REVUE DE STATISTIQUE APPLIQUÉE

A. BERDAÏ

B. GAREL

Performances d'un test d'homogénéité contre une hypothèse de mélange gaussien

Revue de statistique appliquée, tome 42, nº 1 (1994), p. 63-79

<http://www.numdam.org/item?id=RSA_1994__42_1_63_0>

© Société française de statistique, 1994, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 1994, **XLII** (1), 63-79

PERFORMANCES D'UN TEST D'HOMOGÉNÉITÉ CONTRE UNE HYPOTHÈSE DE MÉLANGE GAUSSIEN

A. Berdaï (1), B. Garel (2)

(1) Laboratoire de Statistique et Probabilités,
URA-CNRS DO745,
Université Paul Sabatier
(2) GAPSE – ENSEEIHT,
2 rue Camichel,
31071 Toulouse Cedex

RÉSUMÉ

Dans ce papier, nous étudions les performances d'un test d'homogénéité contre une hypothèse de mélange de deux lois normales unidimensionnelles, de moyennes θ_1 et θ_2 inconnues et d'écarts types connus et égaux $\sigma_1 = \sigma_2$. Nous reprenons dans ce cas particulier, la démarche de Ghosh et Sen(1985) qui ont proposé un test asymptotiquement localement minimax. Malheureusement, aucune tabulation n'a été obtenue par ces auteurs.

Dans le cas où θ_1 est supposée connue, une borne suggérée par Davies (1977) permet une tabulation approchée très satisfaisante. Dans le cas où θ_1 est inconnue, diverses simulations semblent indiquer qu'on puisse utiliser la même borne, en modifiant la statistique de test.

Mots clés : *Mélange gaussien; Test du rapport des maximums de vraisemblance; Loi asymptotique; Maximum de processus gaussien.*

SUMMARY

In this paper perfomances of a likelihood ratio test for testing homogeneity (*i.e.* no mixture) against a mixture of two distinct normal distributions with unknown mean θ_1 and θ_2 and known standard deviations $\sigma_1 = \sigma_2$ are evaluated. We follow Ghosh and Sen (1985) who proposed a locally asymptotically minimax test. Unfortunately no tabulation is given by these authors. When θ_1 is known a bound suggested by Davies(1977) allows to find approximate percentage point. With a slight modification of the test statistic we use the same value when θ_1 is unknown.

AMS 1980 subject classification numbers : *62E20, 62E25, 62F05.*

1. Introduction et notations

Soit $X_1, ..., X_n$ un échantillon d'une variable aléatoire (v.a.) X à valeurs dans $\mathbb{R}$, dont la loi admet la densité :

$$
\begin{aligned}
f(x, \psi') &= \sum_{i=1}^{k} \Pi_i f_i(x, \theta_i) \quad \text{avec} \\
\forall i = 1, ..., k, 0 < \Pi_i < 1 \quad &\text{et} \quad \sum_{i=1}^{k} \Pi_i = 1.
\end{aligned}
\tag{1.1}
$$

$f_i(., \theta_i)$ est une densité de probabilité appartenant à une famille paramétrique dépendant du paramètre θ_i; Π_i est la probabilité qu'un point de l'échantillon suive la loi de densité $f_i(., \theta_i)$.

On appellera les $\Pi_i, i = 1, ..., k$ les proportions du mélange fini à k composants. $\Pi = (\Pi_1, ..., \Pi_k)', \theta = (\theta_1, ..., \theta_k)'$ et on note ψ le vecteur de tous les paramètres distincts intervenant dans le mélange : $\psi = (\Pi', \theta')'$.

1.1 Test du rapport des maximums de vraisemblance.

On souhaite mettre en œuvre un test pour identifier le nombre de composants dans un mélange. Pour cela, on utilise le test du rapport des maximums de vraisemblances (TRMV) :

$$
\lambda_n = \frac{\underset{\psi \epsilon H_0}{\text{Sup}} \left(\prod_{i=1}^{n} f(x_i, \psi') \right)}{\underset{\psi \epsilon H}{\text{Sup}} \left(\prod_{i=1}^{n} f(x_i, \psi') \right)}
\tag{1.2}
$$

où l'hypothèse H_0 correspond à l'homogénéité et H_1 à un véritable mélange, avec $H = H_0 \cup H_1$. L'utilisation de λ_n est asymptotique. Si la théorie usuelle est valable, alors, la loi de $-2\text{Log}\lambda_n$ sous H_0 est asymptotiquement un chi-deux, de degré de liberté égal au nombre de contraintes à imposer sous H_1 pour obtenir H_0 *cf.* Titterington, Smith et Makov(1985). Malheureusement, les conditions habituelles de régularité ne sont pas vérifiées pour $-2\text{Log}\lambda_n$. En effet, outre un problème de calcul du bon degré de liberté du chi-deux, la valeur du paramètre p correspondant à l'homogénéité $p = 0$, est sur la frontière de l'espace des paramètres et non à l'intérieur de celui-ci, comme on le suppose dans la théorie classique.

Self et Liang(1987) ont étudié les propriétés asymptotiques de l'estimateur du maximum de vraisemblance (EMV) et du rapport des vraisemblances quand la vraie valeur du paramètre est sur la frontière de l'espace des paramètres.

1.2 Identifiabilité d'un mélange.

Notons également que s'ajoute ici, un problème d'identifiabilité du mélange; en effet : si on désigne par $U_{a,b}$ la loi uniforme sur $[a, b] \subset \mathbb{R}$, on remarque que : $U_{0,1}(x) = (1-p)U_{p,1}(x) + pU_{0,p}(x)$ pour tout $p\epsilon]0, 1[$, et donc l'égalité

$$(1-p_1)U_{p_1,1}(x) + p_1 U_{0,p_1}(x) = (1-p_2)U_{p_{2},1}(x) + p_2 U_{0,p_2}(x)$$

n'implique pas que $p_1 = p_2$.

