}\right)\right],\\ P_{2i} &= \frac{\pi_{j_2}}{\sqrt{\det\Sigma_{j_2}}}\exp\left[-\frac{1}{2}\left(y_i-\mu_{j_2}\right)'\Sigma_{j_2}^{-1}\left(y_i-\mu_{j_2}\right)\right].\end{aligned}$$

L'expression de $\left|\frac{\partial g_{kk+1}\left(\theta^{(k)}\right)}{\partial\theta^{(k)}}\right|$ *est présentée en page 87. L'expression définitive de la probabiliité d'acceptation du mouvement de séparation/combinaison s'écrit donc de la manière attendue. L'expression finale de la probabilité d'acceptation reprend celle de Richardson et Green [1997], exepté le terme* $(k+1)$ *provenant de la condition d'ordre sur les moyennes.*

■

Nous concluerons cette section par une remarque. Le résultat de Hartigan [1985] nous dit que le rapport des vraisemblances tend vers $+\infty$ lorsque n tend vers $+\infty$. Ceci a une influence sur la probabilité d'acceptation d'un tel type de mouvement. On peut donc s'attendre à ce que des composants contenant un grand nombre de données soient peu combinés mais plutot séparés.

Chapitre VI

Illustrations et résultats

1 Introduction

Nous présentons dans ce chapitre un panel de résultats concernant la mise en oeuvre pratique de l'algorithme à sauts réversibles présenté au chapitre précédent. L'étude de jeux de données multivariées présente l'inconvenient majeur de demander beaucoup de temps de calcul. L'aspect évolutif du nombre de paramètres à étudier nous contraint de plus à utiliser un langage de programmation relativement simple, disposant aussi de bonnes bibliothèques adaptées aux méthodes MCMC.

Nous avons choisi d'utiliser le logiciel Matlab pour nos simulations, ceci malgré sa relative "lenteur". En effet, nous avons estimé qu'elle était compensée par la facilité de programmation ainsi que par l'aspect graphique de ce logiciel. Peter Green avait choisi une approche en Fortran 77, ainsi que Matthew Stephens. Cela leur permettait d'effectuer des simulations beaucoup plus rapidement que nous, mais sans posséder cet aspect convivial qui permet un accès plus facile aux programmes. En effet, nos programmes sont aisément utilisables par d'autres personnes et seront très prochainement disponibles sur le web. Signalons aussi que nous avons fait un usage abondant de Martinez et Martinez [2002] en utilisant notamment un programme nommé "csdfplot" spécialement adapté au tracé des mélanges gaussiens.

Nous avons effectué toutes nos simulations sur un PC 1.6 Ghrtz. Les temps de calculs des programmes utilisés pour ce chapitre peuvent varier de 30 minutes à 6 heures ; les plus longs étant évidemment ceux où l'a priori sur le nombre de composants est élevé.

Il existe relativement peu de jeux de données multivariées classiques provenant de mélanges gaussiens. Afin de faciliter les comparaisons, nous avons utilisé quelques données déjà étudiées par Stephens [2000a], notamment les données "Geyser". De plus, afin de vérifier la qualité d'estimation obtenue par l'algorithme, nous avons réutilisé l'échantillon de test "échantillon 1" présenté de façon détaillée à la section 4.2 du chapitre II. L'aspect sur lequel nous mettons l'accent est l'étude de l'impact de l'a priori du nombre de composants sur l'estimation finale. Le nombre de composants a posteriori aura donc une importance primordiale.

Nous utilisons plusieurs graphiques afin de visualiser le fonctionnement de l'algorithme. Comme au chapitre II, nous regardons classiquement l'évolution de certains paramètres au cours du temps, en particulier les variations du nombre de composants. De ces graphiques, nous tirons une estimation des diverses densités a posteriori en utilisant une méthode par noyaux gaussiens. Après avoir appliqué les algorithmes du chapitre IV pour enlever le label switching, nous sommes en mesure d'utiliser les estimations fournies par le mode a posteriori de ces lois. L'estimation des proportions est obtenue par la moyenne a posteriori, celle-ci donnant des estimations plus adaptées à notre contexte. Une fois obtenus les paramètres estimés, nous pouvons visualiser les mélanges bivariés de deux façons. La première est fournie par la routine csdfplot abordée plus haut, la seconde par l'estimation à noyaux. Csdfplot fournit la moyenne et le poids de chaque

composant ainsi qu'une ellipse de confiance pour chacun d'eux.

Nous sommes alors en mesure d'étudier l'algorithme de manière globale tout en précisant ses capacités d'exploration de l'espace d'état (mixing). Nous pourrons en particulier constater que la convergence paraît beaucoup plus difficile à atteindre dans le cas multivarié que dans le cas univarié.

Nous commençons par analyser le comportement de l'algorithme sur des données générées aléatoirement. Nous avons utilisé des échantillons bivariés de taille 200. Dans un premier temps nous reprenons l'échantillon ayant servi à illustrer les algorithmes de Gibbs et de Metropolis Hastings (cf figures II.5 et II.6). Un des composants est ici relativement éloigné et de faible proportion. Cet échantillon nous sert à décrire les capacités d'exploration et de détection. Pour le second échantillon analysé, nous revenons au problème du label switching, en analysant comment 3 composants de proportions égales sont détectés.

Avant de poursuivre, nous pouvons signaler que le fait d'utiliser une loi a priori uniforme sur le nombre de composants fait converger l'algorithme vers des modèles de dimension élevée, en atteignant rapidement le nombre de composants maximal fixé à 30. On peut donc considérer que l'a priori défini sur le nombre de composants peut difficilement être non informatif. Un modèle alternatif de lois a priori pour le nombre de composants, nommé "variable kappa" à été étudié par Stephens [1997]. Ce dernier consiste à ne plus considérer les hyperparamètres sur les variances comme fixés, mais à les munir de lois a priori non informatives. Le modèle hiérarchique que nous avons utilisé dans cette thèse a fait l'objet d'une comparaison avec le modèle "variable kappa" dans Stephens [1997]. Les deux approches se comportent différemment quant à l'estimation du nombre de composants. Le modèle "variable kappa" a notamment tendance à surestimer le nombre de composants. Pour des raisons pratiques liées au temps et à la programmation, nous avons choisis de n'étudier ici que le modèle utilisé par Richardson et Green [1997].

Avant de poursuivre notre étude, nous pouvons signaler que nous l'avons volontairement restreinte à certains aspects. Nous considérons essentiellement la loi a posteriori du nombre de composants, le mélange final obtenu par notre procédure, et les lois a posteriori sur les moyennes permettant d'évaluer l'effet de "label switching". Nous aborderons aussi de manière non-exhaustive la convergence de l'algorithme. Nous avons choisi ici de ne pas aborder certains aspects comme l'évolution du paramètre β ainsi que les lois a posteriori sur les proportions. Des simulations plus approfondies sont nécessaires afin d'évaluer avec précision les capacités de l'algorithme.

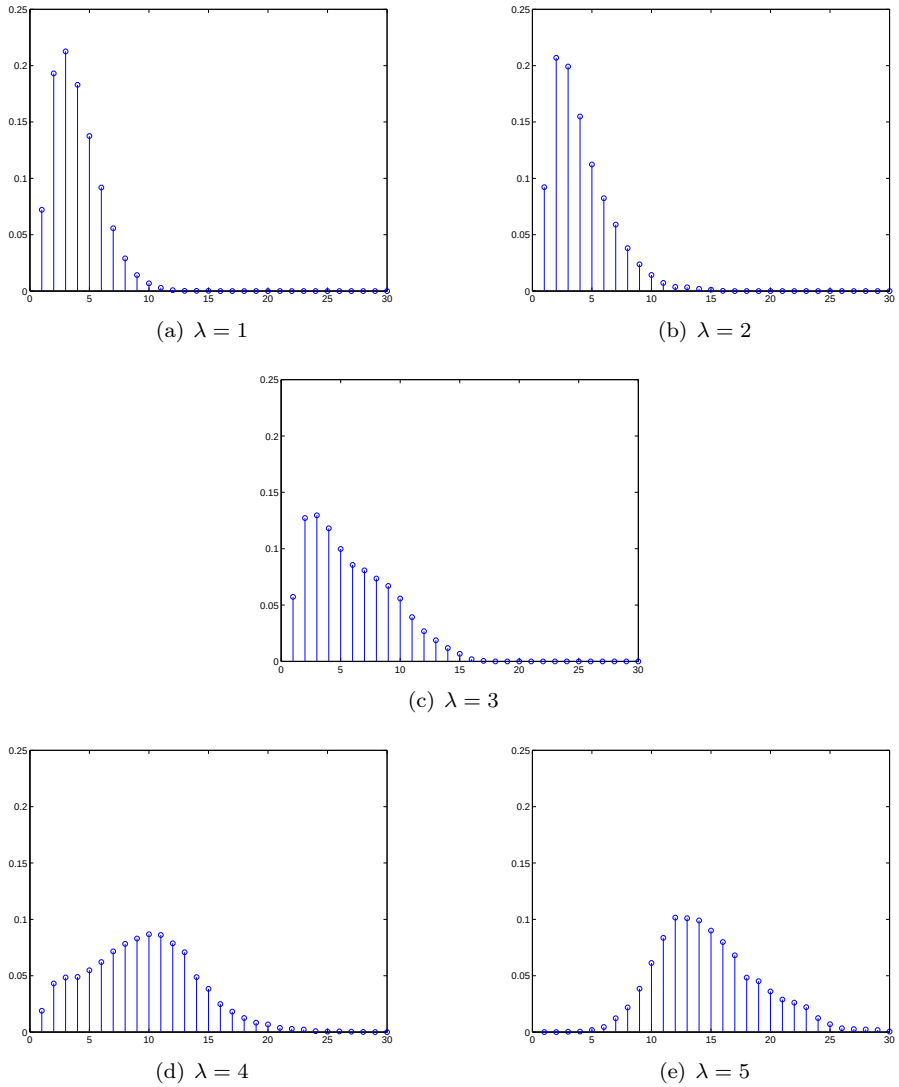
2 Echantillon simulé à 3 composants dont un de faible proportion

Il s'agit de l'échantillon généré comme indiqué page 49 et nommé "échantillon 1". Il comporte 200 données issues d'un mélange à trois composants. Le composant de moyenne $\begin{bmatrix} 4 \\ 4 \end{bmatrix}$ est celui possédant la plus faible proportion (0.1).

Nous utilisons comme loi a priori sur le nombre de composants la loi de Poisson tronquée qui est donnée par :

$$\text{pour } k \geqslant 1, \quad \mathbb{P}(K = k) = \frac{1}{e^{\lambda} - 1} \frac{\lambda^k}{k!}.$$

A la figure VI.1, on peut visualiser la loi a posteriori du nombre de composants. Nous avons effectué 25000 itérations dans chaque cas, en éliminant les 5000 premières comme temps de chauffe. Si cela semble suffire pour $\lambda = 1, 2$ ou 3, pour lesquels le nombre de composants trouvé est satisfaisant, cela semble en revanche nettement moins suffisant pour des valeurs élevées de λ. Nous en concluons que la convergence est plutôt lente dans ce cas, d'autant que le nombre important de paramètres à simuler pour chaque boucle ralentit considérablement le processus. Apparemment, comme on peut le constater sur les figures VI.2 l'algorithme semble se déplacer très facilement sur l'espace d'état puisque presque 80% des changements de dimension sont acceptés. Comme



(a) $\lambda = 1$ (b) $\lambda = 2$

(c) $\lambda = 3$

(d) $\lambda = 4$ (e) $\lambda = 5$

FIG. VI.1 – Lois a posteriori du nombre de composants pour les données "échantillon 1", et pour plusieurs valeurs du paramètre λ.

plusieurs changements de dimension sont possibles (mouvements de naissance/mort et séparation/combinaison) pour chaque itération, nous sommes presque systématiquement en présence d'un changement de dimension. Au vu de ce phénomène, notre impression est qu'autoriser l'algorithme à changer de dimension plusieurs fois lors d'une même itération, peut conduire à des biais pour les lois a posteriori. En effet, il arrive très souvent pendant les simulations, que la mise à jour de la totalité des paramètres ne soit pas effectuée entre deux changements de dimension. L'exploration de l'espace d'état par l'algorithme à sauts réversibles est donc sujet à des changements de dimension systématiques. Afin de garantir la mise à jour de tous les paramètres après chaque changement de modèle, nous avons décidé d'effectuer 5 itérations de Gibbs à vide (une itération s'entendant comme la mise à jour successive de tous les paramètres selon les lois conditionnelles a posteriori). C'est-à-dire que nous ne retenons que la dernière valeur générée avant de poursuivre le cours normal de l'algorithme. Un autre choix possible consiste à n'autoriser qu'un seul changement de dimension par itération, et donc à n'effectuer qu'un seul des deux mouvements possibles (séparation/combinaison ou naissance/mort). Nous envisageons de tester

nombre de composants	2	3	4	5	6	7	8	9	10
AIC	1853	1611	1619	1606	1624	1631	1633	1640	1621
BIC	1871	1638	1654	1651	1678	1694	1705	1720	1710

TAB. VI.1 – Critères AIC et BIC pour les données "échantillon 1" et pour différents nombres de composants.

cette solution pour la suite de nos travaux.

Une autre constatation est que l'algorithme effectue d'assez longues excursions vers des espaces (ou modèles) de dimension élevée, restant parfois plus de 500 itérations dans des modèles à plus de 10 composants. Cette dernière remarque nous permet de mettre l'accent sur le très grand nombre d'itérations nécessaires à une bonne exploration de l'espace d'état. Lorsque l'a priori sur le nombre de composants est élevé, l'algorithme doit explorer une très grande quantité de modèles de dimensions de plus en plus grandes, et nécessitant donc de plus en plus d'itérations. Ce nombre dépassant largement nos capacités de calculs disponibles, nous avons été contraints de restreindre nos simulations à des valeurs de λ peu élevées.

En étudiant la figure VI.3 on constate aussi un défaut de qualité d'estimation des paramètres par l'algorithme à sauts réversibles. Ce problème d'estimation est directement dû au phénomène de label switching et donc à la façon dont les deux méthodes dont nous disposons arrivent à l'éliminer. La figure VI.4 nous donne le détail des lois a posteriori sur les moyennes obtenues par ces deux méthodes, ainsi que celle avec la présence de label switching. On peut ici constater une meilleure gestion du phénomène par la méthode de Stephens que par la méthode de Celeux. Avec l'algorithme proposé par Stephens, la multimodalité a quasiment disparu des lois a posteriori des deux premiers composants, et celui-ci a même réussi à reconstituer la bonne loi a posteriori et à sélectionner les bons modes pour le troisième composant. En constatant la multimodalité des lois a posteriori, on se rend bien compte qu'utiliser une estimation par moyennes a posteriori n'est pas optimal. Etant alors contraints d'utiliser les modes de la loi a posteriori, nous sommes obligés de constater que, même après avoir éliminé au maximum le phénomène de label switching, nous pouvons être en présence de plusieurs modes, notamment pour le composant le moins visité (celui ayant souvent la plus faible proportion). Dans le cas présenté ici, le mode retenu par le programme en utilisant la méthode de Celeux avait déjà été sélectionné pour d'autres composants. Il aurait donc fallu sélectionner les modes correspondant aux valeurs 4 et 6. Ce qui aurait été pressenti par un analyste. On peut donc d'ores et déjà imaginer une amélioration future de l'algorithme permettant de sélectionner des modes non encore utilisés, sans pour autant qu'ils soient les plus importants. En effet, la méthode proposée par Stephens donne généralement de très bons résultats, comme dans le cas de la figure VI.4 , mais ne peut garantir la sélection du bon mode. Un moyen de résoudre ce problème consiste à programmer une méthode de recuit simulé afin de maximiser la loi a posteriori obtenue pour le jeu de paramètres dans son ensemble, c'est-à-dire $p\left(k, \pi, z^{(n)}, \mu, \Sigma, \beta \mid y^{(n)}\right)$. On obtient alors directement l'estimation de tous les paramètres correspondant au maximum local trouvé.

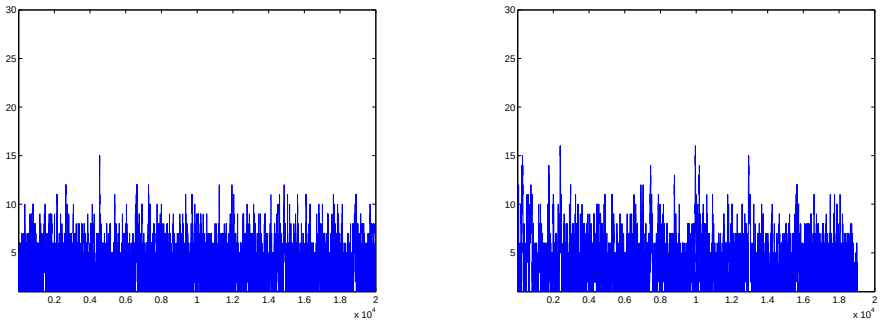
Afin de poursuivre et de compléter l'analyse de ce jeu de données, nous fournissons les lois a posteriori pour la variance (cf figure VI.5) correspondant au composant de moyenne $(-5, 3)$, ainsi que celle correspondant aux proportions des composants (figure VI.6). Ceci nous permet d'illustrer un autre aspect majeur de l'algorithme à sauts réversibles appliqué aux mélanges gaussiens multivariés. Le label switching est en effet présent sur chacun des paramètres à estimer, y compris les variances et les proportions. La gestion de ce phénomène semble cependant être plus délicate pour les proportions que pour les moyennes ou les variances, cette dernière remarque nous ayant conduits à estimer les proportions par la moyenne a posteriori.

Une question qui se pose lorsque l'on observe les résultats de l'algorithme à sauts réversibles est de savoir si celui-ci arrive à faire mieux que certaines méthodes classiques. Dans ce but nous avons calculé les valeurs des critères AIC et BIC pour différents nombres de composants. Les résultats sont fournis au tableau VI.1. A la figure VI.7 nous donnons le graphique des estimations obtenues en ayant fait tourner l'algorithme EM pour le nombre de composants trouvé

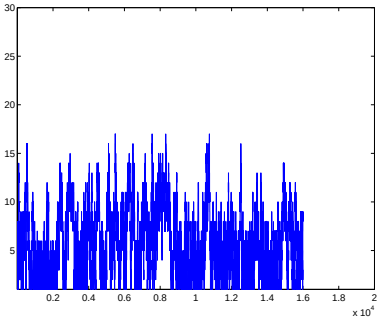
en minimisant le critère en question. Pour les données analysées ici, le critère BIC détecte bien 3 composants, alors que le critère AIC en détecte 5. Ceci met en lumière les critiques répandues dans la littérature sur la validité de ces critères, et permet de relativiser la difficulté de l'algorithme à sauts réversibles à estimer le nombre de composants ainsi que les paramètres du mélange. La méthodologie consistant à utiliser le mélange obtenu par la combinaison BIC-EM représente une alternative aisément programmable et très rapide. L'algorithme à sauts réversibles semble cependant fournir des résultats intéressants quant à l'estimation du nombre de composants, ainsi que pour l'estimation des paramètres. Les mélanges obtenus paraissent suffisamment adaptés à la réalité de l'échantillon pour pouvoir considérer cette méthode comme étant digne d'intérêt.