Pour résoudre le problème d'identifiabilité ci-dessus, on pose la définition suivante : soit

$$\mathcal{H} = \{(1-p)f(x, \theta_1) + pf(x, \theta_2), p\epsilon[0, 1], \theta_i \epsilon \Theta_i, i = 1, 2\}\,; \tag{1.3}$$

on dit que $\mathcal{H}$ est une famille à composants identifiables, ou en bref $\mathcal{H}$ est identifiable, si et seulement si pour $p \neq 0$, $p \neq 1$, $\theta_1 \neq \theta_2$, l'égalité :

$$(1-p)f(x, \theta_1) + pf(x, \theta_2) = (1-p^*)f(x, \theta_1^*) + p^* f(x, \theta_2^*) \tag{1.4}$$

implique $(p = p^*, \theta_1 = \theta_1^*$ et $\theta_2 = \theta_2^*)$ ou $(p = 1 - p^*, \theta_1 = \theta_2^*$ et $\theta_2 = \theta_1^*)$.

Une condition nécessaire et suffisante pour que $\mathcal{H}$ soit identifiable dans le cas général d'un mélange à k composants est donnée par Yakowitz et Spragins (1968). Les mélanges gaussiens, exponentiels, de Cauchy sont identifiables, par contre, les mélanges binomiaux et uniformes ne le sont pas en général.

Remarquons que même si les densités sont identifiables au sens de la définition ci-dessus, les paramètres ne le sont pas (on renvoie à Kotz et Johnson, vol.4, pour les notions d'identifiabilité).

Sous H_1 par exemple, $f(x, (p, \theta_1, \theta_2)) = f(x, (1-p, \theta_2, \theta_1))$ pour tous p, θ_1, θ_2. Une façon de contourner cette difficulté est d'étudier les classes de paramètres constituées des valeurs pour lesquelles les densités sont égales. On identifie le paramètre à sa classe et on étudie le maximum de vraisemblance sur l'espace quotient.

Parmi les espaces de paramètres équivalant à l'espace quotient, on choisit celui sur lequel on maximise la vraisemblance. Par exemple, sous H_1, on prendra l'espace des paramètres $\psi = (p, \theta_1, \theta_2)$ tels que :

$$p\epsilon[0, \frac{1}{2}] \text{ et } \| \theta_2 - \theta_1 \| \geqslant \varepsilon > 0 \tag{1.5}$$

condition de séparation des paramètres sous H_1 proposée par Ghosh et Sen (1985).

Remarquons que si Θ_1 et Θ_2 sont deux intervalles bornés de $\mathbb{R}$, tels que $\bar{\Theta}_1 \cap \Theta_2 = \emptyset$, alors, l'ordre défini sur $\mathbb{R}$ nous permet de prendre comme espace équivalent l'ensemble des paramètres $\psi = (p, \theta_1, \theta_2)$ tels que :

$$p\epsilon[0, 1] \quad \text{et} \quad | \theta_2 - \theta_1 | \geqslant \varepsilon > 0 \tag{1.6}$$

Un résultat concernant la convergence de L'EMV de θ étudié sur l'espace quotient nous assurant l'identifiabilité des paramètres a été établi par Redner (1981).

1.3 Les approches successives.

Une des premières études dans ce contexte, a été proposée par Wolfe (1971), où il considère le TRMV pour tester l'hypothèse nulle H_0 : «mélange à k' composants» contre l'hypothèse alternative H_1 :«mélange à k composants» $(k > k')$. Il suggère, à partir des simulations effectuées dans le cas où $k' = 1$ et $k = 2$, d'approcher la loi de $-2c\text{Log}\lambda_n$ par un χ^2_ν où ν est deux fois la différence du nombre de paramètres dans les deux hypothèses, sans compter les proportions du mélange, avec $c = (n - 2 - \frac{1}{2}k)/n$.

Titterington, Smith et Makov (1985) considèrent le TRMV pour un mélange de deux densités quelconques connues, avec des proportions inconnues. Ils montrent qu'asymptotiquement sous l'hypothèse d'homogénéité, $-2\text{Log}\lambda_n = 0$ avec une probabilité $\frac{1}{2}$ et avec la même probabilité, est distribuée comme un chi-deux avec 1 degré de liberté. De façon équivalente, la loi de $-2\log\lambda_n$ est celle de $\{\text{Sup}(0, Y)\}^2$ où $Y \in N(0, 1)$ où $N(m, \sigma^2)$ représente la loi normale de moyenne m et de variance σ^2 : voir McLachlan et Basford (1988, p. 22).

Hartigan (1985; a,b) obtenait le même résultat dans un cas particulier.

Une analyse des problèmes posés, de leur lien avec le problème d'identifiabilité et de leur complexité topologique a été réalisée par Ghosh et Sen(1985). Ils ont étudié le comportement asymptotique de $-2\text{Log}\lambda_n$ pour un mélange à deux composants éléments d'une famille paramétrique quelconque. Ils ont obtenu la loi limite de $-2\text{Log}\lambda_n$ pour θ_1 et θ_2 inconnus, mais identifiables où θ_2 est dans un ensemble compact. Ils montrent que $-2\text{Log}\lambda_n$ est distribuée comme une certaine fonction $[\text{Max}(0, \sup_{\theta_2} T(\theta_2))]^2$ où T(.) est un processus gaussien dont l'espace des indices est l'espace des valeurs du paramètre θ_2, de moyenne nulle et dont la covariance dépend de la vraie valeur de θ_1 sous H_0 et

$$\text{Var}(T(\theta_2)) = 1 \quad \text{sous} \quad H_0, \forall\theta_2.$$

Pour le cas où le vecteur des paramètres ψ est supposé non identifiable, ils imposent une condition de séparation sur la valeur de ψ sous H_0 et H_1. Ils ont montré que le test obtenu est asymptotiquement localement minimax pour un mélange à deux composants, éléments d'une famille exponentielle.

Une bibliographie sur les problèmes de mélanges a été donnée par Titterington(1990). Il présente les méthodes les plus importantes et les plus récentes sur le sujet.

2. Test d'homogénéité de deux lois normales unidimensionnelles

Le problème consiste à tester l'hypothèse nulle

$$H_0 :'' f(x, \psi) = g(x - \theta_1)'' \quad \text{ou encore} \quad ''p = 0'' \tag{2.1}$$

postulant une homogénéité, contre l'hypothèse alternative

$$H_1 :'' f(x, \psi) = (1 - p)g(x - \theta_1) + pg(x - \theta_2), p\epsilon]0, 1[, \theta_1 \neq \theta_2 '' \tag{2.2}$$

postulant un véritable mélange, dans le cas particulier où g est la densité d'une loi normale centrée réduite. On note par $f(x,\psi^0)$ la vraie densité. Ainsi, sous H_0, la densité homogène est $f(x,\psi^0) = g(x-\theta_1^0)$ avec $\psi^0 = (0,\theta_1^0,\theta_2), \theta_2$ quelconque.

2.1 Application de la théorie de Ghosh et Sen (1985).

Soit $E = (X_1, ..., X_n)$ un échantillon de taille n d'une variable aléatoire réelle (v.a.r) X de densité : $f(X,\psi) = (1-\theta_0)g(X-\theta_1) + \theta_0 g(X-\theta_2), \theta_0 \epsilon [0,1], \theta_1 \epsilon \Theta_1$ ouvert relativement compact de $\mathbb{R}$, $\theta_2 \epsilon [b,c], (b,c) \epsilon \mathbb{R}^2$ et g est la densité d'une loi normale centrée réduite. Soit a un réel strictement inférieur à b. On suppose que :

$$\theta_1 < a \qquad \forall \theta_1 \epsilon \Theta_1 \tag{2.3}$$

condition de séparation des paramètres sous H_1.

Soit λ_n la statistique de test du rapport des maximums de vraisemblance. Alors on a

$$-2\text{Log}\lambda_n = 2\left[\text{Sup}_{H_1} L_n(\psi) - \text{Sup}_{H_0} L_n(\psi)\right] \tag{2.4}$$

où L_n est la fonction Log-vraisemblance de E.

Soit $\psi^0 = (0,\theta_1^0,\theta_2)$ la vraie valeur du paramètre sous H_0, et

$$\eta = (\eta_0,\eta_1,\eta_2) \quad \text{avec} \quad \theta_0 = \frac{\eta_0}{\sqrt{n}}, \theta_1 = \theta_1^0 + \frac{\eta_1}{\sqrt{n}}$$

$$\text{et} \quad \theta_2 = \eta_2.$$

La justification de cette paramétrisation se trouve dans Ghosh et Sen (1985, section 3, p. 799). Le problème est de contrôler le reste R_n du développement de Taylor à l'ordre deux de $L_n(\psi)$ autour de ψ^0, ce qui rejoint la démarche de Wald(1943).

Pour cela, Ghosh et Sen (1985) ont procédé en trois étapes :

la première étape consiste à atteindre le maximum de $L_n(\psi)$ sur un voisinage compact inclu dans Θ_1 :

$(i.e)$: il existe un voisinage compact C de θ_1^0 tel que :

$$\sup_{(\theta_0,\theta_1)\epsilon[0,1]\times\Theta_1} L_n(\psi) - \sup_{(\theta_0,\theta_1)\epsilon[0,1]\times C} L_n(\psi) = o_{p,n}(1), \forall \eta_2. \tag{2.5}$$

avec $o_{p,n}(1)$ est une quantité qui converge en probabilité vers 0, uniformément par rapport à η_2, quand n tend vers l'infini.

La deuxième étape consiste à atteindre le maximum de $L_n(\psi)$ sur un voisinage quelconque, mais fixé de (O,θ_1^0).

La troisème étape consiste à atteindre le maximum de $L_n(\psi)$ sur un $n^{-\frac{1}{2}}$ voisinage borné dans le plan (η_0,η_1).

Ainsi, on montre que R_n est un $o_{p,n}(1)$ uniformément pour η_2 sur ce $n^{-\frac{1}{2}}$ voisinage borné. Par application de la démarche de Ghosh et Sen (1985) au cas du mélange : $(1-\theta_0)N(\theta_1,1)+\theta_0 N(\theta_2,1)$ sous la condition (2.3), on a :

$$-2\mathrm{Log}\lambda_n = \left(\sup_{\eta_2} T_n(\eta_2)\right)^2 .\mathbb{1}_{\left(\sup_{\eta_2} T_n(\eta_2)\geqslant 0\right)} + o_{p,n}(1), \tag{2.6}$$

$$T_n(\eta_2) = \frac{n^{-\frac{1}{2}}\sum_{i=1}^{n}\left[\left(\frac{e^{-\frac{1}{2}(X_i-\eta_2)^2}}{e^{-\frac{1}{2}(X_i-\theta_1^0)^2}}-1\right)+(\theta_1^0-\eta_2)(X_i-\theta_1^0)\right]}{(e^{(\eta_2-\theta_1^0)^2}-(\eta_2-\theta_1^0)^2-1)^{\frac{1}{2}}} \tag{2.7}$$

et $T_n(\eta_2)\xrightarrow[n\to+\infty]{\mathcal{L}_{H_0}} N(0,1)$, où $\mathcal{L}_{H_0}$ représente la convergence en loi sous H_0.

Ghosh et Sen (1985) ont montré que le test obtenu est asymptotiquement localement minimax. Pour fixer les idées, disons qu'un test minimax à un niveau α donné est un test dont la plus mauvaise performance est meilleure que celle des autres tests de niveau α.

Pour une définition complète dans un contexte asymptotique local, on renvoie à Ghosh et Sen (1985, section 4, p. 803).