Afin de préciser les capacités d'exploration de notre algorithme, nous avons tracé le cusumplot (cf Mengersen et al. [1999] par exemple) pour les premières coordonnées de chacune des trois moyennes avec $\lambda = 1$. Ce graphique consiste à étudier la sous-chaîne obtenue par l'algorithme pour chaque composant en traçant l'évolution des coordonnées du paramètre d'intérêt après les avoir préalablement standardisées. La comparaison avec le graphique similaire obtenu avec un échantillon généré selon une loi uniforme (non tracé) permet de confirmer le bon comportement exploratoire de la chaîne. Les graphiques de la figure (VI.8) permettent de plus de constater que la chaîne effectue quelques excursions vers des valeurs très éloignées.

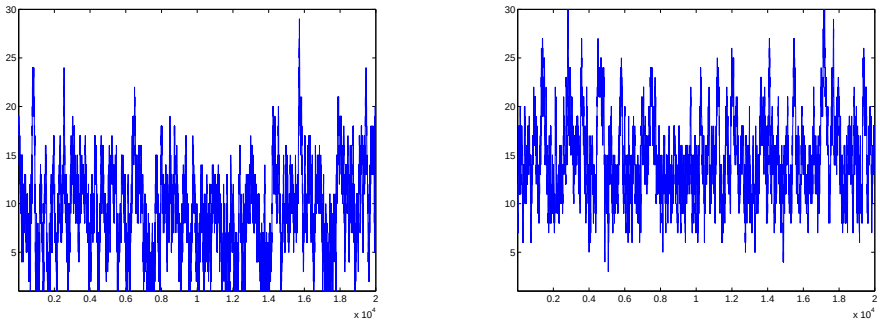
Lorsque l'on regarde tout simplement la disposition des points générés pour la première moyenne en superposition avec le csdfplot du mélange original (cf VI.9), nous avons la confirmation que l'algorithme génère des valeurs suffisamment éloignées pour une bonne exploration de l'espace d'état. La convergence peut être évaluée par la figure (VI.10). L'évolution de la probabilité a posteriori pour le nombre de composants en fonction des itérations nous permet de penser que l'algorithme est (dans ce cas précis) assez peu éloigné de la convergence. Pour des valeurs plus élevées de λ, les graphiques obtenus (non représentés ici) montrent de plus grands signes d'instabilité.



(a) $\lambda = 1$ (b) $\lambda = 2$

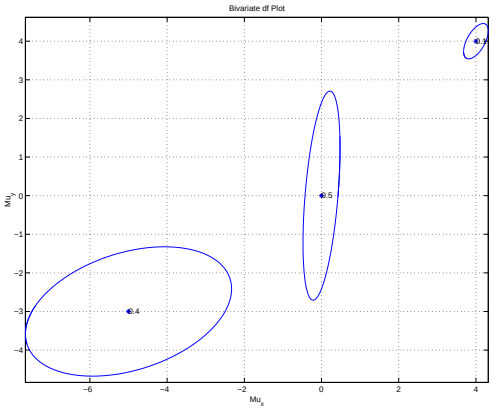


(c) $\lambda = 3$

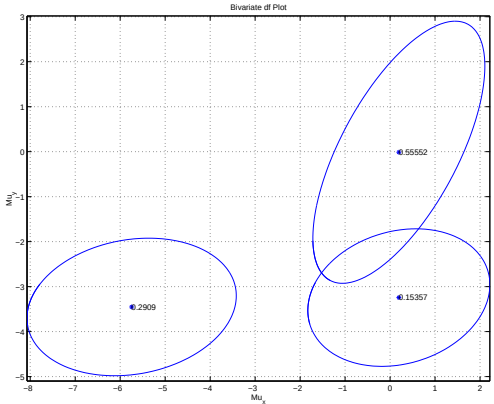


(d) $\lambda = 4$ (e) $\lambda = 5$

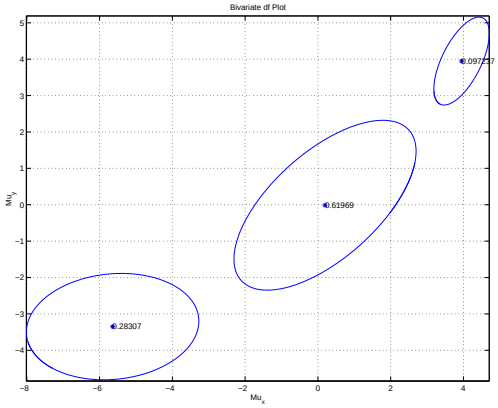
FIG. VI.2 – Evolution du nombre de composants pour les données "échantillon 1", et pour plusieurs valeurs du paramètre λ.



(a) Csdfplot du mélange original.

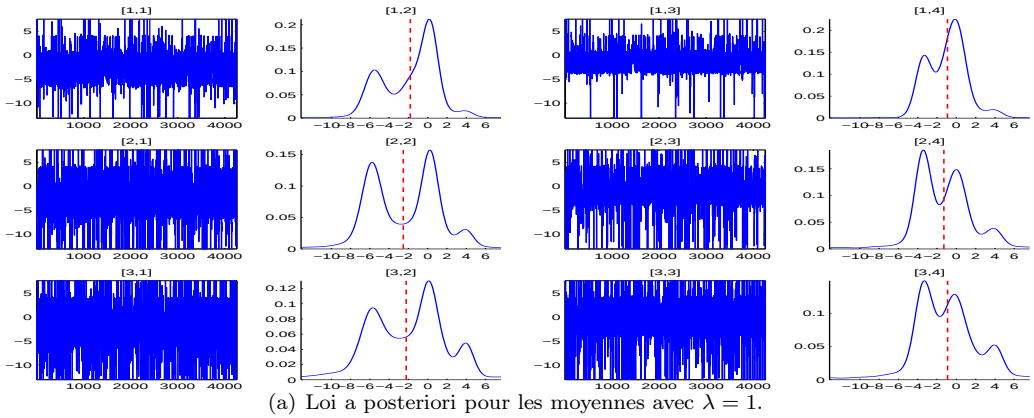


(b) Csdfplot du mélange estimé pour $\lambda = 1$ en enlevant le label switching par la méthode de Celeux.

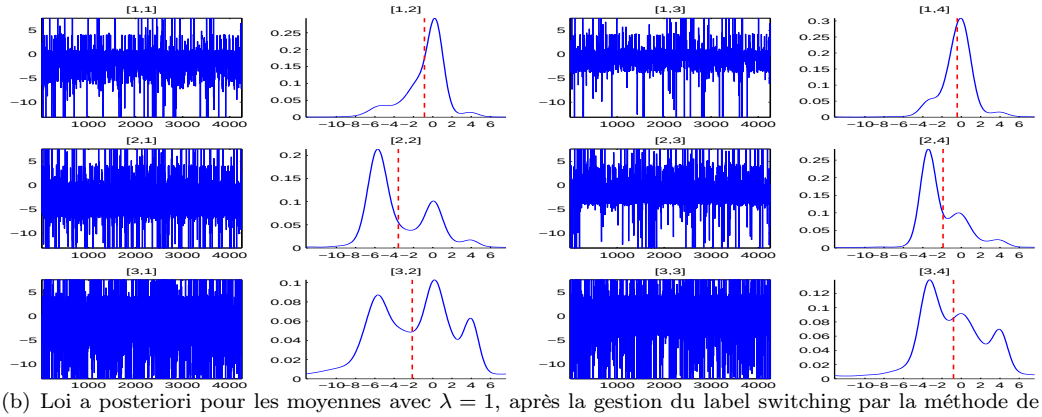


(c) Csdfplot du mélange estimé pour $\lambda = 1$ en enlevant le label switching par la méthode de Stephens.

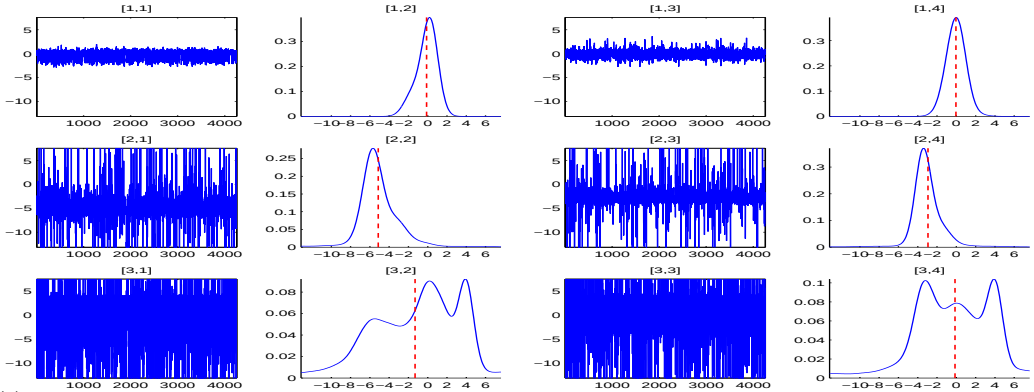
FIG. VI.3 – Csdfplot du mélange original et des deux estimations différentes selon la méthode utilisée pour enlever le label switching.



(a) Loi a posteriori pour les moyennes avec $\lambda = 1$.



(b) Loi a posteriori pour les moyennes avec $\lambda = 1$, après la gestion du label switching par la méthode de Celeux.



(c) Loi a posteriori pour les moyennes avec $\lambda = 1$, après la gestion du label switching par la méthode de Stephens.

FIG. VI.4 – Comparaison des lois a posteriori obtenues pour les moyennes en fonction de la méthode utilisée.

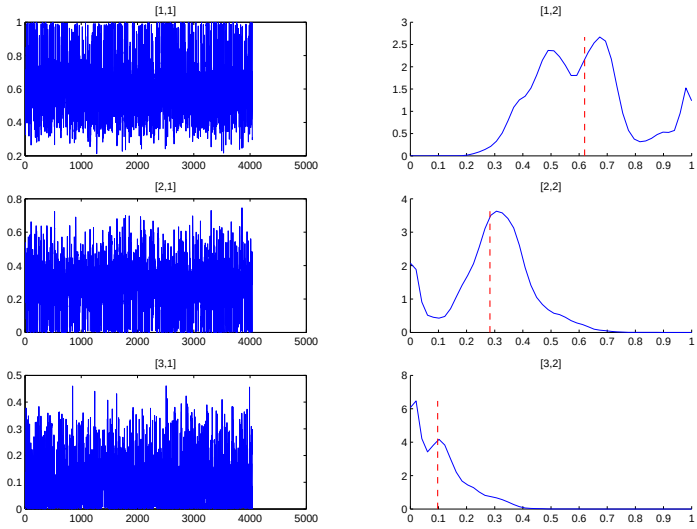


FIG. VI.5 – Loi a posteriori pour les proportions obtenue en utilisant l'algorithme de Stephens avec $\lambda = 1$.

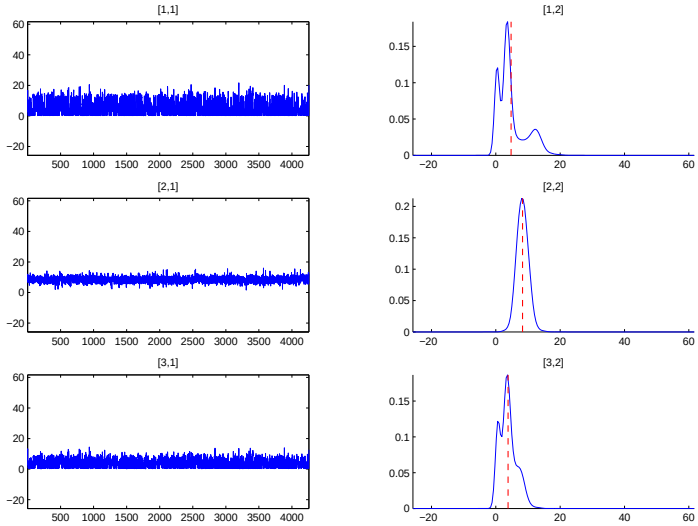
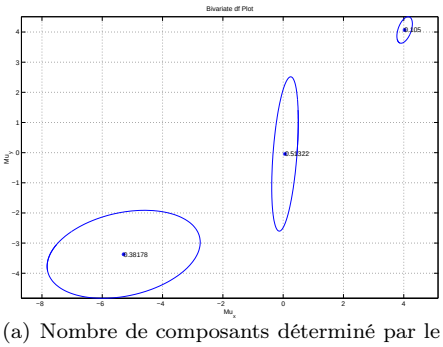
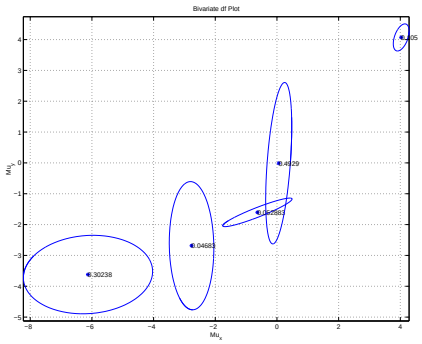


FIG. VI.6 – Loi a posteriori pour la variance du premier composant obtenue en utilisant l'algorithme de Stephens avec $\lambda = 1$.

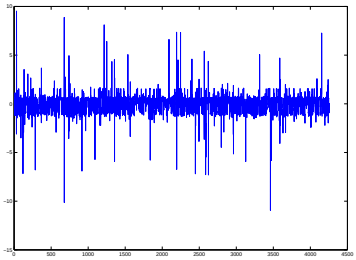


(a) Nombre de composants déterminé par le critère BIC.

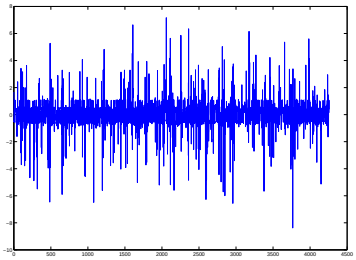


(b) Nombre de composants déterminé par le critère AIC.

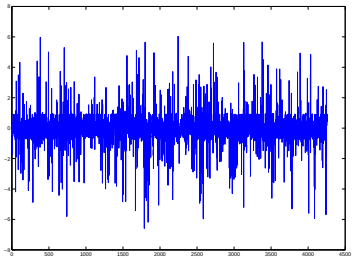
FIG. VI.7 – Csdfplot du mélange estimé pour les données "échantillon 1" en utilisant le nombre de composants trouvé par les critères AIC et BIC, puis avec les paramètres estimés par l'algorithme EM.



(a) Cusumplot pour la première coordonnée de μ_1



(b) Cusumplot pour la première coordonnée de μ_2



(c) Cusumplot pour la première coordonnée de μ_3

FIG. VI.8 – Evolution des premières coordonnées des trois moyennes moyennes standardisées (Cusumplot) pour $\lambda = 1$.

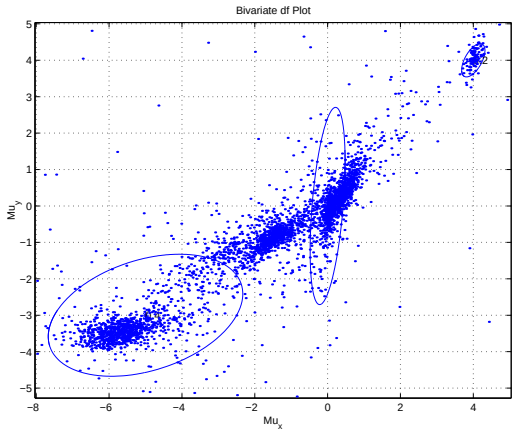
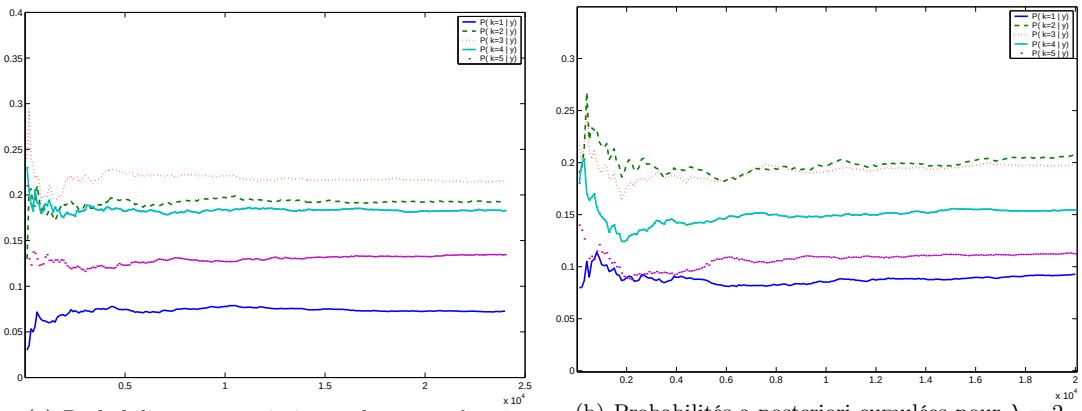


FIG. VI.9 – Csdfplot du mélange original superposé avec le nuage de points des valeurs visitées par l'algorithme pour la première moyenne avec $\lambda = 2$.

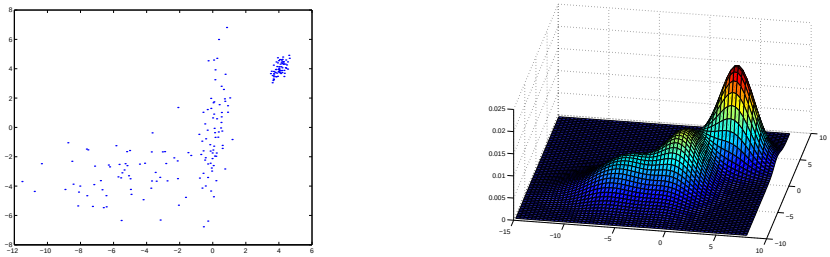


(a) Probabilités a posteriori cumulées pour $\lambda = 1$

(b) Probabilités a posteriori cumulées pour $\lambda = 2$

FIG. VI.10 – Evolution des probabilités a posteriori pour le nombre de composants en fonction du nombre d'itérations (en abscisse), et pour différentes valeurs de λ.