Malheureusement, cette théorie reste incomplète à cause de la difficulté de déterminer la loi du $\sup_{\eta_2} T_n(\eta_2)$. Cependant, comme nous allons le voir, en considérant un cas particulier, il est possible de proposer une statistique de test et une tabulation.

2.2 Cas du mélange $(1-\theta_0)N(0,1)+\theta_0 N(\theta_2,1); \theta_0,\theta_2$ *inconnus.*

Remarquons que dans le cas du mélange $(1-\theta_0)N(0,1)+\theta_0 N(\theta_2,1)$, θ_0,θ_2 inconnues, la vraie valeur de θ_1 sous H_0 est $\theta_1^0=0$.

La statistique du TRMV ne peut être obtenue en remplaçant θ_1^0 par 0 dans (2.7) pour un mélange $(1-\theta_0)N(\theta_1,1)+\theta_0 N(\theta_2,1)$ $\theta_0,\theta_1,\theta_2$inconnues. Cela résulte du nombre de paramètres inconnus présents dans les deux cas de mélange. Ainsi, on est amené à faire un calcul similaire dans ce cas particulier.

Soit un échantillon $X_1,...,X_n$ d'une v.a.r. X de densité $f(X,\psi)=(1-\theta_0)g(X)+\theta_0 g(X-\theta_2)$ avec $\psi=(\theta_0,\theta_2)$, $\theta_2\epsilon[b,c]$, $\theta_0\epsilon[0,1]$ et g est la densité d'une loi normale centrée réduite. Posons :

$$\begin{cases} \theta_0=\frac{\eta_0}{\sqrt{n}}, \theta_2=\eta_2. \\ \eta=(\eta_0,\eta_2) \end{cases} \tag{2.8}$$

On suppose que $|\ \theta_2\ |\geqslant \varepsilon>0$: condition d'identification des paramètres sous H_1.La fonction de Log-vraisemblance est :

$L_n(\psi) \;=\; \sum_{i=1}^{n} \mathrm{Log}\left[(1-\theta_0)g(X_i)+\theta_0 g(X_i-\theta_2)\right]$ et on note $L_n(\psi)$ par $V_n(\eta)$. On s'intéresse à l'étude de

$$-2\mathrm{Log}\lambda_n = \mathrm{Sup}[2(L_n(\eta_2) - L_n(H_0))] = \sup_{\eta_2}[-2\mathrm{Log}\lambda_n(\eta_2)]$$

avec $L_n(\eta_2) = \underset{\theta_0\in[0,1]}{\mathrm{Sup}}\; L_n(\psi), L_n(H_0) = \sum_{i=1}^{n}\mathrm{Log}(g(X_i))$ et

$$-2\mathrm{Log}\lambda_n(\eta_2) = 2[L_n(\eta_2) - L_n(H_0)]. \tag{2.9}$$

Par un développement de Taylor à l'ordre deux de $V_n(\eta)$ autour de $(0,\eta_2)'$, pour n suffisamment grand, on obtient :

$$V_n(\eta) = V_n((0,\eta_2)) + \eta_0\left[n^{-\frac{1}{2}}\sum_{i=1}^{n} a_i(\eta_2)\right] + \frac{1}{2}\eta_0^2\left[-\frac{1}{n}\sum_{i=1}^{n} a_i^2(\eta_2)\right] + o_{p,n}(1)$$

avec $a_i(\eta_2) = \dfrac{e^{-\frac{1}{2}(X_i-\eta_2)^2}}{e^{-\frac{1}{2}X_i^2}} - 1$ et $\frac{1}{n}\sum_{i=1}^{n} a_i^2(\eta_2) \xrightarrow[n\to+\infty]{} \mathrm{var}_{H_0}(a_i(\eta_2)) = e^{\eta_2^2}-1.$
et donc :

$$V_n(\eta) = V_n((0,\eta_2)) + \eta_0(n^{-\frac{1}{2}}\sum_{i=1}^{n} a_i(\eta_2)) - \frac{1}{2}\eta_0^2(e^{\eta_2^2}-1) + o_{p,n}(1) \tag{2.10}$$

Ainsi, pour η_2 fixé :

$$L_n(\eta_2) = \underset{\theta_0\in[0,1]}{\mathrm{Sup}}\; L_n(\psi) = V_n((0,\eta_2)) + \underset{\eta_0\geqslant 0}{\mathrm{Sup}}\left[\eta_0(n^{-\frac{1}{2}}\sum_{i=1}^{n} a_i(\eta_2)) - \frac{1}{2}\eta_0^2(e^{\eta_2^2}-1)\right] + o_{p,n}(1).$$

Notons par η_0^s la valeur pour laquelle la fonction h définie par :
$h(x) = (n^{-\frac{1}{2}}\sum_{i=1}^{n} a_i(\eta_2))x - \frac{1}{2}(e^{\eta_2^2}-1)x^2$ est maximale. On obtient

$$\eta_0^s = n^{-\frac{1}{2}}\sum_{i=1}^{n} a_i(\eta_2)/(e^{\eta_2^2}-1) \text{ et}$$

$$\begin{cases} \text{si } \eta_0^s < 0, & \sup_{\eta_0 \geqslant 0} h(\eta_0) = 0. \\ \text{si } \eta_0^s \geqslant 0, & sup_{\eta_0 \geqslant 0} h(\eta_0) = \eta_0^s n^{-\frac{1}{2}} \sum_{i=1}^{n} a_i(\eta_2) - \frac{1}{2}(\eta_0^s)^2(e^{\eta_2^2} - 1). \end{cases} \tag{2.11}$$

D'où, en posant :

$$T_n(\eta_2) = n^{-\frac{1}{2}} \sum_{i=1}^{n} a_i(\eta_2)/(e^{\eta_2^2} - 1)^{\frac{1}{2}}. \tag{2.12}$$

on voit que le signe de $T_n(\eta_2)$ est celui de η_0^s et on a :
- Si $T_n(\eta_2) < 0, L_n(\eta_2) = L_n(H_0) + o_{p,n}(1)$ et donc $-2\text{Log}\lambda_n(\eta_2) = o_{p,n}(1)$.