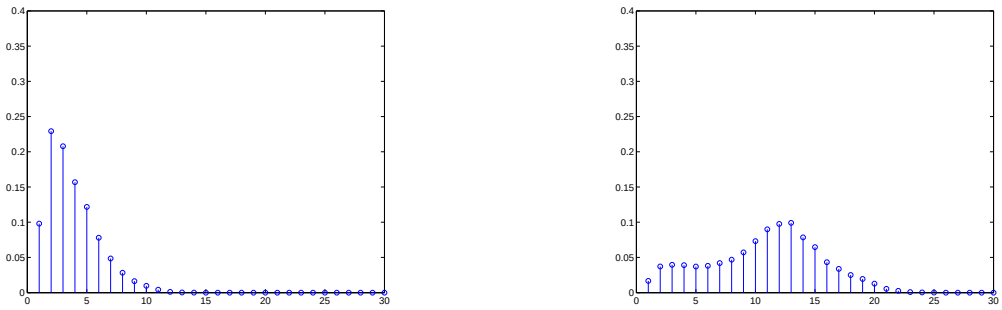
3 Echantillon simulé à 3 composants de proportions identiques



(a) Données "échantillon 2 "

(b) Estimation par noyaux gaussiens des données "échantillon 2"

FIG. VI.11 – Visualisation des données "échantillon 2".



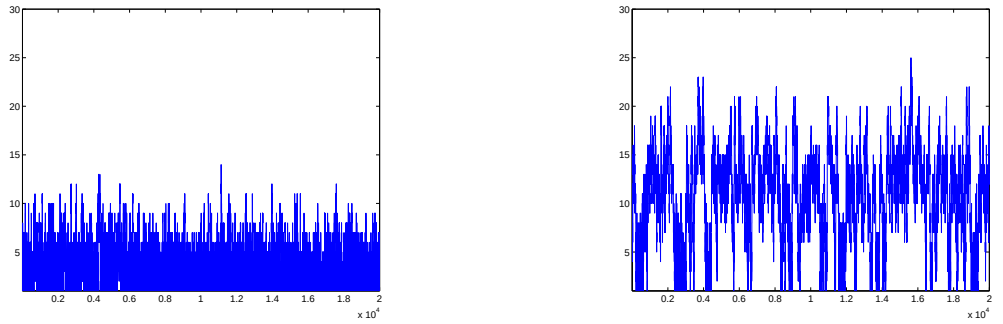
(a) $\lambda = 1$

(b) $\lambda = 4$

FIG. VI.12 – Lois a posteriori du nombre de composants pour les données "échantillon 2", et pour plusieurs valeurs du paramètre λ.

L'échantillon analysé ici est presque identique au précédent, excepté que tous ses composants ont la même proportion. Nous donnons à nouveau les lois a posteriori obtenues pour le nombre de composants pour différentes valeurs de λ en figure VI.12. Comme on pouvait le supposer, la présence de label switching est bien plus importante lorsque les composants ont le même poids. En effet, en examinant les lois a posteriori pour les moyennes et pour $\lambda = 1$ (figure VI.16), on constate une très forte multimodalité menant à des lois quasi identiques pour toutes les coordonnées. On peut d'ailleurs remarquer qu'obtenir ce type de similitude peut être considéré comme un critère de convergence de l'algorithme à sauts réversibles. En effet, son absence peut signifier que la chaîne a insuffisamment exploré l'ensemble des modes de la loi a posteriori. Ayant observé cela, on est en mesure d'évaluer une nouvelle fois les capacités des deux algorithmes dont nous disposons pour éliminer ce phénomène de label switching. On constate à nouveau que la méthode proposée par Stephens est la plus performante. Celle-ci a pu en effet éliminer la plus grande partie du label switching présent tout en sélectionnant les bons modes de la loi a posteriori. Nous donnons à la figure VI.15 l'estimation du mélange obtenue par l'étude de la chaîne à 3 composants avec $\lambda = 1$ (remarquons cependant que le composant le plus probable dans ce cas était 2). L'estimation fournie est nettement plus correcte en ce qui concerne les moyennes et les variances que celles de la figure VI.3 obtenues pour l'"échantillon 1" qui possédait les mêmes moyennes et variances.

Des composants également répartis semblent donc plus favorables à l'estimation malgré la présence accrue de label switching. On peut donc s'attendre à ce que des composants relativement



(a) $\lambda = 1$ (b) $\lambda = 4$

FIG. VI.13 – Evolution du nombre de composants pour les données "échantillon 2", et pour plusieurs valeurs du paramètre λ.

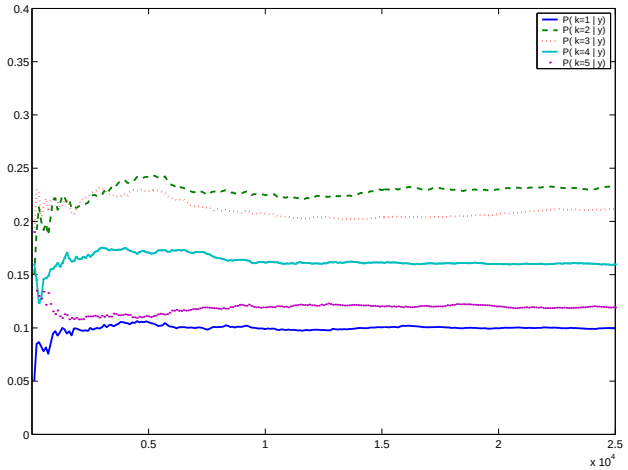


FIG. VI.14 – Evolution des probabilités a posteriori pour le nombre de composants des données "échantillon2" en fonction du nombre d'itérations (en abscisse), et pour $\lambda = 1$.

peu représentés, ayant une proportion inférieure ou égale à 0.1, soient peu visités et donc difficiles à repérer compte tenu de la présence de label switching ayant aussi pour conséquence le mélange des observations.

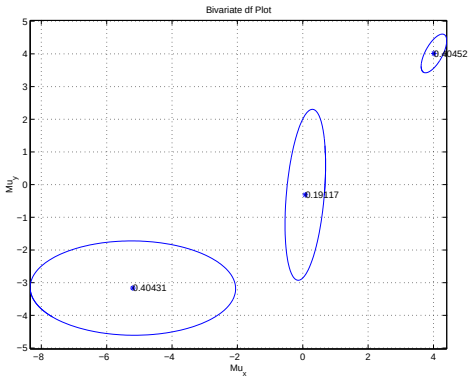
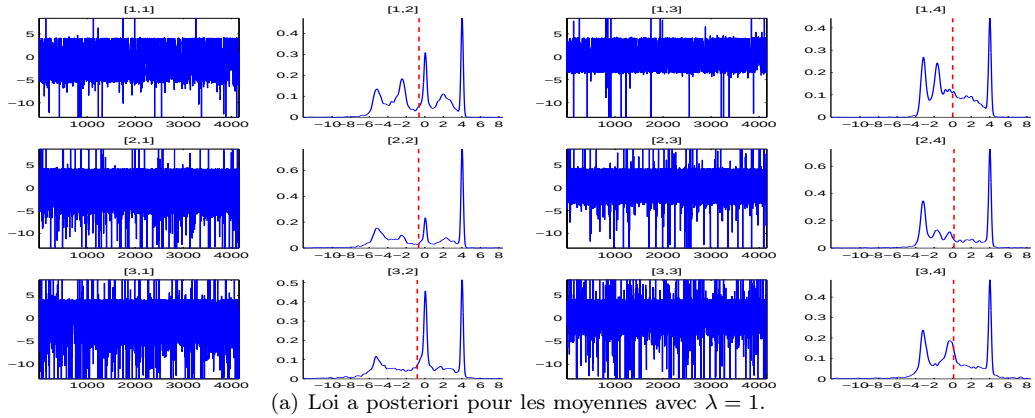
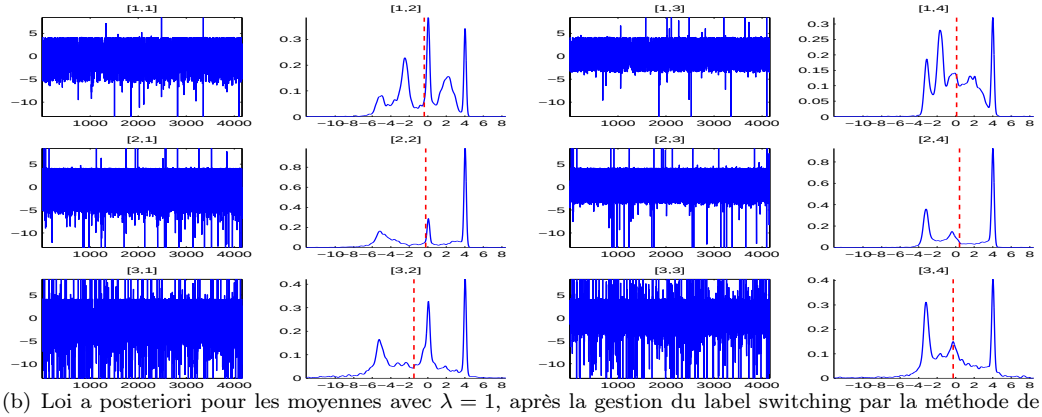


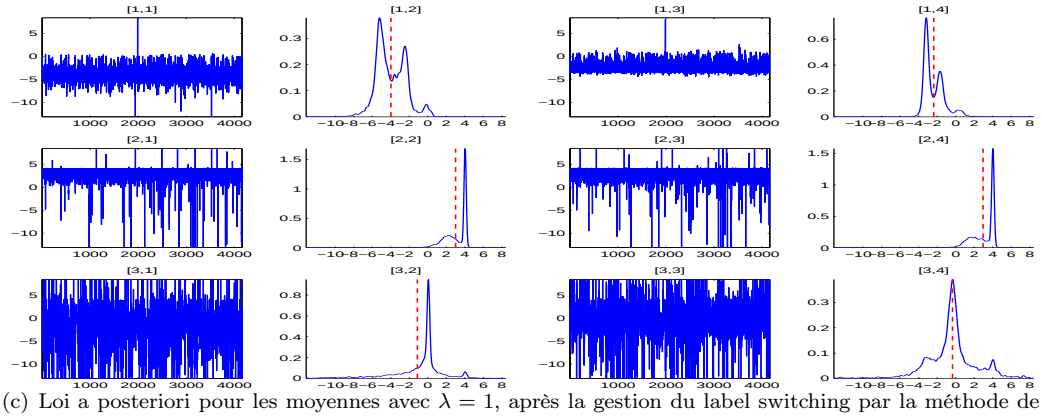
FIG. VI.15 – Csdfplot du mélange estimé pour $\lambda = 1$ en enlevant le label switching par la méthode de Stephens.



(a) Loi a posteriori pour les moyennes avec $\lambda = 1$.



(b) Loi a posteriori pour les moyennes avec $\lambda = 1$, après la gestion du label switching par la méthode de Celeux.



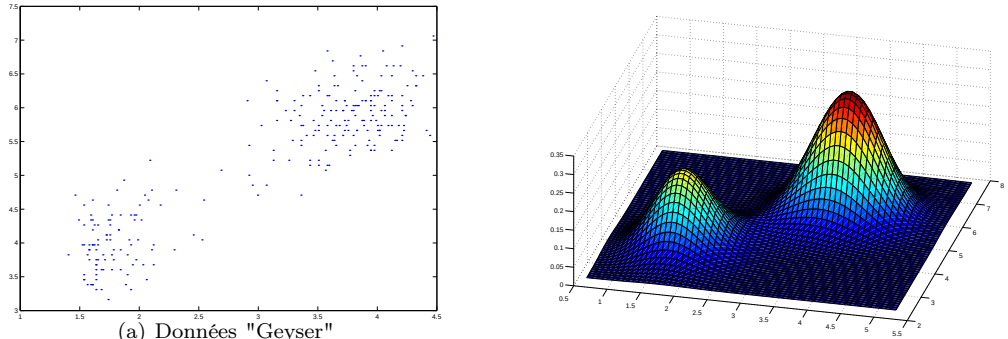
(c) Loi a posteriori pour les moyennes avec $\lambda = 1$, après la gestion du label switching par la méthode de Stephens.

FIG. VI.16 – Comparaison des lois a posteriori obtenues pour les moyennes, pour les données "échantillon 2", en fonction de la méthode utilisée.

nombre de composants	2	3	4	5	6	7	8	9	10
AIC	802	797.8	707.7	813	806	821	833	845	859
BIC	782	768	757	764	747	752	755	757	760

TAB. VI.2 – Critères AIC et BIC pour les données "geyser" et pour différents nombres de composants.

4 Données "Geyser"



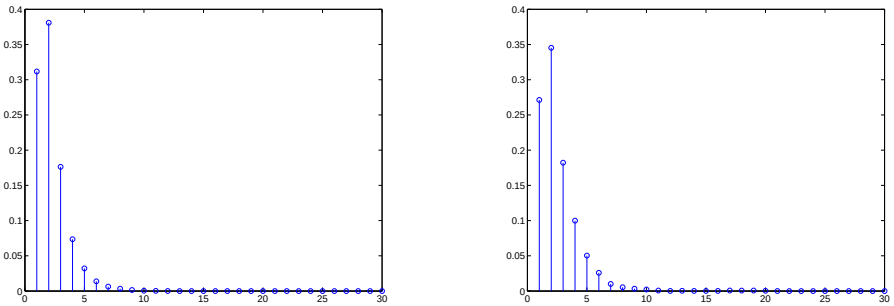
(a) Données "Geyser"

(b) Estimation par noyaux gaussiens des données "Geyser"

FIG. VI.17 – Visualisation des données "geyser".

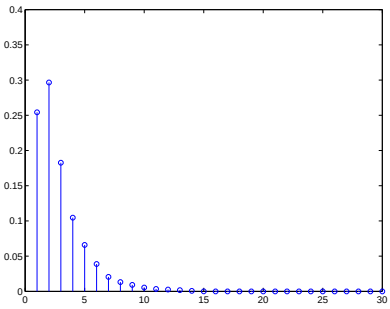
A la figure VI.17, nous pouvons visualiser les données geyser ainsi que leur densité estimée par la méthode des noyaux gaussiens. Il s'agit de 272 mesures sur des éruptions du geyser Old Faithful du Yellowstone National Park. Ce sont des données bivariées comportant une mesure de durée de l'éruption associée au temps d'attente jusqu'à l'éruption suivante. Remarquons que pour des raisons purement techniques nous avons normalisé les données. Cet échantillon est fréquemment étudié dans la littérature car la dépendance intrinsèque des données en rend l'étude difficile. En effet, les critères usuels basés sur une minimisation de la vraisemblance pénalisée détectent 4 composants pour le critère AIC, et 6 composants pour BIC (cf figure VI.2). Les mélanges obtenus par l'algorithme EM en initialisant ce dernier avec le nombre de composants trouvés sont fournis à la figure VI.24. Si l'on considère des critères de vraisemblance, le nombre de composants apparait comme étant incertain malgré la présence de deux amas de points nettement séparés. Les lois a posteriori pour le nombre de composants obtenus dans nos simulations sont présentées à la figure VI.18. Toutes les valeurs testées pour la loi a priori ont fourni une probabilité élevée pour un modèle à deux composants, ce qui peut sembler satisfaisant. On peut aussi examiner le graphique de l'évolution des lois a posteriori du nombre de composants en fonction des itérations (cf figure VI.20) qui montre un comportement satisfaisant.

L'analyse des évolutions à la figure VI.19 confirme aussi les capacités importantes d'exploration de l'espace d'état de l'algorithme. En effet, des modèles à plus de dix composants sont régulièrement visités et plus de 80% des itérations de l'algorithme donnent lieu à un changement de dimension. La très bonne exploration de l'espace d'état est confirmée par le nuage de points représentant les valeurs visitées par la chaîne correspondant à la première moyenne pour $\lambda = 1$ (cf figure VI.21). On est en présence d'une très grande instabilité des données générées avec un nuage diffus, permettant toutefois de dégager trois amas de points. On peut rapprocher cette configuration de la figure VI.22 représentant les lois a posteriori pour les moyennes. Les deux estimations à noyaux de la loi a posteriori pour la première moyenne (première ligne de VI.22(a)) montrent la présence d'un mode très marqué accompagné de deux petits modes. L'amas central de la figure VI.21 représente le mode principal, alors que les deux amas plus petits (correspondants aux estimations des centres des amas de points des données geyser) représentent les deux

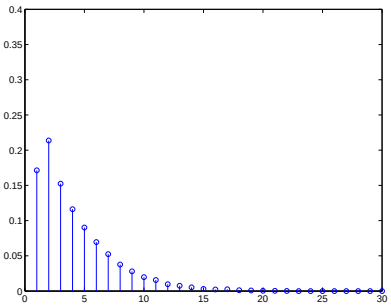


(a) $\lambda = 1$

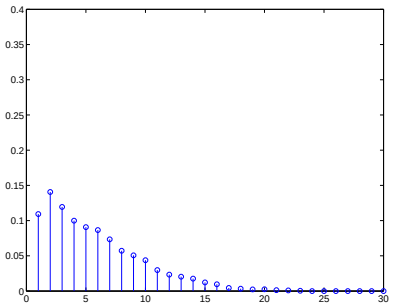
(b) $\lambda = 2$



(c) $\lambda = 3$



(d) $\lambda = 4$



(e) $\lambda = 5$

FIG. VI.18 – Lois a posteriori du nombre de composants pour les données Geyser, et pour plusieurs valeurs du paramètre λ.

modes adjacents. L'algorithme génère donc dans ce cas précis des valeurs situées dans une zone ayant peu de signification réelle. Ce phénomène est à mettre en parallèle avec la mauvaise estimation des proportions. Les mélanges trouvés par notre algorithme pour $\lambda = 1$ sont représentés en figure VI.23 pour différentes méthodes de gestion du "label switching". On peut vérifier qu'un des composants est majoritaire avec une proportion supérieure à 0.95. Ce phénomène provient aussi de la gestion du "label switching". Si un des deux composants est correctement re-indicé, et donc correctement estimé, ce n'est pas le cas pour le second. Les données simulées appartenant à l'amas de points situé au centre sont donc allouées au second composant, qui possède donc une proportion très élevée, étant devenu majoritaire. La présence de ce mode très marqué peut probablement s'expliquer par la présence, dans les données originales, de quelques points situés exactement entre les deux amas principaux. Ces points semblent "attirer" les éléments générés par la chaîne de Markov. Ainsi, lorsque $k = 2$, la chaîne représentant les valeurs de la première moyenne se retrouve attirée au centre, et il en est de même pour la chaîne représentant les valeurs de la seconde moyenne. Un moyen de confirmer cette intuition consisterait à enlever ces quelques

valeurs puis à relancer l'algorithme.