$$\text{- Si } T_n(\eta_2) \geqslant 0, L_n(\eta_2) = L_n(H_0) + \frac{1}{2}\left(\frac{n^{-\frac{1}{2}} \sum_{i=1}^{n} a_i(\eta_2)}{(e^{\eta_2^2} - 1)^{\frac{1}{2}}}\right)^2 + o_{p,n}(1)$$

et donc $-2Log\lambda_n(\eta_2) = T_n^2(\eta_2) + o_{p,n}(1)$.
D'où :

$$-2\text{Log}\lambda_n(\eta_2) = T_n^2(\eta_2)\mathbb{1}_{(T_n(\eta_2)\geqslant 0)} + o_{p,n}(1) \tag{2.13}$$

avec

$$T_n(\eta_2) \xrightarrow{\mathcal{L}_{H_0}} N(0,1), \forall \eta_2.$$

et donc

$$-2\text{Log}\lambda_n = (\sup_{\eta_2} T_n(\eta_2))^2.\mathbb{1}_{(\sup_{\eta_2} T_n(\eta_2)\geqslant 0)} + o_{p,n}(1). \tag{2.14}$$

2.3 Modèle de R.B. Davies (1977)

Soit un échantillon $X_1, ..., X_n$ d'une v.a. X, dont la loi admet une densité $f(X, \xi, \mu)$ dépendant de deux paramètres ξ et μ et telle que $f(X, 0, \mu)$ ne dépend pas de μ, $\mu \epsilon [b, c]$.
On veut tester H_0 :$'' \ \xi = 0''$ contre H_1 :$'' \ \xi > 0''$. Posons

$$Z_n(\mu) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left[\frac{\partial}{\partial \xi} \left\{ \text{Log} f(X_i, \xi, \mu) \right\} / \gamma(\mu) \right]_{\xi=0} \tag{2.15}$$

où $(\gamma(\mu))^2 = \text{var}\{[\frac{\partial}{\partial \xi}\text{Log}f(X_i, \xi, \lambda)]_{\xi=0}\}$ calculé sous H_0.

Supposons que sous $H_0, \text{Cov}(Z_n(\mu_1), Z_n(\mu_2)) = \rho(\mu_1, \mu_2)$.

De la définition de Z_n, on a : $\rho(\mu, \mu) = 1, \forall \mu$.

Soit $D(Z_n; \xi, \mu)$ la distribution du processus $Z_n(.)$ quand ξ et μ sont des paramètres donnés. Alors, sous des conditions de régularité, on a :

$$D(Z_n; n^{-\frac{1}{2}}\xi, \mu) \xrightarrow[n \to +\infty]{} D(Z; \xi, \mu)$$

où Z(.) est un processus gaussien, de moyenne

$$E_{\xi_0, \mu_0}[Z(\mu_1)] = \xi_0 \gamma(\mu_0) \rho(\mu_1, \mu_0)$$

et de covariance

$$\text{Cov}(Z(\mu_1), Z(\mu_2)) = \rho(\mu_1, \mu_2).$$

On peut obtenir une formule asymptotiquement équivalente à (2.15), en remplaçant $Z_n(\mu)$ par :

$$S_n(\mu) = \sqrt{n}\gamma(\mu)\hat{\xi}_n(\mu) \tag{2.16}$$

ou

$$W_n(\mu) = \{-2\text{Log}(\lambda_n(\mu))\}^{\frac{1}{2}}\text{sgn}(\hat{\xi}_n(\mu)) \tag{2.17}$$

où $\widehat{\xi_n}(\mu)$ est l'estimateur du maximum de vraisemblance de ξ quand μ est connue : *cf.* Davies (1977 p. 248). Le test proposé dans ce modèle a une région critique de la forme $\{\sup\limits_{b \leqslant \mu \leqslant c} Z(\mu) > k\}$.

Malheureusement, on n'est pas capable de calculer $P_{\xi_0, \mu_0}\{\sup\limits_{b \leqslant \mu \leqslant c} Z(\mu) > k\}$ même pour $\xi_0 = 0$. Cependant, une borne de cette quantité a été proposée par R. B. Davies (1977) se basant sur l'approche théorique de Cramer et Leadbetter (1967) :

$$P_{H_0}\{\sup_{b \leqslant \mu \leqslant c} Z(\mu) > k\} \leqslant \Phi(-k) + \frac{1}{2\Pi}\exp(-\frac{1}{2}k^2)\int_b^c (-\rho_{11}(\mu))^{\frac{1}{2}} d\mu \tag{2.18}$$

où $\rho_{11}(\mu) = [\frac{\partial^2}{\partial u^2}\rho(u, \mu)]_{u=\mu}$ et Φ est la fonction de répartition d'une loi normale centrée réduite.

Ce qui nous donne une borne du niveau de signification du test.

Généralement, on suppose que μ prend des valeurs discrètes $\mu_1, ..., \mu_m$ et que les propriétés du modèle sont satisfaites.

D'une façon analogue au cas continu, on a :

$$P_{H_0}\left\{\sup_{1 \leqslant i \leqslant m} Z(\mu_i) > k\right\} \leqslant \Phi(-k) + \sum_{i=2}^{m} P_{H_0}\{Z(\mu_{i-1}) < k, Z(\mu_i) > k\}$$

autrement dit :

$$P_{H_0}\left\{\operatorname*{Sup}_{1\leqslant i\leqslant m} Z(\mu_i) > k\right\} \leqslant \Phi(-k) + \sum_{i=2}^{m} P_{H_0}\{Z(\mu_{i-1}) < k\} - \sum_{i=2}^{m} P_{H_0}\{Z(\mu_{i-1}) < k, Z(\mu_i) < k\} \tag{2.19}$$

Ainsi, dans le cas du mélange $(1-\theta_0)N(0,1)+\theta_0 N(\theta_2,1), \theta_0, \theta_2$ inconnues, on est bien dans les conditions du test de R. B. Davies et par suite, par une discrétisation de l'intervalle $[b,c]$ en $\{\theta_2^1, ..., \theta_2^m\}$, on a :

$$P_{H_0}\left\{\mathrm{Sup}_{1\leqslant i\leqslant m} T_n(\theta_2^i) > k\right\} \leqslant \Phi(-k) + \sum_{i=2}^{m} P_{H_0}\left\{T_n(\theta_2^{i-1}) < k\right\} - \sum_{i=2}^{m} P_{H_0}\left\{T_n(\theta_2^{i-1}) < k, T_n(\theta_2^i) < k\right\}$$

avec T_n défini au §2 : (2.12).
la probabilité bivariée $P_{H_0}\left\{T_n(\theta_2^{i-1}) < k, T_n(\theta_2^i) < k\right\}$ est complètement évaluée par la donnée de :

$\rho(\theta_2^{i-1}, \theta_2^i) = \mathrm{Cov}(T_n(\theta_2^{i-1}), T_n(\theta_2^i)) = (e^{(\theta_2^{i-1})^2}-1)^{-\frac{1}{2}}(e^{(\theta_2^i)^2}-1)^{-\frac{1}{2}}(e^{\theta_2^{i-1}\theta_2^i}-1)$
pour $i = 1, ..., m$.

3. Résultats et simulations

Dans une première étude, nous avons vérifié par des simulations la qualité de la borne de R. B. Davies(1977), en discrétisant l'intervalle [b,c]. L'intervalle [b,c] a été fixé égal à [1,2], et on a pris $m = 11, \theta_1 = 1.0, \theta_2 = 1.1, \theta_3 = 1.2, \theta_4 = 1.3, \theta_5 = 1.4, \theta_6 = 1.5, \theta_7 = 1.6, \theta_8 = 1.7, \theta_9 = 1.8, \theta_{10} = 1.9$ et $\theta_{11} = 2.0$.
On peut trouver dans Davies (1977) un autre choix de discrétisation de [b,c].

On simule le vecteur gaussien multidimensionnel $\begin{pmatrix} T(\theta_1) \\ \cdot \\ \cdot \\ \cdot \\ T(\theta_m) \end{pmatrix}$ de moyenne nulle et de matrice variance-covariance :

$$\Sigma = \left((e^{\theta_i^2}-1)^{-\frac{1}{2}}(e^{\theta_j^2}-1)^{-\frac{1}{2}}(e^{\theta_i\theta_j}-1)\right)_{\substack{1\leqslant i\leqslant m\\ 1\leqslant j\leqslant m}} \tag{3.1}$$

On réitère cette simulation 10 000 fois, ainsi on obtient un échantillon de 10 000 individus suivant le processus discrétisé. Pour chaque individu, on retient la composante maximale de ce vecteur, ce qui nous permet d'obtenir une approximation de la loi du $\sup_{1 \leqslant i \leqslant m} T(\theta_i)$ et donc de déterminer la valeur du k correspondant à

$P_{H_0}\left\{\sup_{\theta \epsilon [b,c]} T(\theta) > k\right\} = \alpha$, pour α fixé.

La deuxième étape consiste à calculer la borne de Cramer et Leadbetter transformée par R. B. Davies (1977) :

$$\Phi(-k) + \sum_{i=2}^{m} P_{H_0}\{T(\theta_{i-1}) < k\} - \sum_{i=2}^{m} P_{H_0}\{T(\theta_{i-1}) < k, T(\theta_i) < k\}$$
$$= 1 + (m-2)\Phi(k) - \sum_{i=2}^{m} P_{H_0}\{T(\theta_{i-1}) < k, T(\theta_i) < k\} \qquad (3.2)$$

Le tableau 1 donne les valeurs des seuils de rejet de H_0 pour des niveaux de signification de première espèce fixé à 5%, 10%, 20% et 25%, ainsi que la valeur de la borne de R.B.Davies correspondante :

TABLEAU 1

$P_{H_0}\left\{\operatorname{Sup}_{1 \leqslant i \leqslant m} T(\theta_i) > k\right\}$	k	Valeur de la borne de Davies
5%	1,9234	4,97%
10%	1,5797	9,81%
20%	1,1334	20,38 %
25%	0,9685	25,60 %

Par une étude complète des différentes valeurs des seuils de rejet de H_0, pour des valeurs de niveaux de première espèce fixés de 1% jusqu'à 100%, on remarque que la borne de R. B. Davies qui constitue un majorant théorique de $P_{H_0}\{\operatorname{Sup}_\theta \epsilon [b,c] T(\theta) > k\}$ est très proche de cette quantité, ce qui nous permet d'approcher les quantiles du sup du processus et donc d'avoir une bonne approximation de la fonction de répartition de $\operatorname{Sup} \theta \epsilon [b,c] T(\theta)$: *cf.* BerdaïA. (1991).

Bien naturellement, ces résultats nous ont incités à utiliser la borne de Davies pour obtenir les valeurs critiques du test proposé ci-dessus : cas de mélange $(1-p)N(0,1) + pN(\theta,1)$ avec p, θ inconnues et dont la statistique est :

$$\left(\operatorname*{Sup}_{\theta\epsilon[b,c]} T_{n(\theta)}\right)^2 \mathbb{1}_{\left(\operatorname*{Sup}_{\theta\epsilon[b,c]} T_n(\theta \geqslant 0\right)} \quad o\grave{u} T_n(\theta) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left(\frac{e^{-\frac{1}{2}(X_i-\theta)^2}}{e^{-\frac{1}{2}X_i^2}} - 1 \right) \Big/ (e^{\theta^2}-1)^{\frac{1}{2}}$$

Une première série de simulations a été effectuée pour θ_1 connue en utilisant l'échantillon recentré par rapport à θ_1.