5 Conclusion

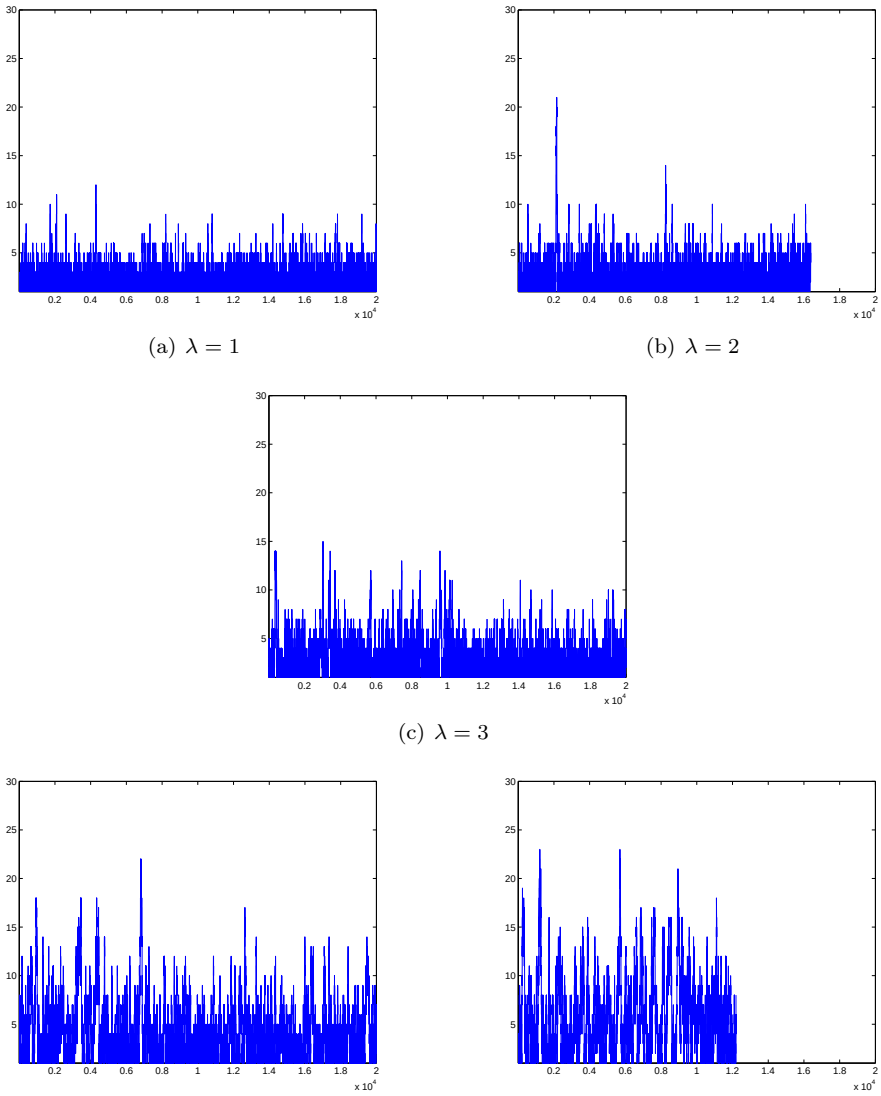
Les simulations effectuées dans ce chapitre ont été présentées et obtenues à l'aide du logiciel MATLAB. Nous avons déjà souligné les limitations et les avantages de ce logiciel, mais nous revenons ici sur la nécessité d'une programmation dans un langage compilé.

En effet, la lenteur de calcul de MATLAB ne nous a pas permis de pousser nos investigations suffisamment loin pour atteindre la convergence de notre algorithme. Si l'on peut affirmer que pour $\lambda = 1$ nous étions proches de la convergence, pour $\lambda > 2$ nous en sommes certainement assez loin. Les résultats obtenus, quoique prometteurs, nécessittent cependant une analyse plus approfondie. En premier lieu, de plus amples études de convergence doivent être menées avec un nombre d'itérations sensiblement plus élevé. Il est tout aussi nécessaire de vérifier la validité des calculs en faisant tourner l'algorithme sans données pour vérifier que l'on retrouve bien la loi a priori.

D'autres méthodes sont disponibles pour éliminer l'effet de "label switching" et doivent également être testées. Nous avons aussi signalé plusieurs améliorations possibles du schéma proposé par Richardson et Green [1997]. Voici brièvement quelques pistes en vue de futurs travaux :

- Il est possible de n'autoriser qu'un seul changement de dimension.
- On peut spécifier des mouvements différents.
- Il est envisagé d'utiliser la paramétrisation des variances introduite par Banfield et Raftery [1993].
- On peut implémenter une procédure de recuit simulé pour déterminer le mode de la loi a posteriori de manière globale.

Certaines de ces voies sont en cours d'exploration mais sont trop peu avancées pour rentrer dans le cadre de cette thèse. La multiplicité des approches et améliorations possibles rendent la méthode à sauts réversibles appliquée aux mélanges gaussiens attrayante et prometteuse.



(a) $\lambda = 1$ (b) $\lambda = 2$

(c) $\lambda = 3$

(d) $\lambda = 4$ (e) $\lambda = 5$

FIG. VI.19 – Evolution du nombre de composants pour les données Geyser, et pour plusieurs valeurs du paramètre λ.

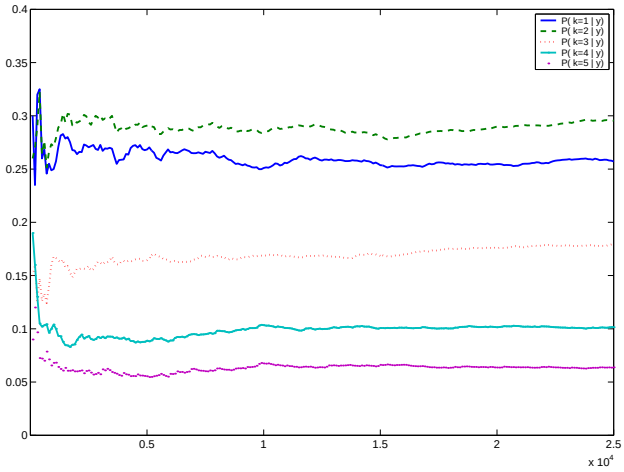


FIG. VI.20 – Evolution des probabilités a posteriori pour le nombre de composants des données geyser en fonction du nombre d'itérations (en abscisse), et pour $\lambda = 3$.

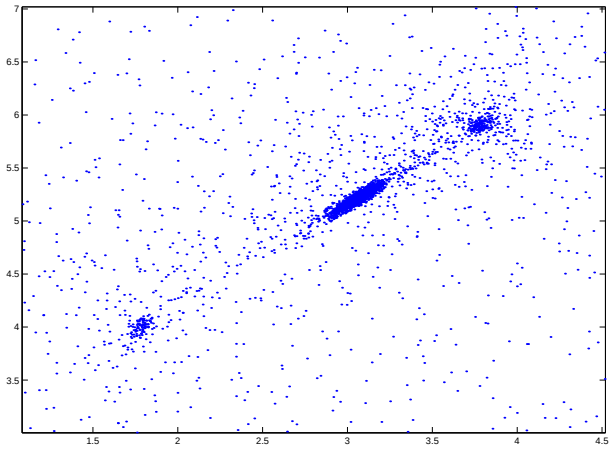
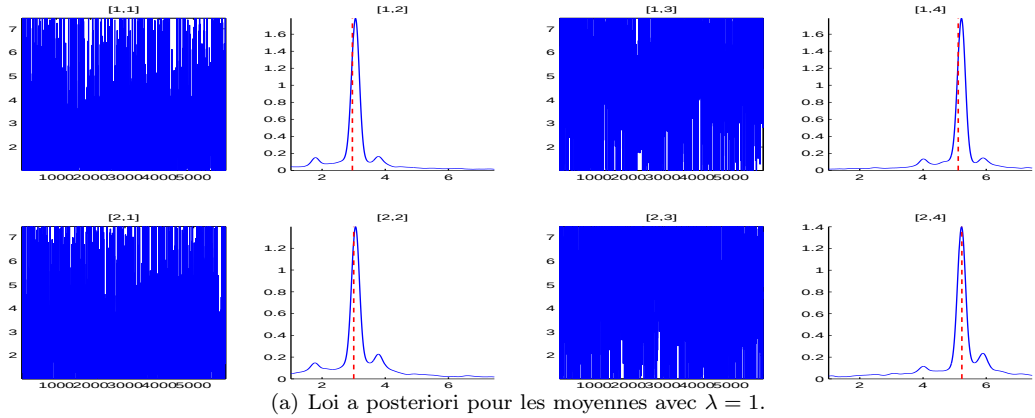
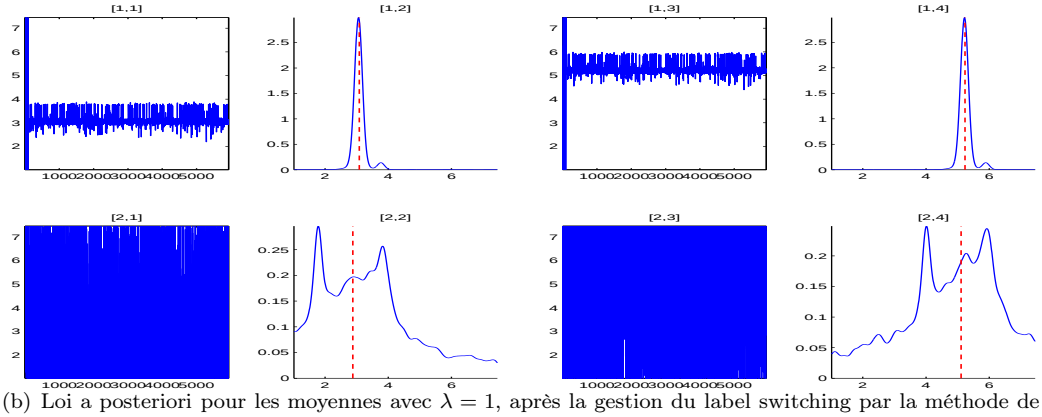


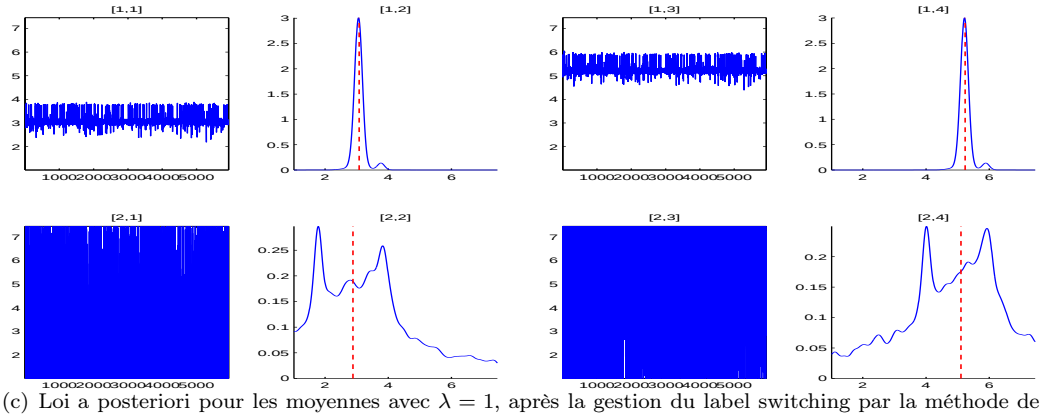
FIG. VI.21 – nuage de points des valeurs visitées par l'algorithme pour la première moyenne avec $\lambda = 1$.



(a) Loi a posteriori pour les moyennes avec $\lambda = 1$.

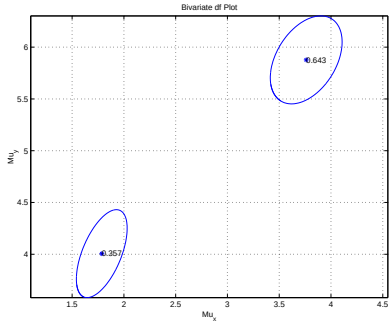


(b) Loi a posteriori pour les moyennes avec $\lambda = 1$, après la gestion du label switching par la méthode de Celeux.

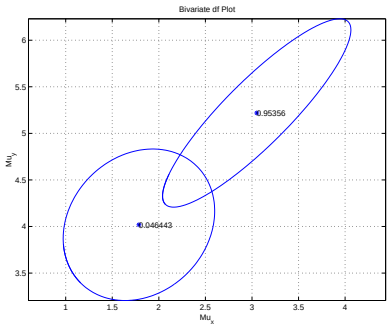


(c) Loi a posteriori pour les moyennes avec $\lambda = 1$, après la gestion du label switching par la méthode de Stephens.

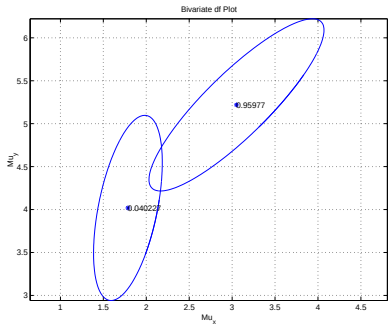
FIG. VI.22 – Comparaison des lois a posteriori obtenues pour les moyennes, pour les données Geyser, selon la méthode utilisée.



(a) Csdfplot du mélange estimé par l'algorithme de Gibbs pour k=2.

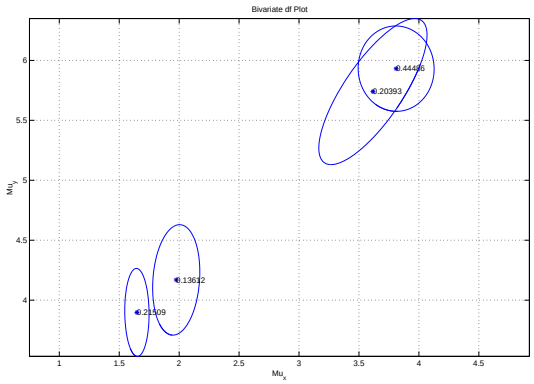


(b) Csdfplot du mélange estimé pour $\lambda = 1$ en enlevant le label switching par la méthode de Celeux.

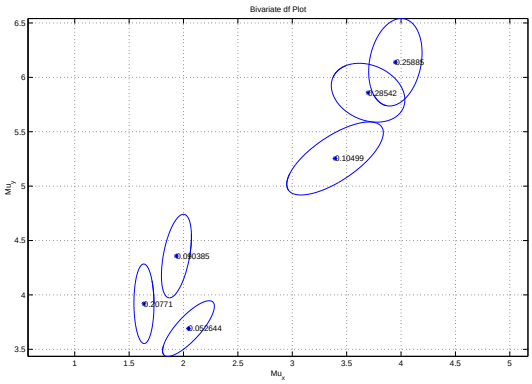


(c) Csdfplot du mélange estimé pour $\lambda = 1$ en enlevant le label switching par la méthode de Stephens.

FIG. VI.23 – Csdfplot du mélange original et des deux estimations différentes obtenues pour l'échantillon "Geyser" et selon la méthode utilisée pour enlever le label switching.



(a) Nombre de composants déterminé par le critère BIC.



(b) Nombre de composants déterminé par le critère AIC.

FIG. VI.24 – Csdfplot du mélange estimé pour les données "Geyser" en utilisant le nombre de composants trouvé par les critères AIC et BIC, puis avec les paramètres estimés par l'algorithme EM.

Chapitre VII

Fiabilité des systèmes : qualification des I.L.S.

Le travail qui suit a été réalisé à la suite d'un contrat pour les Services Techniques de la Navigation Aérienne (S.T.N.A.) avec la collaboration de Véronique Font (société I.X.I). Le travail a porté sur des problèmes de fiabilité et de qualification d'appareils. Il nous a permis de développer un concept théorique intéressant en statistique appliquée : les probabilités de confiance. Il nous a permis également d'écrire un article publié à la Revue de Statistique Appliquée (d'Estampes et al. [2003]). Celui-ci fait l'objet de la première section.

Ce travail a fait l'objet de trois rapports techniques (un en janvier 2001 et deux en février 2002). A la suite de ces trois rapports, nous avons développé un programme Matlab qui permet entre autres de tracer des plans de fiabilité en fonction de différents paramètres (risque de première espèce, risque de deuxième espèce,...). La seconde section fournit une note technique sur le programme créé.

1 Test séquentiel : Niveau de confiance après acceptation

Écrit par Ludovic d'Estampes[a], Bernard Garel[a] et Guillaume Saint Pierre[a]

[a]Groupe Statistique et Probabilités, LEN7, ENSEEIHT, B.P.7122,
2, rue Camichel 31071 Toulouse Cedex 7, France.

Résumé : Nous développons ici la notion de niveau de confiance après acceptation lors d'un test séquentiel du rapport des probabilités de Wald. Les notions de borne de confiance et d'intervalle de confiance après acceptation sont bien connues. La notion de niveau de confiance apparaît comme une notion duale de la précédente. Ces concepts sont illustrés dans le cadre d'un problème de fiabilité des systèmes concernant la qualification de systèmes d'atterrissage aux instruments ou Instrument Landing System (ILS).

Mots-clés : *test séquentiel, loi exponentielle, borne et niveau de confiance.*

Abstract : In this paper we develop the notion of confidence level at acceptance, during a probability ratio sequential test of Wald. The notions of confidence limits and confidence interval at acceptance are well known. The notion of confidence level appears as the dual of the former. These concepts are illustrated in a reliability of technical systems problem, namely the Instrument Landing System (ILS) certification.

Keywords : *sequential testing, exponential distribution, confidence bounds and probabilities.*

1.1 Introduction

La notion de test séquentiel du rapport des probabilités a été introduite par Wald durant la seconde guerre mondiale. Mais son livre ne fut publié qu'en 1947 (cf Wald [1947]).

Contrairement aux tests classiques, le nombre d'observations nécessaires à un test séquentiel n'est pas fixé a priori. En revanche, les risques de première espèce et de seconde espèce sont, quant à eux, fixés.

La technique des tests séquentiels est peu enseignée dans le supérieur. Cependant, elle est très utilisée en fiabilité et plus généralement pour tout problème de certification, comme par exemple celle de médicaments nouveaux. En effet cette technique permet une économie de mesures et par là même de temps, de l'ordre de 50% en moyenne.

L'une des phases de la qualification est un test paramétrique entre deux hypothèses simples associées à deux valeurs possibles du paramètre. Une fois l'hypothèse de qualification acceptée, il est possible de calculer des limites de confiance inférieures et supérieures pour la vraie valeur inconnue du paramètre. Mais les industriels souhaitent en fait pouvoir répondre à la question suivante : avec quelle probabilité le paramètre inconnu dépasse-t-il une valeur déterminée ?