Le tableau 2 (resp. le tableau 3) donne l'erreur de première espèce au niveau nominal 5% sous H_0 (resp. de deuxième espèce au niveau nominal 5%), pour 100 simulations :

TABLEAU 2
Echantillons provenant de N(1,1).
Erreurs de première espèce au niveau 5% pour 100 simulations sous H_0.

Taille de l'échantillon	50	100	150	200	250	300	400	500
Erreur de première espèce	4%	6%	7%	7%	6%	6%	5%	5%

TABLEAU 3
Echantillons provenant de $0.8N(1,1) + 0.2N(2.1)$
Erreurs de deuxième espèce au niveau 5% pour 100 simulations

Taille de L'échantillon	50	100	150	200	250	300	400	500
Erreurs de deuxième espèce	44%	15%	10%	4%	2%	3%	1%	0%

4. Extension

Nous souhaitons maintenant traiter le cas où θ_1 est inconnue. Notons que dans ce cas, les développements de R. B. Davies ne sont plus directement applicables.

On peut envisager le mélange $(1-p)N(\theta_1,1) + pN(\theta_2,1)$ comme une contamination d'ordre p de $N(\theta_1,1)$ par $N(\theta_2,1)$.

Il nous a paru naturel d'utiliser un estimateur robuste de θ_1 : $\hat{\theta}_1$ et par suite de travailler comme si θ_1 était connue.

En centrant par rapport à $\hat{\theta}_1$, on a utilisé le test du mélange $(1-p)N(0,1) + pN(\theta,1), p, \theta$ inconnus. Pour cela, on a calculé divers estimateurs robustes de θ_1; mais, nous nous sommes surtout intéressés à ceux de Huber et de Andrews : voir D. F. Andrews *et al.*(1972). Ainsi, on a mis en application le test statistique du mélange $(1-p)N(\theta_1,1) + pN(\theta_2,1)$ où p, θ_1, θ_2 sont supposés inconnus. Les étapes du test sont alors les suivantes :

$(E1)$: Estimer θ_1 par les estimateurs robustes de Huber et de Andrews ainsi que par la moyenne : $\widehat{\theta}_1$.

$(E2)$: Centrer l'échantillon par rapport à $\widehat{\theta}_1$.

$(E3)$: Effectuer le test statistique à l'aide de la borne de Davies sur l'échantillon centré.

4.1 Résultats de simulations.

Nous avons réalisé un premier groupe de simulations sous l'hypothèse d'homogénéité H_0 afin d'établir des tableaux d'erreurs de première espèce. Puis un second groupe de simulations, cette fois sous l'hypothèse alternative, nous a permis d'évaluer la puissance du test.

4.1.1 Simulations sous l'hypothèse H_0 *:*$'' \ f \equiv N(\theta_1, 1)''$.

Le tableau 4 (resp.5) donne l'erreur de première espèce au niveau nominal 5% pour 100 simulations sous H_0, pour le test d'un mélange $(1-p)N(\theta_1, 1) + pN(\theta_2, 1), p, \theta_1, \theta_2$ inconnus, appliqué à des échantillons de tailles variant entre 50 et 400, provenant d'une loi normale N(0,1) (resp. N(10,1)).

TABLEAU 4

Echantillons provenant de N(0,1)

Erreurs de première espèce au niveau 5% pour 100 simulations sous H_0.

Taille de l'échantillon n Type d'estimateurs	50	100	150	200	250	300	400
Andrews	2%	2%	5%	3%	7%	6%	5%
Huber	6%	1%	5%	5%	3%	5%	1%
Moyenne	1%	5%	2%	3%	4%	1%	3%

TABLEAU 5

Echantillons provenant de N(10,1)

Erreurs de première espèce au niveau 5% pour 100 simulations.

Taille de l'échantillon n Type d'estimateurs	50	100	150	200	250	300	400
Andrews	3%	3%	4%	7%	6%	7%	4%
Huber	4%	5%	3%	2%	3%	4%	5%
Moyenne	4%	1%	1%	5%	4%	3%	4%

4.1.2 Simulations sous l'hypothèse H_1.

Les tableaux suivants 6,7,8 et 9 présentent l'erreur de deuxième espèce au niveau nominal 5% pour 100 simulations, pour le test de mélange $(1-p)N(\theta_1, 1) + pN(\theta_2, 1), p, \theta_1, \theta_2$ inconnues : appliqué à des échantillons provenant de 4 mélanges, pour les cas suivants supposés intéressants : mélanges déséquilibrés et $| \theta_2 - \theta_1 |$ suffisamment petit.

Tableau 6 : $p = 0.8$, $\theta_1 = 0.2$ et $\theta_2 = 1.8$

Tableau 7 : $p = 0.7$, $\theta_1 = 0.2$ et $\theta_2 = 1.8$

Tableau 8 : $p = 0.8$, $\theta_1 = 0.5$ et $\theta_2 = 1.8$

Tableau 9 : $p = 0.8$, $\theta_1 = 0.5$ et $\theta_2 = 1.5$

TABLEAU 6

Echantillons provenant du mélange 0.8 N(0.2,1) + 0.2 N(1.8,1)

Erreurs de deuxième espèce au niveau réel 5% pour 100 simulations sous H_1.

Taille de l'échantillon n Type d'estimateurs	50	100	150	200	250	300	400
Andrews	57%	40%	27%	11%	4%	4%	1%
Huber	53%	37%	23%	16%	9%	4%	3%
Moyenne	57%	42%	24%	17%	6%	6%	4%

TABLEAU 7

Echantillons provenant du mélange $0.7N(0.2, 1) + 0.3N(1.8, 1)$

Erreurs de deuxième espèce au niveau réel 5% pour 100 simulations.