Nous commençons par rappeler brièvement comment se calcule une borne inférieure de confiance dans le cas général. Nous introduisons ensuite la notion de niveau de confiance. Puis nous nous plaçons plus précisément dans le cas où la loi de probabilité supposée des observations est exponentielle et développons la notion de niveau de confiance à l'acceptation lors d'un test séquentiel. Enfin, nous effectuons une étude sur données réelles.

1.2 Borne de confiance, niveau de confiance

La notion de borne de confiance est, en général, bien présentée dans les manuels et relève des techniques d'estimation.

On souhaite obtenir des précisions sur la valeur d'un paramètre réel θ inconnu. Lorsqu'il existe une fonction pivotale pour θ (cf [Tassi, 1985, p.223]), le calcul d'une borne ou d'un intervalle de confiance s'effectue très facilement. Dans le cas général, on peut opérer de la façon suivante (cf Mood et Graybill [1963]).

Soit Y une variable aléatoire réelle (V.A.R.) dont la loi de probabilité dépend d'un paramètre θ (on peut penser à un estimateur de θ). Notons $y \mapsto g(y, \theta)$ sa densité. Soit γ fixé $\in [0; 1]$, on peut trouver h tel que

$$\mathbb{P}[Y > h] = \int_h^{+\infty} g(y, \theta)\, dy = \gamma.$$

La valeur de h va dépendre évidemment de la valeur de θ. On suppose $\theta \mapsto h(\theta)$ strictement croissante. Notons y la valeur observée de Y.

Dans le plan $(0; \theta; Y)$, une ligne horizontale passant par le point de coordonnées $(0; y)$ coupe la courbe $h(\theta)$ en un point d'abscisse θ_L. Notons Θ_L la V.A.R. dont les réalisations sont θ_L. On a

$$[\Theta_L < \theta] \Longleftrightarrow [Y < h(\theta)] \ ,$$

d'où

$$\mathbb{P}[\Theta_L < \theta] = \mathbb{P}[Y < h(\theta)] = 1 - \gamma \ .$$

On appelle $]\Theta_L; +\infty[$ l'intervalle de probabilité $1 - \gamma$ à droite pour θ et $]\theta_L; +\infty[$ l'intervalle de confiance à $100(1 - \gamma)\,\%$ à droite pour θ. La valeur θ_L est alors appelée borne de confiance inférieure à $100(1 - \gamma)\,\%$.

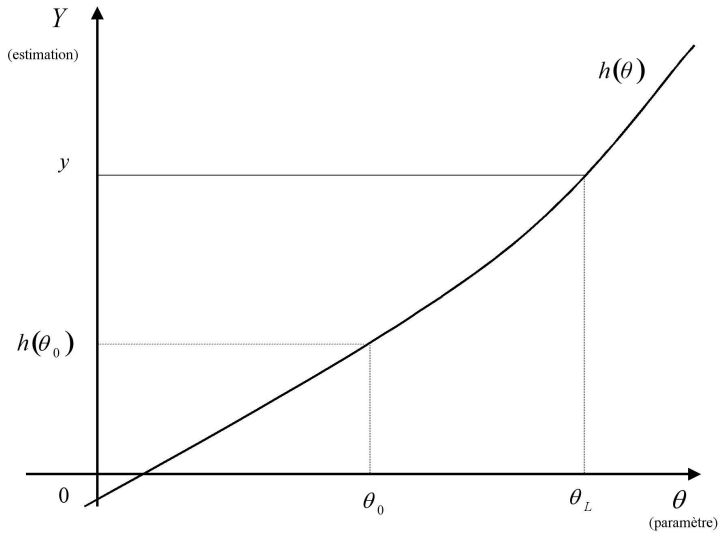


FIG. VII.1 – Lien entre paramètre et estimation

Il n'est pas forcément nécessaire de calculer explicitement la fonction h. En effet, θ_L est la valeur de θ telle que $h(\theta_L) = y$ et donc pour laquelle

$$\int_y^{+\infty} g(u, \theta)\, du = \gamma. \tag{VII.1}$$

S'il est possible d'exprimer simplement $\int_y^{+\infty} g(u, \theta)\, du$ et de résoudre l'équation en θ alors la racine θ_L est la borne inférieure souhaitée.

Inversement, étant donnée y une observation de Y et θ_0 une valeur fixée pour θ, l'équation (VII.1) peut être résolue en γ. C'est ce que nous proposons et donnons à cet effet la définition suivante.

Définition 8 Soit θ_0 une valeur fixée a priori et $\gamma(\theta_0)$ la solution de (VII.1) en γ. On appelle niveau de confiance que θ soit supérieur à θ_0 après observation de y la valeur $1 - \gamma(\theta_0)$ ainsi trouvée.

Essayons maintenant de donner une interprétation probabiliste à cette notion nouvelle. On sait que la notion d'intervalle de confiance peut s'interpréter de façon rigoureuse en adoptant un point de vue fréquentiste : si l'on répétait un grand nombre de fois l'expérience et si on notait à chaque fois l'intervalle ainsi trouvé, alors dans $100(1-\gamma)\,\%$ des cas en moyenne, la vraie valeur de θ se trouverait dans cet intervalle. Une autre interprétation est possible dans un contexte bayésien. Mais il faut alors disposer d'une probabilité a priori sur θ (cf Lecoutre [1997]).

L'interprétation du niveau de confiance défini plus haut ne se fait qu'en revenant à la notion d'intervalle de confiance : si l'on avait cherché une borne inférieure de confiance à $100(1-\gamma_0)\,\%$ après observation de y, on aurait trouvé la valeur θ_0 dont on est parti pour trouver γ_0.

La valeur $1 - \gamma_0$ représente donc la probabilité de recouvrement de $]\theta_0; +\infty[$, intervalle de confiance à droite pour θ.

La notion de niveau de confiance apparaît alors comme duale de celle de borne de confiance ou d'intervalle de confiance.

Ces calculs peuvent s'appliquer dans le cas où la loi de probabilité de Y est discrète. Mais dans ce cas, on ne trouvera pas forcément de valeur de θ qui réponde exactement à la question.

1.3 Test séquentiel tronqué

Nous allons nous placer maintenant dans le cadre des plans de tests séquentiels de la norme américaine (cf Department of Defense (USA) [1998]) concernant le temps moyen de bon fonctionnement (Mean Time Between Failures, MTBF) d'équipements électroniques.

Pour le problème qui nous intéresse, ces équipements électroniques sont des Instrument Landing Systems qui, au sol, sont de deux sortes :

- le «localiser», situé en extrémité de la piste d'atterrissage d'un aéroport et qui sert à déterminer un plan vertical ;
- le «glidepath», situé sur le côté droit de la piste. Il sert à déterminer une surface courbe.

L'intersection du plan et de la surface courbe constitue le rail électronique le long duquel l'avion doit évoluer pour son atterrissage.

Ces équipements étant utilisés pour guider l'avion par des conditions de visibilité réduite (le pilote ne voyant pas la piste jusqu'à une distance pouvant être très faible), la disponibilité du signal pendant la phase d'approche et d'atterrissage doit être de haut niveau. Les interruptions d'émission ne doivent pas dépasser une certaine probabilité d'émergence réglementaire. Ces éventuelles interruptions d'émission sont qualifiées d'outages et l'on parle de temps moyen entre deux outages (MTBO).

Pour la séquence des outages nous utilisons un modèle de processus de Poisson homogène. Si on note $X_1, \ldots, X_r$ la suite des temps de bon fonctionnement entre deux outages, les VAR X_i, $i = 1, \ldots, r$ sont alors considérées comme indépendantes et de même loi exponentielle de densité

$$x \longmapsto f(x, \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \mathbb{1}_{\mathbb{R}^+}(x) . \tag{VII.2}$$

Nous rappelons maintenant quelques éléments de la théorie de Wald. On désire tester $H_0 : \theta = \theta_0$ contre l'hypothèse $H_1 : \theta = \theta_1$, avec $\theta_1 < \theta_0$. Dans le cadre séquentiel la taille de l'échantillon n'est plus fixée. On rajoute des données x_i une par une et on se demande à chaque fois si on peut accepter ou non H_0 ou si d'autres observations sont nécessaires.

L'équivalent séquentiel de la règle de décision de Neyman-Pearson est le théorème de Wald (cf Ghosh [1970] ou Siegmund [1985]) qui nous indique de poursuivre la collecte de données tant que :

$$B < L_r = \prod_{i=1}^{r} \frac{f(x_i, \theta_1)}{f(x_i, \theta_0)} < A. \tag{VII.3}$$

On stoppe la procédure en choisissant H_0 dès que $L_r \leqslant B$ ou H_1 dès que $A \leqslant L_r$. Cette règle de décision s'appelle le test séquentiel du rapport des probabilités (TSRP). Les constantes A et B sont déterminées à partir des risques α et β.

Le risque α représente la probabilité de refuser l'équipement alors qu'il est opérationnel. Le risque β représente la probabilité d'accepter l'équipement alors que celui-ci est défectueux. La valeur α est généralement qualifiée de risque du producteur alors que β représente le risque de l'acheteur.

Selon la théorie de Wald, les constantes A et B sont approchées respectivement par $\frac{1-\beta}{\alpha}$ et $\frac{\beta}{1-\alpha}$.

A chaque étape m de la procédure du test séquentiel, on peut définir trois régions

- D_m^1 la région de rejet,
- D_m^0 la région d'acceptation,
- D_m la région de continuation.

On remarque que D_{m-1} est l'union disjointe de D_m, D_m^0, D_m^1. Soit θ fixé. Pour tout test séquentiel du rapport des probabilités et tout $m \geqslant 1$, on peut écrire :

$$\begin{aligned} &\mathbb{P}_\theta(\text{continuer la procédure à l'étape } m) \\ &\quad + \quad \mathbb{P}_\theta(\text{décider } H_0 \text{ avant ou à l'étape } m) \\ &\qquad + \quad \mathbb{P}_\theta(\text{décider } H_1 \text{ avant ou à l'étape } m) = 1 . \end{aligned}$$

Soit $(X_1, \ldots, X_m)$ une séquence d'observations. On note

$$\begin{aligned} C_m &= \mathbb{P}_\theta\left((X_1, \ldots, X_m) \in D_m\right) \\ \mathbb{P}_m^0 &= \mathbb{P}_\theta\left((X_1, \ldots, X_m) \in D_m^0\right) \\ \text{et } \mathbb{P}_m^1 &= \mathbb{P}_\theta\left((X_1, \ldots, X_m) \in D_m^1\right). \end{aligned}$$

A l'étape 1, on a $C_1 + \mathbb{P}_1^0 + \mathbb{P}_1^1 = 1$. A l'étape 2, on obtient $C_2 + \mathbb{P}_2^0 + \mathbb{P}_2^1 = C_1$, puis d'une façon générale

$$\mathbb{P}_\theta\left((X_1, \ldots, X_m) \in D_m\right) + \sum_{n=1}^{m} \mathbb{P}_\theta\left((X_1, \ldots, X_n) \in D_n^0\right) + \sum_{n=1}^{m} \mathbb{P}_\theta\left((X_1, \ldots, X_n) \in D_n^1\right) = 1.$$

En faisant tendre m vers $+\infty$, on montre que le test s'arrête en un nombre fini d'étapes avec la probabilité 1. A la dernière étape, on obtient alors $\mathbb{P}_{m_0}^0 + \mathbb{P}_{m_0}^1 = C_{m_0-1}$.

Wald a introduit ensuite deux notions : la courbe d'efficacité, qui représente la probabilité d'accepter H_0 en fonction de θ, et le nombre moyen d'observations nécessaires pour terminer le test en fonction de θ (ou Average Sample Number, noté ASN). Wald a calculé diverses approximations de ces courbes.

Pour la densité exponentielle (VII.2), les inégalités (VII.3) s'écrivent

$$B < \left(\frac{\theta_0}{\theta_1}\right)^r \exp\left\{-\sum_{i=1}^{r} x_i \left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right)\right\} < A$$

où r est le nombre d'outages. A condition de poser $t = \sum_{i=1}^{r} x_i$ et $d = \frac{\theta_0}{\theta_1}$, cette double inégalité peut s'écrire

$$\frac{\ln B}{\ln d} + \frac{\left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right)}{\ln d} t < r < \frac{\ln A}{\ln d} + \frac{\left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right)}{\ln d} t.$$

Ces dernières inégalités correspondent exactement au test de Wald et supposent que la dernière observation soit allée jusqu'au temps exact où s'est produit l'outage. Or, si l'on écrit les inégalités en privilégiant le temps, on obtient

$$\frac{r \ln d - \ln A}{\frac{1}{\theta_1} - \frac{1}{\theta_0}} < t < \frac{r \ln d - \ln B}{\frac{1}{\theta_1} - \frac{1}{\theta_0}}.$$

Il est alors possible de prendre une décision dès que l'une des frontières est franchie et l'on obtient une version dite continue du TSRP. Lorsque ce test accepte H_0, il le fait en un temps inférieur à celui du TSRP de Wald. Mais il ne s'agit plus du test de Wald et certaines modifications sont nécessaires.

Les bornes A et B retenues sont alors (cf Dvoretsky et al. [1963]) :

$$A = \frac{(1-\beta)(d+1)}{2\alpha d} \quad \text{et} \quad B = \frac{\beta}{1-\alpha}. \tag{VII.4}$$

Bien que le test séquentiel s'arrête en un temps fini avec la probabilité 1, il peut aléatoirement nécessiter un nombre d'observations supérieur à celui d'un test classique. On se fixe alors un temps t_0 au delà duquel le test ne doit pas continuer, ainsi qu'un nombre r_0 d'outages qui ne doit pas être dépassé. En général, on se fixe r_0 comme étant la taille d'un test classique de niveau α et de puissance maximale $1-\beta$ pour tester H_0 contre H_1. D'après Epstein et Sobel [1955], cette valeur r_0 est le plus petit des entiers r tels que

$$\frac{\chi^2_{2r,\alpha}}{\chi^2_{2r,1-\beta}} \geqslant \frac{\theta_1}{\theta_0} = \frac{1}{d} \tag{VII.5}$$

où $\chi^2_{2r,\alpha}$ est le quantile d'ordre α d'un chi-deux à $2r$ degrés de liberté et le bon choix du temps de troncature est

$$t_0 = \frac{\theta_0}{2}\chi^2_{2r_0,\alpha} \ . \tag{VII.6}$$

En général, il est plus commode de travailler en temps standardisé t' qui s'évalue en multiple de θ_1. Le temps $t' = \frac{t}{\theta_1}$ est alors le temps standardisé correspondant au temps réel t. Puis, en posant

$$h'_0 = -\frac{\ln B}{1-\frac{1}{d}},\ h'_1 = \frac{\ln A}{1-\frac{1}{d}},\ s' = \frac{\ln d}{1-\frac{1}{d}},$$

la règle en temps standardisé s'écrit

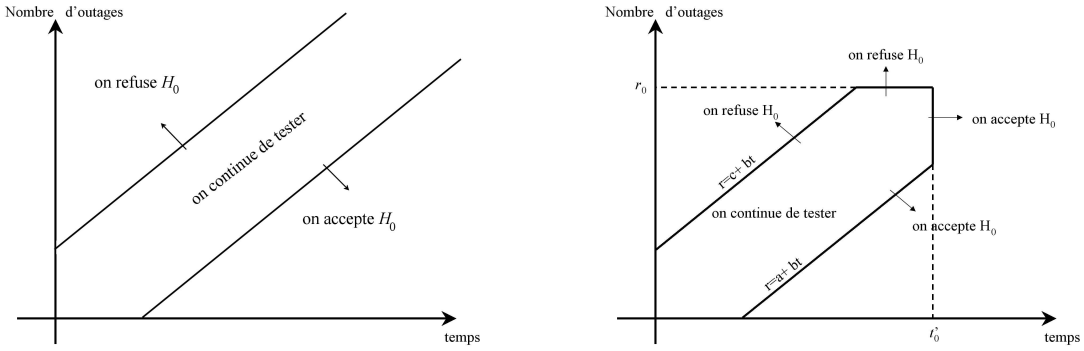
- si $t' \geqslant rs' + h'_0$, on accepte H_0,
- si $t' \leqslant rs' - h'_1$, on rejette H_0,
- si $rs' - h'_1 < t' < rs' + h'_0$, on continue d'observer l'équipement.

Afin de tenir compte de la troncature, cette règle est complétée par

- on accepte H_0 si le temps $t_0^{'}$ s'est écoulé sans atteindre r_0 outages,
- on accepte H_1 si r_0 outages se sont produits pendant une durée inférieure à $t_0^{'}$.

Dans le cas des tests tronqués, les approximations de la fonction d'efficacité et de la fonction ASN ne sont plus valables. Il est toutefois possible d'en calculer la valeur exacte à l'aide des probabilités $\mathbb{P}^0_m$, $\mathbb{P}^1_m$ et C_m, $m \geqslant 1$. Le calcul à chaque étape de ces probabilités en fonction de θ, s'appelle la méthode directe d'Aroian (cf Aroian [1968]).

Ci-dessous sont représentées graphiquement dans le plan (t',r) les frontières du TSRP dans le cas ordinaire et dans le cas tronqué. La bande de continuation est plus étroite dans le cas tronqué et surtout se termine obligatoirement avant ou en (t'_0, r_0).



(a) Test séquentiel classique

(b) Test séquentiel tronqué

FIG. VII.2 – Frontières du test séquentiel des rapports des probabilités.

1.4 Calcul d'un niveau de confiance après acceptation pour une loi exponentielle

Nous allons maintenant décrire les méthodes permettant d'obtenir les limites inférieures de confiance standardisées à $100\,(1-\gamma)\,\%$ et les niveaux de confiance lorsque l'hypothèse H_0 a été acceptée.

1.4.1 Probabilités de continuation

Nous construisons ici la V.A. qui va nous permettre d'exploiter les résultats du paragraphe 1.2. Bien que cette construction soit relativement naturelle, elle s'avère essentielle pour une traduction rigoureuse de l'équation (VII.1).

On note t'_{A_i} le temps standardisé d'acceptation après i outages et t'_{R_j} le temps standardisé de rejet après j outages. Ces temps sont en fait les abscisses des points d'intersection entre la frontière d'acceptation et la droite $r = i$ pour le premier et entre la frontière de rejet et la droite $r = j$ pour le second. Soient T_A la V.A. représentant le temps standardisé où le test conduit à accepter l'équipement et $N(T_A)$ le nombre d'outages survenus pendant le temps d'observation T_A. Ainsi T_A est une variable aléatoire discrète qui prend les valeurs t'_{A_i} pour i variant de 0 à $r_0 - 1$ lorsque le test se termine par une acceptation et nous lui affectons la valeur $+\infty$ lorsque le test se termine par un rejet.