Taille de l'échantillon n Type d'estimateurs	50	100	150	200	250	300	400
Andrews	54%	32%	15%	11%	6%	0%	2%
Huber	66%	24%	16%	8%	7%	3%	0%
Moyenne	64%	28%	17%	7%	2%	3%	1%

On observe que, pour ces simulations, l'utilisation d'estimateurs robustes n'apporte pas une amélioration significative des performances du test.

Ceci peut s'expliquer par la grande taille des échantillons qui rend moins sensible la moyenne à la présence de valeurs aberrantes, et, surtout par le faible écart entre les moyennes θ_1 et θ_2.

TABLEAU 8
Echantillons provenant du mélange 0.8N(0.5,1)+0.2N(1.8,1)
erreurs de deuxième espèce au niveau réel 5% pour 100 simulations

Taille de l'échantillon n Type d'estimateurs	50	100	150	200	250	300	400	500
Andrews	76%	64%	58%	48%	34%	26%	13%	8%
Huber	77%	68%	60%	46%	38%	33%	2%	5%
Moyenne	89%	66%	64%	48%	42%	31%	24%	13%

TABLEAU 9
Echantillons provenant du mélange 0.8N(0.5,1)+0.2N(1.5,1)
erreurs de deuxième espèce au niveau réel 5% pour 100 simulations.

Taille de l'échantillon n Type d'estimateurs	50	100	150	200	250	300	400	500	1000	2000
Andrews	93%	78%	80%	84%	69%	68%	60%	54%	25%	3%
Huber	84%	77%	84%	69%	74%	62%	54%	54%	22%	8%
Moyenne	94%	95%	85%	79%	73%	75%	67%	54%	19%	1%

4.2 Conclusion.

Les résultats obtenus par ce test apparaissent sensibles à la taille de l'échantillon n, à la valeur de $D = \mid \theta_2 - \theta_1 \mid$ (pour σ=1) et aux proportions du mélange.

Par ailleurs la puissance de ce test s'approche de la puissance maximale et apparait moins sensible à D lorsque n devient assez grand (Tableaux 8 et 9). Ces résultats montrent que les puissances obtenues en tenant compte de D, p et n, en adoptant la moyenne comme estimateur de θ_1 sont relativement proches de ceux établis par Mendell, Thode et Finch (1991), qui ont observé ce phénomène en se basant sur l'étude de la valeur de $D = \mid \theta_2 - \theta_1 \mid / \sigma$. Ils ont montré également que la puissance du TRMV d'un mélange gaussien univarié dépend de p quand $p < 0.2$ et $p > 0.8$.

Une des applications intéressantes d'un test du type ci-dessus concerne la détection de gêne majeur dans le cadre d'une amélioration génétique.

L'idée de base est que, sous l'hypothèse de présence d'un gêne majeur, la distribution des observations est un mélange de distributions, ce qui n'est pas le cas lorsque le gêne majeur est absent : *cf.* Thode, Finch et Mendell (1988).

REMERCIEMENTS

Les auteurs remercient P. CAZES et le Comité de rédaction pour les corrections et les améliorations qu'ils leur ont suggérées.

REFERENCES

ANDREWS D.F., BICKEL P.J., HAMPEL F.R., HUBER P.J., ROGERS W.M., TUKEY J.W. (1972), Robust Estimates of Location. Princeton University Press.

BERDAÏ A. (1991), Une étude asymptotique et numérique de la loi limite du test de mélange de deux lois contre une seule. Technical report, S.B.I.A., I.N.R.A. Toulouse.

CRAMER H., LEADBETTER M.R. (1967), Stationnary and related stochastic Process. New York : Wiley.

DAVIES R. B. (1977), Hypothesis testing when a nuisance parameter is present only under the aternative. Biometrika,64, 247-254.

DAVIES R. B. (1987), Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika,74, 33-43.

GHOSH J. K., SEN P.K. (1985), On the asymptotic performance of the Log likelihood ratio statistic for the mixture model and related results. Proc. Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer (vol.II), L.M. Le Cam and R.A. Olshen (Eds.). Monterey : Wadsworth, p. 789-806.

HARTIGAN J. A.(1985a), Statistical theory in clustering. J.Classification, 2, 63-76.

HARTIGAN J. A.(1985b), A failure of likelihood asymptotics for normal mixtures, Proc. Berkeley Conference in honor of Jerzy Neymann and Jack Kiefer (vol.99), L. M. Le Cam and R.A. Oshen (Eds). Monterey

KOTZ S., JOHNSON N.L. (1983), Encyclopedia of Statistical Sciences, vol.4, p. 3-4.

McLACHLAN G.J., BASFORD K.E. (1988), Mixture models : Inference and applications to clustering. Marcel Dekker, New York.

MENDELL N.R., THODE H.C., FINCH S.J., (1991), The likelihood ratio test for two component normal mixture problem : power and sample size analysis. Biometrics, 47, 1143-1148.

REDNER R.A.(1981), Note on the consistency of the maximum likelihood estimate for non identifiable distributions. Ann. Statist., 9, 225-228.

SELF S.G., LIANG K.Y. (1987), Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non standard conditions. JASA, vol. 82, N°398, 605-610.

THODE H.C., FINCH S.J., MENDEL N.R. (1988), Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals. Biometrics, 44, 1195-1201.

TITTERINGTON D.M. (1990), Some recent research in the analysis of mixture distributions. Statistics, 21,4, 619-641.

TITTERINGTON D.M., SMITH A.F.M., MAKOV U.E. (1985), Statistical analysis of finite mixture distributions. Wiley, London.

YAKOWITZ S.J., SPRAGINS J.D. (1968), On the identifiability of finite mixtures. IEEE Trans. Inform. Th., IT-16, 330-338.

WALD A. (1943), Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Am. Math. Soc, 54, 426-482.

WOLFE J.H. (1971), A Monte Carlo study of the sampling distribution of the likelihhod ratio for mixtures of multinomial distributions. Tech. Bull. STB,72-2, Nav. Pers & Tran. Res. Lab., San Diego.