La fonction $N(\cdot)$ représente la fonction de comptage du processus de Poisson homogène de la séquence des outages. Pour tout $t > 0$, $N(t)$ suit une loi de Poisson de paramètre $\frac{t}{\theta}$. On estime alors $\theta' = \frac{\theta}{\theta_1}$ par la V.A.R.

$$Y = \frac{T_A}{N(T_A)} \mathbb{1}_{[T_A < +\infty]}.$$

Ceci revient à estimer θ' par $\frac{t'_{A_i}}{i}$ quand le test se termine par une acceptation au temps t'_{A_i} après i outages et par 0 si le test se termine par un rejet. Notons $\theta'_{L,\gamma,i}$ la borne inférieure de confiance à $100(1-\gamma)\,\%$ de θ' lorsque le test s'est terminé par une acceptation avec i outages. La borne inférieure $\theta_{L,\gamma,i}$ de confiance à $100(1-\gamma)\,\%$ de θ est égale à $\theta_1 \theta'_{L,\gamma,i}$.

D'après l'équation (VII.1), cette borne inférieure de confiance est solution de

$$\sum_{s/\frac{t'_{A_s}}{s} \geqslant \frac{t'_{A_i}}{i}} \mathbb{P}\left(Y = \frac{t'_{A_s}}{s}; \theta'_{L,\gamma,i}\right) + \mathbb{P}\left(Y = 0; \theta'_{L,\gamma,i}\right) \mathbb{1}_{\left\{0 \geqslant \frac{t'_{A_i}}{i}\right\}} = \gamma, \tag{VII.7}$$

où $\mathbb{P}\left(D; \theta'\right)$ est la probabilité de l'événement D lorsque $\theta'\theta_1$ est la vraie valeur du paramètre. Il est facile de montrer que la séquence $\left(\frac{t'_{A_s}}{s}\right)_s$ est décroissante et que

$$\left\{Y = \frac{t'_{A_s}}{s}\right\} = \left\{\left[N(T_A) = s\right] \ \cap \ \left[T_A = t'_{A_s}\right]\right\} \ .$$

L'équation (VII.7) s'écrit alors

$$\gamma = \sum_{s=0}^{i} \mathbb{P}\left(\left(N(T_A), T_A\right) = \left(s, t'_{A_s}\right); \theta'_{L,\gamma,i}\right),$$

où $\mathbb{P}\left(\left(N(T_A), T_A\right) = \left(s, t'_{A_s}\right); \theta'\right)$ désigne la probabilité que le test se termine par une acceptation au temps réel $t'_{A_s} \times \theta_1$ lorsque $\theta' \times \theta_1$ est la vraie valeur du paramètre. Afin de simplifier les notations, nous poserons pour la suite

$$\mathbb{P}\left(\left(N(T_A), T_A\right) = \left(s, t'_{A_s}\right); \theta'_{L,\gamma,i}\right) = \mathbb{P}\left(\left(s, t'_{A_s}\right); \theta'_{L,\gamma,i}\right) \ .$$

Proposition 3 Soit $\left(t'_{(k)}\right)_k$ une suite de temps de fin de test standardisés, distincts et rangés dans l'ordre croissant telle que $t'_{(0)} = 0$ et $t'_{(j)} = t'_0$. On a :

$$\mathbb{P}\left(\left(i, t'_{(k)}\right); \theta'\right) = e^{-\frac{t'_{(k)}}{\theta'}} \left(\frac{1}{\theta'}\right)^i c'\left(i, t'_{(k)}\right), \tag{VII.8}$$

où les $c'\left(i, t'_{(k)}\right)$ sont des coefficients indépendants de θ'.

Preuve de la proposition 3
Nous reprenons ici les notations de Illig [2001]. On pose pour $l = 1, \ldots, j$,

$$\Delta_l = t'_{(l)} - t'_{(l-1)}.$$

On a alors

$$\left]0, t'_{(k)}\right] = \cup_{l=1}^{k} \left]t'_{(l-1)}, t'_{(l)}\right] \text{ et } \sum_{l=1}^{k} \Delta_l = t'_{(k)}.$$

Soit $\left(t'_{(k)}, i\right)$, $k \leqslant j$, *un point de la région de continuation ou de la frontière d'acceptation. Pour une valeur* θ' *du paramètre standardisé, la probabilité que* δ_l *outages surviennent dans l'intervalle de temps* $\left]t'_{(l-1)}, t'_{(l)}\right]$ *est égale à*

$$\mathbb{P}\left(\delta_l \text{ outages dans } \left]t'_{(l-1)}, t'_{(l)}\right]; \theta'\right) = \mathbb{P}\left(N\left(t'_{(l)}\theta_1\right) - N\left(t'_{(l-1)}\theta_1\right) = \delta_l; \theta'\right).$$

Or $N\left(t'_{(l)}\theta_1\right) - N\left(t'_{(l-1)}\theta_1\right)$ *suit une loi de Poisson de paramètre* $\frac{\Delta_l}{\theta'}$. *On obtient donc*

$$\boxed{\mathbb{P}\left(\delta_l \text{ outages dans } \left]t'_{(l-1)}, t'_{(l)}\right]; \theta'\right) = e^{-\frac{\Delta_l}{\theta'}} \frac{\left(\frac{\Delta_l}{\theta'}\right)^{\delta_l}}{\delta_l!}.}$$

Si maintenant on considère un k-uplet $(\delta_1, \ldots, \delta_k)$ *d'entiers positifs tels que* $\sum_{l=1}^{k} \delta_l = i$ *et ne conduisant pas à une terminaison du test avant* $t'_{(k)}$ *alors la probabilité qu'il y ait i outages pendant le temps* $t'_{(k)}$ *avec pour chaque l,* δ_l *outages dans l'intervalle* $\left]t'_{(l-1)}, t'_{(l)}\right]$ *sans que le test ne se termine avant* $t'_{(k)}$ *est égale à*

$$\begin{aligned}
& \mathbb{P}\left((\delta_1, \ldots, \delta_k), \text{ pas de terminaison avant } t'_{(k)}; \theta'\right) \\
= \; & \mathbb{P}\left(\forall l = 1, \ldots, k, \ \delta_l \text{ outages dans } \left]t'_{(l-1)}, t'_{(l)}\right]; \theta'\right) \\
= \; & \mathbb{P}\left(\forall l = 1, \ldots, k, \ N\left(t'_{(l)}\theta_1\right) - N\left(t'_{(l-1)}\theta_1\right) = \delta_l; \theta'\right).
\end{aligned}$$

Par indépendance des accroissements de N, on obtient alors

$$\begin{aligned}
& \mathbb{P}\left((\delta_1, \ldots, \delta_k), \text{ pas de terminaison avant } t'_{(k)}; \theta'\right) \\
= \; & \prod_{l=1}^{k} e^{-\frac{\Delta_l}{\theta'}} \frac{\left(\frac{\Delta_l}{\theta'}\right)^{\delta_l}}{\delta_l!} \\
= \; & \left[\prod_{l=1}^{k} e^{-\frac{\Delta_l}{\theta'}} \left(\frac{1}{\theta'}\right)^{\delta_l}\right] \left[\prod_{l=1}^{k} \frac{(\Delta_l)^{\delta_l}}{\delta_l!}\right] \\
= \; & e^{-\frac{t'_{(k)}}{\theta'}} \left(\frac{1}{\theta'}\right)^{i} \prod_{l=1}^{k} \frac{(\Delta_l)^{\delta_l}}{\delta_l!}.
\end{aligned}$$

Par conséquent, la probabilité que i outages surviennent pendant un temps $t'_{(k)}$ *pour une valeur* θ' *du paramètre standardisé est*

$$\begin{aligned}
\mathbb{P}\left(\left(i, t'_{(k)}\right); \theta'\right) &= \sum_{S} \mathbb{P}\left((\delta_1, \ldots, \delta_k), \text{ pas de terminaison avant } t'_{(k)}; \theta'\right) \\
&= e^{-\frac{t'_{(k)}}{\theta'}} \left(\frac{1}{\theta'}\right)^{i} \sum_{S} \prod_{l=1}^{k} \frac{(\Delta_l)^{\delta_l}}{\delta_l!},
\end{aligned}$$

où S désigne l'ensemble des k-uplets tels que $\sum_{l=1}^{k} \delta_l = i$ *et ne conduisant pas à une terminaison du test avant l'instant* $t'_{(k)}$.
En posant,

$$c'\left(i, t'_{(k)}\right) = \sum_{S} \prod_{l=1}^{k} \frac{(\Delta_l)^{\delta_l}}{\delta_l!},$$

on obtient la relation voulue. ■

L'équation (VII.8) nous permet d'obtenir les coefficients $c'\left(i, t'_{(k)}\right)$ en fonction de $\mathbb{P}\left(\left(i, t'_{(k)}\right); \theta'\right)$ que l'on sait calculer lorsque θ' est fixé (cf section suivante). On est alors en mesure d'écrire l'équation (VII.7) sous la forme suivante :

$$\boxed{\gamma = \sum_{s=0}^{i} c'\left(s, t'_{A_s}\right) e^{-\frac{t'_{A_s}}{\theta'_{L,\gamma,i}}} \left(\frac{1}{\theta'_{L,\gamma,i}}\right)^s .} \qquad \text{(VII.9)}$$

1.4.2 Calcul des coefficients $c'(i, t'_{(k)})$

Nous calculons les probabilités $\mathbb{P}\left(\left(i, t'_{(k)}\right); \theta'\right)$ pour la valeur $\theta' = 1$. Nous en déduisons alors les coefficients $c'\left(i, t'_{(k)}\right)$ par l'équation

$$c'\left(i, t'_{(k)}\right) = e^{t'_{(k)}}\, \mathbb{P}\left(\left(i, t'_{(k)}\right); 1\right) .$$

Les paramètres α, β et d ayant été fixés, il est possible de calculer les droites de confiance ainsi que r_0 et t'_0. Les temps standardisés correspondant aux intersections des lignes d'outages avec les frontières d'acceptation et de rejet sont ensuite calculés et rangés en ordre croissant $t'_{(1)} \leqslant t'_{(2)} \leqslant \cdots \leqslant t'_0$. On complète alors le quadrillage par des verticales passant par ces points d'intersection. Apparaissent alors dans la bande de continuation des points spécifiques de coordonnées $\left(t'_{(j)}, i\right)$, appelés points de continuation.

Le calcul des probabilités de continuation pour la valeur $\theta' = 1$ se fait à l'aide d'une procédure récursive.

La $j^{\text{ème}}$ étape de cette procédure consiste à calculer aux points d'abscisse $t'_{(j)}$ les probabilités en question. Pour cela, on énumère pour chacun de ces points les trajectoires qui y aboutissent (cf Fig.VII.3), afin de calculer leurs probabilités. On part des points d'abscisses $t'_{(j-1)}$ pour lesquels on a fait le calcul des probabilités de continuation à l'étape $j-1$ et on calcule le nombre d'outages nécessaires pour parvenir en $\left(t'_{(j)}, i\right)$.

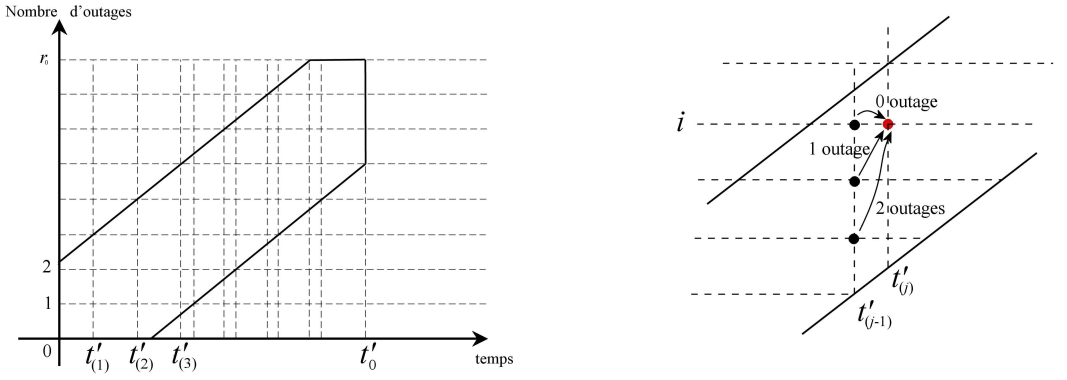


FIG. VII.3 – Etude des trajectoires

Parmi les points de coordonnées $\left(t'_{(j-1)}, k\right)$, $k \leqslant i$, on sélectionne ceux qui appartiennent à la bande de continuation et on note $\mathcal{K}$ l'ensemble de ces points. La probabilité de continuation

en $\left(t'_{(j)}, i\right)$ s'obtient alors par la formule :

$$\mathbb{P}\left(\left(i, t'_{(j)}\right); 1\right) = \sum_{k \in \mathcal{K}} \mathbb{P}\left(\left(k, t'_{(j-1)}\right); 1\right) e^{-\left(t'_{(j)} - t'_{(j-1)}\right)} \frac{\left(t'_{(j)} - t'_{(j-1)}\right)^{i-k}}{(i-k)!} \quad .$$

L'initialisation de cette procédure consiste à calculer les probabilités de continuation aux points de coordonnées $\left(t'_{(1)}, k\right)$. On peut voir que cela représente en fait la probabilité d'avoir k outages pendant un temps total $t'_{(1)}$.

On peut ensuite calculer les probabilités d'acceptation aux points de coordonnées $\left(t'_{A_i}, i\right)$ pour $i = 1, \ldots, r_0$. Pour chaque t'_{A_i}, il existe un l_i tel que $t'_{A_i} = t'_{(l_i)}$ et on a

$$\mathbb{P}\left(\left(i, t'_{A_i}\right); 1\right) = \mathbb{P}\left(\left(i, t'_{(l_i)}\right); 1\right) = \mathbb{P}\left(\left(i, t'_{(l_i-1)}\right); 1\right) e^{-\left(t'_{(l_i)} - t'_{(l_i-1)}\right)} \quad .$$

1.4.3 Borne inférieure de confiance

Une fois l'équipement accepté, on cherche à savoir quelle est la borne inférieure de confiance à $100\,(1-\gamma)\,\%$ pour θ au sens de la section 1.2. On est donc amené à résoudre l'équation (VII.9) en θ où γ est fixé par l'utilisateur.

Plusieurs méthodes de résolution numérique sont disponibles (Point fixe, Newton, etc). Nous utilisons ici la méthode de bisection (cf Nougier [1987]).

1.4.4 Niveau de confiance

Ce qui intéresse les ingénieurs, une fois l'équipement accepté, est de savoir quelle confiance ils peuvent avoir dans le fait que le temps moyen entre deux outages dépasse une valeur seuil qu'ils considèrent comme minimale pour la sécurité du fonctionnement des ILS. On retrouve donc la notion de niveau de confiance telle que nous l'avons introduite à la section 1.2.

Ce problème est le problème dual du précédent. C'est la même équation (VII.9) qui va servir, mais au lieu de la résoudre en θ' on va la résoudre en γ. On effectue donc la somme de i termes dépendant chacun des coefficients $c'\left(s, t'_{A_s}\right)$ dont le calcul est rappelé en 1.4.2.

1.5 Etude de cas

Les 21 valeurs ci-dessous[1] représentent les temps observés de bon fonctionnement entre deux outages. Elles concernent un localiser expérimental.

Ces données peuvent surprendre par leur variabilité : on trouve un temps de bon fonctionnement de 3 heures faisant suite à un fonctionnement sans interruption de plus d'une année !

175	1505	4488	2382	16	1165	244
9240	3	6126	4708	3148	3447	5
3554	388	7578	5073	215	2622	2802

Tab. VII.1 – Temps réel, en heures, de bon fonctionnement entre deux outages

Nous savons que le localiser peut détecter des intrus (oiseaux, personnes, véhicules s'interposant entre l'appareil et l'avion) déclenchant alors une interruption de fonctionnement. Toutes les interruptions ne sont donc pas de même nature et ne nécessitent pas le réarmement de l'appareil. Dans le cas présent, nous ignorons la nature réelle des interruptions.

Trois tests séquentiels ont été effectués correspondant à trois types de localisers

[1] These data have been provided by LVNL -Air Traffic Control the Netherlands- and used with their kind permission.

Ces données ont été fournies par LVNL -Air Traffic Control the Netherlands- et sont utilisées avec leur aimable autorisation.

- Localiser de catégorie III, MTBO souhaité > 4000 heures,
- Localiser de catégorie II, MTBO souhaité > 2000 heures,
- Localiser de catégorie I, MTBO souhaité > 1000 heures.

On constate que le problème initial concerne des hypothèses composites. Comme on le fait souvent, afin de conserver la simplicité de la procédure, la norme américaine (cf Department of Defense (USA) [1998]) conseille de se ramener à des hypothèses simples. Il est alors d'usage de fixer θ_1 à la valeur minimale souhaitée pour le MTBO et θ_0 à 2 fois θ_1. Ceci évite de qualifier un appareil dont le MTBO serait trop proche de l'objectif minimal.

D'autre part, pour des raisons évidentes de sécurité, il est préférable d'observer les ILS pendant une durée minimale. Celle-ci est en général égale à un an car le fonctionnement des ILS est aussi lié aux conditions climatiques.

Dans tous les cas, nous avons pris $\alpha = \beta = 0,1$ et $d = 2$. Pour le dernier test (localiser de catégorie I), nous avons également envisagé le test avec un temps minimum d'observations de un an.

Les plans de test, les trajectoires correspondantes ainsi que les niveaux de confiance que θ' dépasse 1 pour chacune des acceptations sont fournis figures VII.4, VII.5, VII.6 et VII.7.

La figure VII.4 correspond au test de $\theta_0 = 8000$ contre $\theta_1 = 4000$. La qualification est refusée. La figure VII.5 correspond au test de $\theta_0 = 4000$ contre $\theta_1 = 2000$. Le test se termine par un refus au bout de 9975 heures et 7 outages. La trajectoire complète du processus a cependant été tracée. Les figures VII.6 et VII.7 concernent le test de $\theta_0 = 2000$ contre $\theta_1 = 1000$, avec un temps minimum d'observation de un an dans le dernier cas.

Dans ces deux derniers cas, la qualification est obtenue après sept outages. Le niveau de confiance que θ soit supérieur à 1000 est de $92,48\%$ dans le premier cas et de $94,21\%$ dans le second. Notons que pour tous ces plans, les valeurs de α, β et d sont les mêmes. Il est donc normal que les valeurs de troncatures en multiples de θ_1 soient les mêmes et que les niveaux de confiance que θ dépasse θ_1 (ou que θ' dépasse 1) soient identiques pour les trois premiers plans.

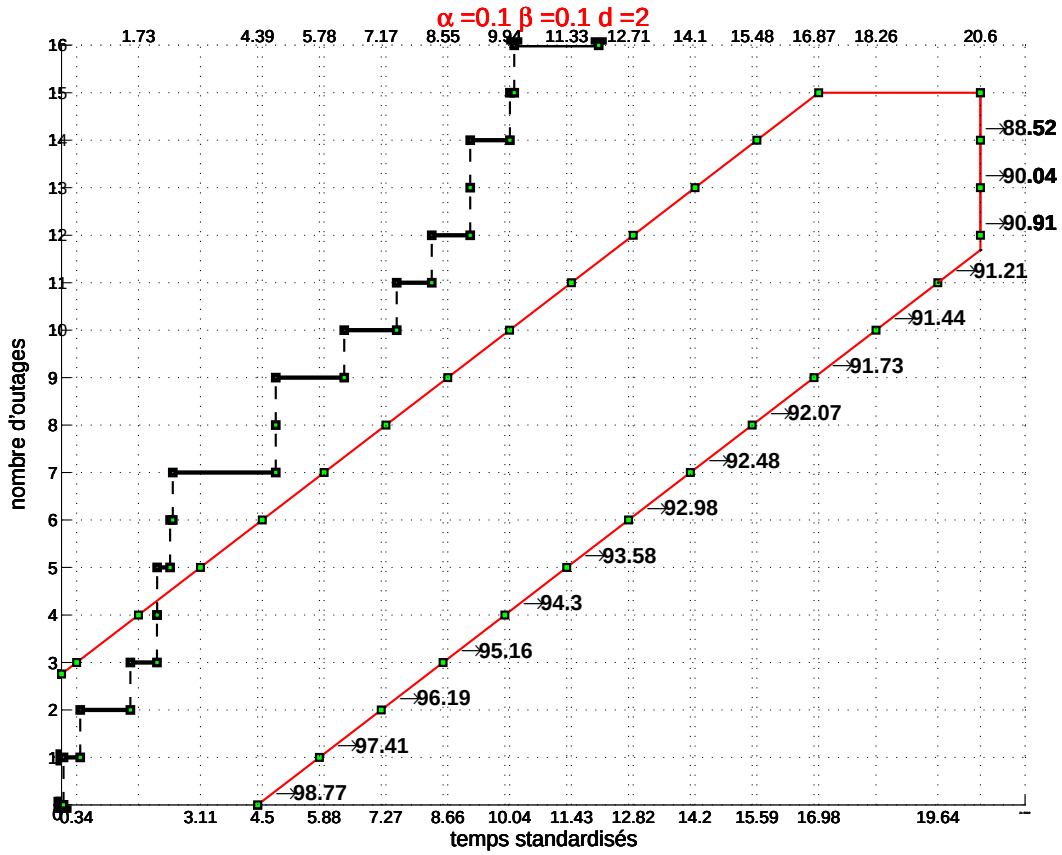


FIG. VII.4 – Plan du test séquentiel pour un localiser de catégorie III avec $\theta_1 = 4000$ heures.

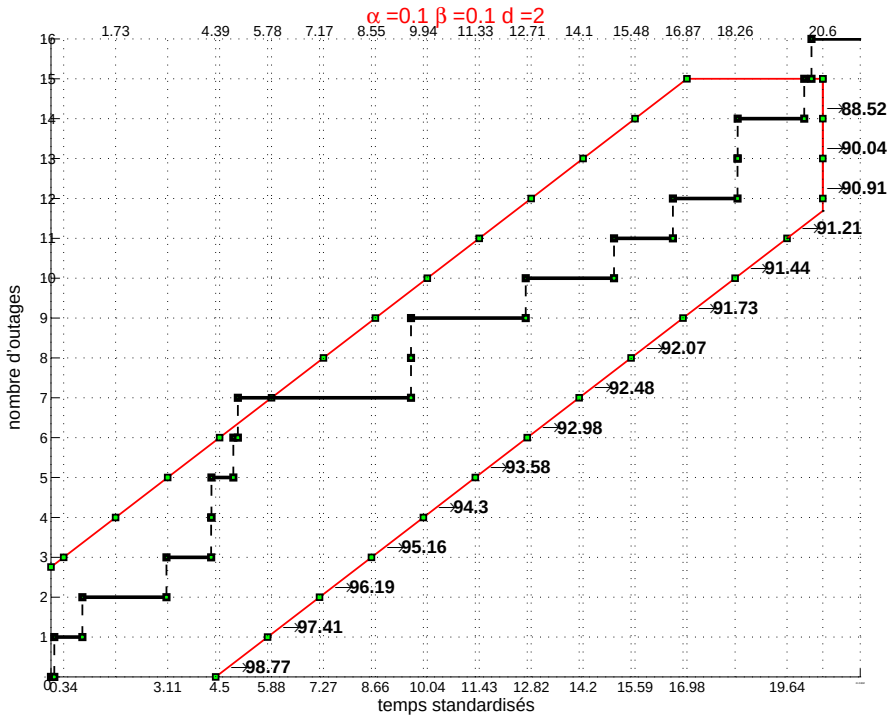


FIG. VII.5 – Plan du test séquentiel pour un localiser de catégorie II avec $\theta_1 = 2000$ heures.

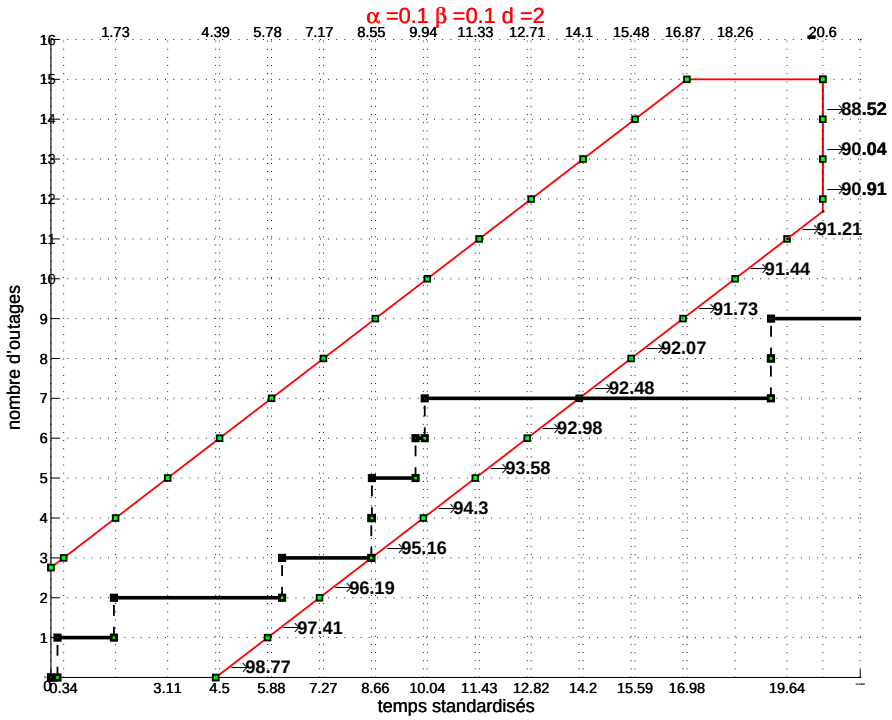


FIG. VII.6 – Plan du test séquentiel pour un localiser de catégorie I avec $\theta_1 = 1000$ heures.

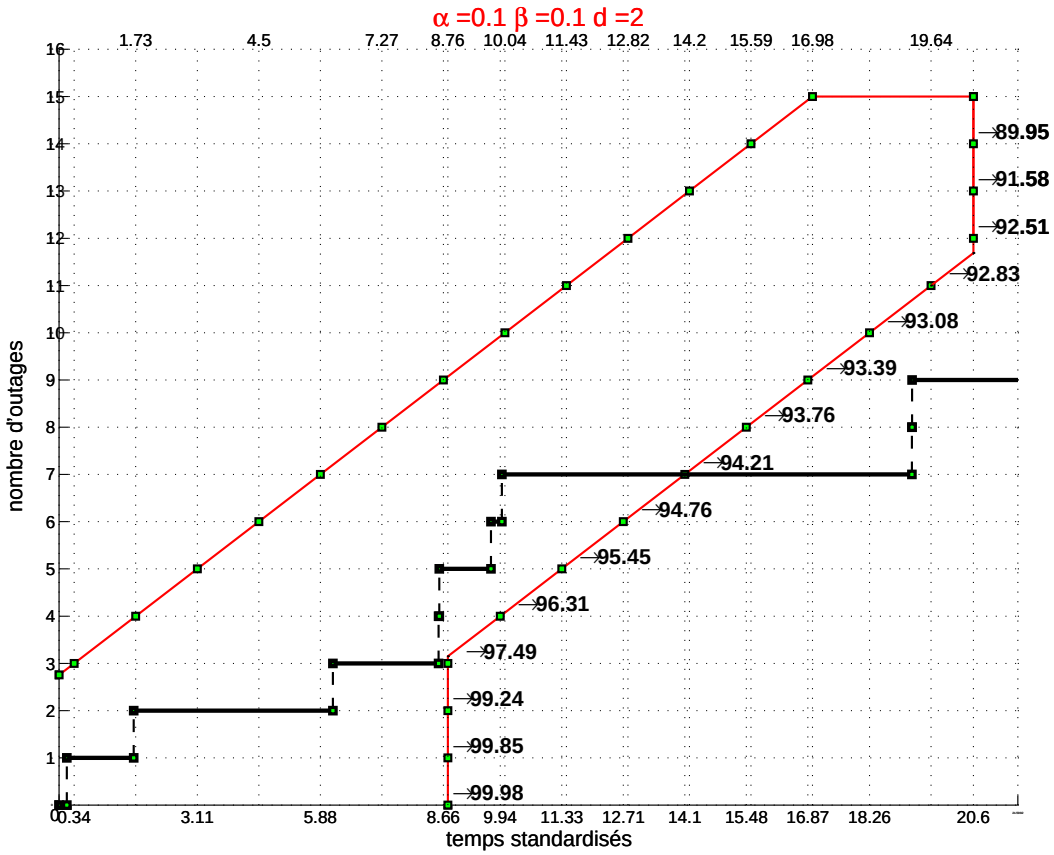


FIG. VII.7 – Plan du test séquentiel pour un localiser de catégorie I avec $\theta_1 = 1000$ heures et un minimum de un an d'observation.

Pour valider les calculs ci-dessus, il faut s'assurer de l'hypothèse de loi exponentielle. Dans le cas présent, la valeur obtenue lors du test du χ^2 conduit à une probabilité excédentaire autour de 8%. L'hypothèse de loi exponentielle est donc acceptée avec des réserves. Notons que lorsque les valeurs 3, 5 et 16 sont écartées, l'ajustement est quasi-parfait. Il est possible que ces données correspondent à la détection d'un intrus. Les temps correspondants pourraient donc être ajoutés aux temps suivants. En l'absence d'information à ce sujet, nous nous sommes refusés à le faire, ce qui aurait changé les réponses aux différents tests.

Notons que l'intervalle de confiance symétrique à 90% sur θ donne $[2034; 4232]$. La borne supérieure est trop peu élevée pour permettre l'acceptation d'un localiser de catégorie II.

Une question se pose alors avec insistance : comment peut-on prendre des décisions aussi importantes avec si peu d'observations ? Dans le cas présent, ceci représente en temps de test une durée minimale de un an et maximale de plusieurs années. Notons que paradoxalement, un appareil parfait ne fournirait aucune observation puisqu'il n'y aurait pas d'interruptions. Sous réserve que les observations suivent bien une loi exponentielle (hypothèse qu'il faut vérifier, avec peut-être des observations en plus grand nombre !), les résultats obtenus sont rigoureux. Il faut toutefois garder à l'esprit la façon dont s'envisage un intervalle de confiance et donc un niveau de confiance.

Pour conclure, précisons qu'il est possible de calculer, de façon analogue, des bornes supérieures après acceptation (cf Bryant et Schmee [1979]) ou après rejet, et donc des intervalles de confiance ainsi que des niveaux de confiance au sens de la section 1.2. Une fois le test effectué, le problème qui se pose quelle que soit la réponse est de savoir que faire du localiser. Nous pensons que ces niveaux de confiance peuvent aider à prendre les décisions.

Les travaux ci-dessus ont été réalisés à l'aide de programmes en Matlab qui seront disponibles sur le web à l'adresse www.enseeiht.fr/len7/index.html.

Remerciements : les auteurs remercient Pierre Cazes et le comité de rédaction pour leur lecture attentive et leurs nombreuses suggestions qui ont permis la rédaction actuelle de ce travail. Ils remercient également Véronique Font membre de la société IXI et Philippe Crébassa pour son accueil au STNA.

Références

Aroian L. A. (1968). *Sequential analysis, direct method.* Technometrics, vol. 10. pages 125–132.

Bryant L. et Schmee J. (1979). *Confidence limits on MTBF for Sequential Test Plans of MIL-STD 781.* Technometrics, vol. 21. pages 33–42.

Department of Defense (USA) (4 1998), *Handbook for reliability test methods, plans, and environments for engineering, development, qualification, and production.* Rapport technique.

Dvoretsky A., Kiefer J. et Wolfowitz J. (1963). *Sequential decision problem for processus with continuous time parameter. testing hypothesis.* Ann. Math. Stat., vol. 24. pages 254–264.

Epstein B. et Sobel M. (1955). *Sequential life tests in the exponential case.* Ann. Math. Statist., vol. 26. pages 82–93.

Ghosh B. K. (1970). *Sequential tests of statistical hypotheses.* Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont.

Illig A. (2001). *Probabilités de confiance après décision lors d'un test séquentiel avec un temps de bon fonctionnement de loi exponentielle.* Mémoire de DEA sous la direction du professeur B. Garel, Université Toulouse III.

Lecoutre B. (1997). *C'est bon à savoir ! et si vous êtiez un bayésien qui s'ignore.* Modulab, INRIA, vol. 8, n°18. pages 81–87.

Mood A. et Graybill F. (1963). *Introduction to the theory of statistics.* Mc Graw-Hill, 2 éd.

Nougier J. P. (1987). *Méthodes de calcul numérique.* Mathématiques pour la Physique. [Mathematics for Physics]. Masson, 3 éd.

Siegmund D. (1985). *Sequential analysis : tests and confidence intervals.* Springer-Verlag.

Tassi P. (1985). *Méthodes statistiques.* Economica.

Wald A. (1947). *Sequential Analysis.* John Wiley & Sons, New York.

2 Programme Matlab : note technique

Dans cette section, nous expliquons comment se servir de l'interface graphique que nous avons créée. Le programme Matlab est accessible sur le web à l'adresse suivante : `http ://www.enseeiht.fr/len7/`. Pour tous les résultats théoriques, nous renvoyons à la section précédente.

Après exécution du programme `testplan.m`, nous obtenons la figure VII.8. Pour obtenir le plan du test séquentiel, il y a alors 3 étapes à effectuer.

Dans la partie «Valeurs initiales», 4 valeurs sont à spécifier : `alpha` qui correspond au risque de première espèce, `beta` qui correspond au risque de seconde espèce, `teta0` qui correspond à la

valeur du paramètre sous l'hypothèse H_0 et `d` qui est égal à $\frac{\theta_0}{\theta_1}$, où θ_1 est la valeur du paramètre sous l'hypothèse H_1.

Dans la partie «Type du test», nous pouvons préciser si un temps minimum est souhaité ou non. La figure VII.9 montre le cas où l'utilisateur désire un temps minimum. On peut voir que le temps minimum peut être spécifié selon trois unités de mesure (en nombre de mois, en nombre d'heures ou en multiple de θ_1, appelé temps standardisé).

L'équation (VII.9), page 123, nous donne la relation entre γ (niveau de confiance) et θ' (borne inférieure de confiance). Dans la partie «Type du test», nous pouvons alors préciser le résultat voulu. Soit l'utilisateur veut calculer les niveaux de confiance : il doit alors préciser la valeur de `tetaprime`. Soit il veut calculer les bornes inférieures de confiance : auquel cas, il doit préciser la valeur de `gamma`. Par défaut, le programme calcule les niveaux de confiance (cf figure VII.8). Il suffit de cliquer sur `Borne inférieure de confiance` pour rentrer la valeur de `gamma` (cf figure VII.10)

Les paramètres de troncature i_0 et t_0, calculés automatiquement grâce aux valeurs initiales (données par les équations (VII.5) et (VII.6), page 120) peuvent être précisés manuellement par l'utilisateur. Pour cela, il suffit de cliquer sur le paramètre à modifier (cf figure VII.11). Remarquons que la valeur de `t0`, comme pour le temps minimum, peut être spécifiée selon les trois unités de mesure (mois/heure/standardisé).

Lorsque chaque partie a été remplie par l'utilisateur, celui-ci peut cliquer sur `Executer` afin que le programme se lance. Le graphique des résultats présente alors le plan du test séquentiel ainsi que les résultats (niveau de confiance ou borne de confiance) calculés pour chaque acceptation (de 0 outage à i_0 outages).

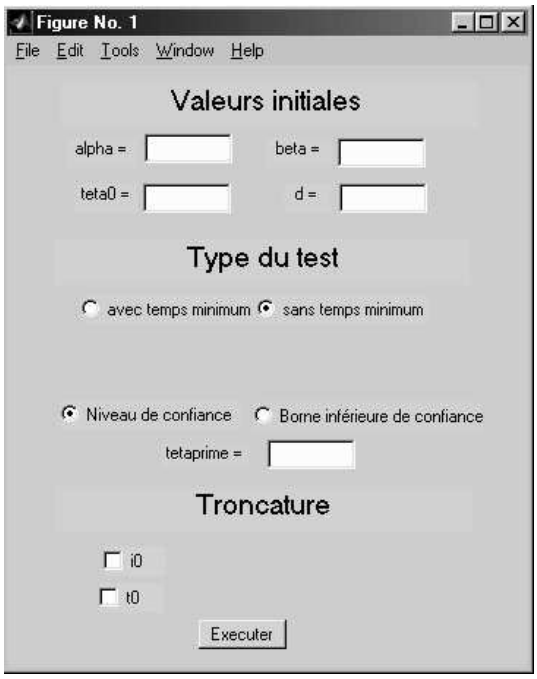


FIG. VII.8 – Programme testplan : interface principale.

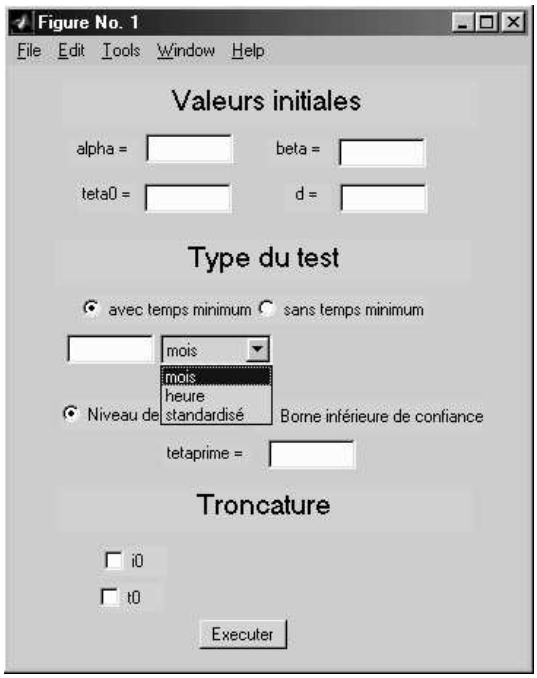


FIG. VII.9 – Programme testplan : avec temps minimum.

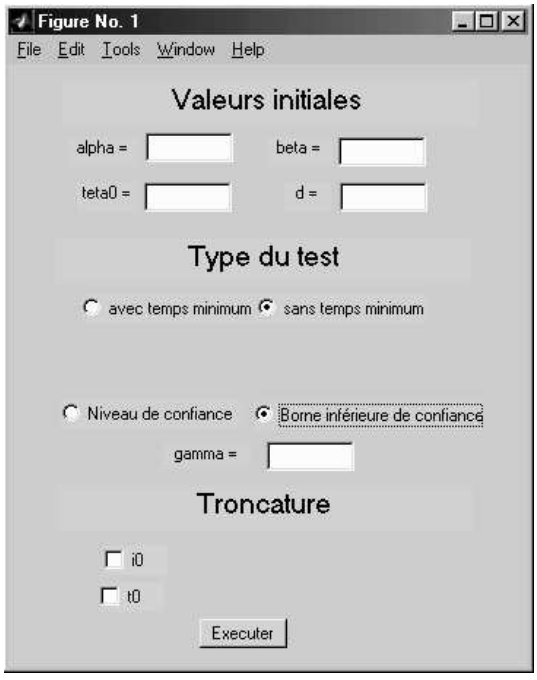


FIG. VII.10 – Programme testplan : borne inférieure de confiance.

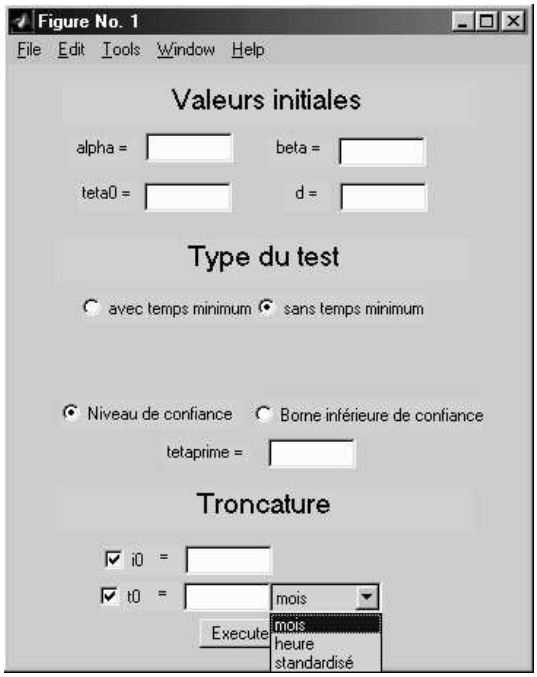


FIG. VII.11 – Programme testplan : troncature.

Bibliographie

Aroian L. A. (1968). *Sequential analysis, direct method.* Technometrics, vol. 10. pages 125–132.

Banfield J. D. et Raftery A. E. (1993). *Model-based Gaussian and non-Gaussian clustering.* Biometrics, vol. 49, n°3. pages 803–821.

Berdaï A. et Garel B. (1996). *Detecting a univariate normal mixture with two components.* Statistics and Decision, vol. 14. pages 35–51.

Bickel P. et Chernoff H. (1995). *Asymptotic distributions of the likelihood ratio statistic in a prototypical non regular problem.* In *Statistics and Probability : a Raghu Raj Bahadur Festchrift (eds J.K. Ghosh, S.K. Mitra, K.R. Parthasaraty and B.L.S. Prakasa Rao)*, New York. Wiley. pages 83–96.

Biernacki C., Celeux G. et Govaert G. (2000). *Assessing a mixture model for clustering with the integrated classification likelihood.* IEEE Transactions on Pattern Analysis an Machine Intelligence, vol. 22. pages 719–725.

Billingsley P. (1968). *Convergence of probability measures.* John Wiley & Sons Inc.

Bryant L. et Schmee J. (1979). *Confidence limits on MTBF for Sequential Test Plans of MIL-STD 781.* Technometrics, vol. 21. pages 33–42.

Celeux G. (1998). *Bayesian inference for mixtures : The label switching problem.* In Compstat98, R. Payne and P. Green (Eds), Physica-Verlag. pages 227–232.

Celeux G. et Govaert G. (1995). *Gaussian parsimonius clustering models.* Pattern recognition, vol. 28. pages 781–793.

Celeux G., Hurn M. et Robert C. P. (2000). *Computational and inferential difficulties with mixture posterior distributions.* J. Amer. Statist. Assoc. (to appear).

Chen H. et Chen J. (2001). *Large sample distribution of the likelihood ratio test for normal mixtures.* Statist. Probab. Lett., vol. 52, n°2. pages 125–133.

Chen M.-H., Shao Q.-M. et Ibrahim J. G. (2000). *Monte Carlo methods in Bayesian computation.* Springer Series in Statistics. Springer-Verlag.

Cheng R. C. H. et Traylor L. (1995). *Non-regular maximum likelihood problems.* J. Roy. Statist. Soc. Ser. B, vol. 57, n°1. pages 3–44.

Chernoff H. (1954). *On the distribution of the likelihood ratio.* Ann. Math. Statistics, vol. 25. pages 573–578.

Cowles M. K. et Carlin B. P. (1996). *Markov chain Monte Carlo convergence diagnostics : a comparative review.* J. Amer. Statist. Assoc., vol. 91, n°434. pages 883–904.

Craigmile P. F. et Titterington D. M. (1997). *Parameter estimation for finite mixtures of uniform distributions.* Comm. Statist. Theory Methods, vol. 26, n°8. pages 1981–1995.

Dacunha-Castelle D. et Gassiat E. (1997). *The estimation of the order of a mixture model.* Bernoulli, vol. 3, n°3. pages 279–299.

Davies R. B. (1977). *Hypothesis testing when a nuisance parameter is present only under the alternative.* Biometrika, vol. 64, n°2. pages 247–254.

Delmas C. (2001). *Distribution of the maximum of a random field and statistical applications (in french).* Université Paul Sabatier, Toulouse III, France.

Dempster A. P., Laird N. M. et Rubin D. B. (1977). *Maximum likelihood from incomplete data via the EM algorithm.* J. Roy. Statist. Soc. Ser. B, vol. 39, n°1. pages 1–38.

Department of Defense (USA) (4 1998), *Handbook for reliability test methods, plans, and environments for engineering, development, qualification, and production.* Rapport technique.

d'Estampes L., Garel B. et Saint Pierre G. (2003). *Test séquentiel : Niveau de confiance après acceptation.* Revue de Statistiques Appliquées, A paraître.

Diaconis P. et Ylvisaker D. (1979). *Conjugate priors for exponential families.* Ann. Statist., vol. 7, n°2. pages 269–281.

Diebolt J. et Robert C. P. (1994). *Estimation of finite mixture distributions through bayesian sampling.* Journal of the Royal Statistical Society B, vol. 56. pages 363–375.

Dvoretsky A., Kiefer J. et Wolfowitz J. (1963). *Sequential decision problem for processus with continuous time parameter. testing hypothesis.* Ann. Math. Stat., vol. 24. pages 254–264.

Epstein B. et Sobel M. (1955). *Sequential life tests in the exponential case.* Ann. Math. Statist., vol. 26. pages 82–93.

Escobar M. D. et West M. (1995). *Bayesian density estimation and inference using mixtures.* J. Amer. Statist. Assoc., vol. 90, n°430. pages 577–588.

Everitt B. (1981). *A monte carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distridutions.* Multivariate Behavioral Research, vol. 16. pages 171–180.

Garel B. (1996). *Théorie asymptotique du test du rapport des vraisemblances d'un mélange à deux composants.* C. R. Acad. Sci. Paris Sér. I Math., vol. 323, n°2. pages 199–202.

Garel B. (2001). *Likelihood ratio test for univariate Gaussian mixture.* J. Statist. Plann. Inference, vol. 96, n°2. pages 325–350.

Garel B. et Goussanou F. (2002). *Removing separation conditions in a 1 against 3-components gaussian mixture problem.* In *Classification, Clustering and data Analysis, Sokolowski A. and Boch H.H.(Eds)*, Berlin. Springer. pages 61–73.

Garel B. (2003). Asymptotic theory of the likelihood ratio test for the identification of a mixture, soumis. 2003.

Gentle J. E. (2002). *Elements of computational statistics.* Statistics and Computing. Springer-Verlag.

Ghosh B. K. (1970). *Sequential tests of statistical hypotheses.* Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont.

Ghosh J. et Sen P. (1985). *On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results.* In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II, Le Cam, L.M., Olshen, R.A. (Eds)*, Wadsworth Statist./Probab. Ser., Monterey. Wadsworth. pages 789–806.

Green P. (1995). *Reversible jump mcmc computation and bayesian model determination.* Biometrika, vol. 82. pages 711–732.

Gupta A. et Nagar D. (2000). *Matrix variate distributions.* Monographs and surveys in pure and applied mathematics Vol 104. Chapman & Hall.

Hartigan J. A. (1985). *A failure of likelihood asymptotics for normal mixtures.* In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II, Le Cam, L.M., Olshen, R.A. (Eds)*, Wadsworth Statist./Probab. Ser., Monterey. Wadsworth. pages 807–810.

Illig A. (2001). *Probabilités de confiance après décision lors d'un test séquentiel avec un temps de bon fonctionnement de loi exponentielle.* Mémoire de DEA sous la direction du professeur B. Garel, Université Toulouse III.

Lecoutre B. (1997). *C'est bon à savoir ! et si vous êtiez un bayésien qui s'ignore.* Modulab, INRIA, vol. 8, n°18. pages 81–87.

Ledoux M. et Talagrand M. (1991). *Probability in Banach spaces.* Springer-Verlag.

Lemdani M. et Pons O. (1999). *Likelihood ratio tests in contamination models.* Bernoulli, vol. 5, n°4. pages 705–719.

Li L. A. et Sedransk N. (1988). *Mixtures of distributions : a topological approach.* Ann. Statist., vol. 16, n°4. pages 1623–1634.

Lindsay B. G. (1995). *Mixture models : Theory, geometry and applications.* IMS.

Liu X. et Shao Y. (2003). *Asymptotics for the likelihood ratio test in a two-component normal mixture model.* Annals of Statistics, vol. 31.

Martinez L., Wendy et Martinez R., Angel (2002). *Computational statistics handbook with matlab.* Chapman & Hall/CRC.

McLachlan G. et Peel D. (2000). *Finite mixture models.* Wiley-Interscience, New York.

McLachlan G. J. et Basford K. E. (1988). *Mixture models*, volume 84 of *Statistics : Textbooks and Monographs.* Marcel Dekker Inc.

McLachlan G. (1987). *On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture.* Appl. Statistics, vol. 36. pages 318–324.

Mengersen K. L., Robert C. P. et Guihenneuc-Jouyaux C. (1999). *MCMC convergence diagnostics : a reviewww.* In *Bayesian statistics, 6 (Alcoceber, 1998)*, pages 415–440. Oxford Univ. Press.

Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H. et Teller E. (1953). *Equations of state calculations by fast computing machines.* Journal of Chemical Physics, vol. 21 or 6. pages 1087–1091.

Mood A. et Graybill F. (1963). *Introduction to the theory of statistics.* Mc Graw-Hill, 2 éd.

Naylor J. et Smith A. (1983). *A contamination model in clinical chemistry : an illustration of a method for the efficient computation of posterior distributions.* The Statistician, vol. 32. pages 82–87.

Nougier J. P. (1987). *Méthodes de calcul numérique.* Mathématiques pour la Physique. [Mathematics for Physics]. Masson, 3 éd.

Pearson K. (1894). *Contributions to the theory of mathematical evolution.* Philosophical Transactions of the Royal Society of London A., vol. 185. pages 71–110.

Posse C. (1998), *Multivariate gaussian based clustering.* Rapport technique.

Raiffa H. et Schlaifer R. (1961). *Applied statistical decision theory.* Division of Research, Graduate School of Business Administration, Harvard University, Boston, Mass.

Redner R. (1981). *Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions.* Ann. Statist., vol. 9, n°1. pages 225–228.

Richardson S. et Green P. (1997). *On bayesian analysis of mixtures with an unknown number of components (with discussion).* Journal of the Royal Statistical Society, B, vol. 59. pages 731–792.

Ripley B. D. (1987). *Stochastic simulation.* John Wiley & Sons Inc.

Robert C. P. (1996a). *Méthodes de Monte Carlo par chaînes de Markov.* Éditions Économica.

Robert C. P. (1996b). *Mixtures of distributions : inference and estimation.* In *Markov chain Monte Carlo in practice*, pages 441–464. Chapman & Hall.

Robert C. P. et Casella G. (1999). *Monte Carlo statistical methods.* Springer-Verlag.

Roeder K. et Wasserman L. (1997). *Practical Bayesian density estimation using mixtures of normals.* J. Amer. Statist. Assoc., vol. 92, n°439. pages 894–902.

Siegmund D. (1985). *Sequential analysis : tests and confidence intervals.* Springer-Verlag.

Solka K., Wegman E., Priebe C., Poston W. et Rogers W. (1998). *Mixture stucture analysis using the akaike criterion and the bootstrap.* Statistics and Computing, vol. 8. pages 177–188.

Stephens M. (2000a). *Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods.* Ann. Statist., vol. 28, n°1. pages 40–74.

Stephens M. (2000b). *Dealing with label switching in mixture models.* J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 62, n°4. pages 795–809.

Stephens M. (1997). *Bayesian methods for mixture of normal distribution.* Magdalen College, Oxford, U.K.

Tassi P. (1985). *Méthodes statistiques.* Economica.

Teicher H. (1963). *Identifiability of finite mixtures.* Ann. Math. Statist., vol. 34. pages 1265–1269.

Tierney L. (1994). *Markov chains for exploring posterior distributions.* Ann. Statist., vol. 22, n°4. pages 1701–1762.

Tierney L. (1996). *Introduction to general state-space Markov chain theory.* In *Markov chain Monte Carlo in practice*, pages 59–74. Chapman & Hall.

Titterington D. M., Smith A. F. M. et Makov U. E. (1985). *Statistical analysis of finite mixture distributions.* John Wiley & Sons Ltd, Chichester.

Waagepetersen R. et Sorensen D. (April 2000), *A tutorial on reversible jump mcmc with a view toward applications in qtl-mapping.* Rapport technique. URL `http://www.math.auc.dk/~rw/revjump21.ps`.

Wald A. (1947). *Sequential Analysis.* John Wiley & Sons, New York.

Wilks S. (1938). *The large sample distribution of the likelihood ratio for testing composite hypotheses.* Ann. Math. Stat., vol. 9. page 60.

Wolfe J. (1971). *A Monte Carlo study of the sampling distribution ratio for mixtures of multinormal distributions.* Technical Bulletin STB, U.S. NAV. Pers. and Train. Res. San Diego, vol. 72, n°2.

Annexe A

Lois utilisées

- **Loi Binomiale :** $\mathcal{B}(n,p) \sim \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} \delta_k$, $p \in]0,1[$,
$$Esp = np \text{ et } Var = np(1-p).$$
- **Loi de Poisson :** $\mathcal{P}(\alpha) \sim \sum_{k=0}^{+\infty} e^{-\alpha} \frac{\alpha^k}{k!} \delta_k$, $\alpha > 0$,
$$Esp = \alpha \text{ et } Var = \alpha.$$
- **loi de Cauchy :** $C(\lambda) \sim \frac{1}{\pi} \frac{\lambda}{\lambda^2 + x^2}$
La loi de Cauchy n'admet ni espérance, ni variance, ni fonction génératrice des moments.
- **loi Gamma :** $\gamma(r,\lambda) \sim \frac{\lambda^r}{\Gamma(r)} e^{-\lambda x} x^{r-1} \mathbb{1}_{\mathbb{R}^+}(x)$, $r > 0$ et $\lambda > 0$
$$Esp = \frac{r}{\lambda} \text{ et } Var = \frac{r}{\lambda^2}.$$
- **Loi Normale sur** R^d **:** $\mathcal{N}_d(m,\Sigma) \sim \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-m)' \Sigma^{-1}(x-m)\right)$, $m \in \mathbb{R}^d$, Σ matrice réelle carrée d'ordre d, symétrique, définie positive.
- **Loi de Wishart :** C'est l'analogue de la loi gamma sur les matrices. Elle représente la distribution de la matrice de covariance d'un échantillon de taille m issu d'une loi Gaussienne à r dimensions de covariance A ($m \geqslant r$). La densité de V sur l'espace des matrices symétriques s'écrit :
$$\begin{aligned} W_r(V; m, A) &\sim K |A|^{-\frac{m}{2}} |V|^{\frac{m-r-1}{2}} \exp\left[-\frac{1}{2} tr\left(A^{-1}V\right)\right] \mathbb{1}_{[V \ def, \text{positive}]} \\ \text{avec} \quad K^{-1} &= 2^{\frac{mr}{2}} \pi^{r(r-1)/4} \prod_{s=1}^{r} \Gamma\left(\frac{m+1-s}{2}\right). \end{aligned}$$
On remarquera que
$$\Gamma(\alpha,\beta) = W_1\left(2\alpha, (2\beta)^{-1}\right).$$
- **Loi de Dirichlet :** On note $\mathcal{D}(\delta_1, ..., \delta_k)$ la distribution de Dirichlet sur le simplexe
$$\{(\pi_1, ..., \pi_{k-1}, 1 - \pi_1 - ... - \pi_{k-1}) : \pi_1 + ... + \pi_{k-1} \leqslant 1\},$$
de densité
$$\mathcal{D}(\delta_1, ..., \delta_k) \sim \frac{\Gamma(\delta_1 + ... + \delta_k)}{\Gamma(\delta_1) ... \Gamma(\delta_k)} \pi_1^{\delta_1 - 1} ... \pi_{k-1}^{\delta_{k-1} - 1} (1 - \pi_1 - ... - \pi_{k-1})^{\delta_k - 1}.$$
Si les δ_i sont égaux à 1, on a alors la distribution uniforme sur le simplexe.