REVUE DE STATISTIQUE APPLIQUÉE

B. GAREL

Note sur l'article : « Performances d'un test d'homogénéité contre une hypothèse de mélange gaussien »

Revue de statistique appliquée, tome 45, n° 1 (1997), p. 97-102

<http://www.numdam.org/item?id=RSA_1997__45_1_97_0>

Rev. Statistique Appliquée, 1997, **XLV** (1), 97-102

NOTE SUR L'ARTICLE : «PERFORMANCES D'UN TEST D'HOMOGÉNÉITÉ CONTRE UNE HYPOTHÈSE DE MÉLANGE GAUSSIEN»

PARU DANS LE NUMÉRO 1 DE LA REVUE DE STATISTIQUE APPLIQUÉE EN 1994

B. Garel

Laboratoire de Statistique et Probabilités (UPS) UMR C55830 et Institut National Polytechnique, ENSEEIHT, 2 rue Camichel, BP 7122, 31071 Toulouse Cedex 7

RÉSUMÉ

Nous apportons ici quelques compléments sur les résultats obtenus dans l'article de Berdaï et Garel(1994). La condition restrictive qui s'y trouvait, condition dite de séparation, peut en effet être levée. Nous donnons la statistique sans condition et indiquons comment calculer les valeurs critiques.

Mots-clés : *mélange gaussien, test de vraisemblance, loi asymptotique, maximum d'un processus gaussien.*

ABSTRACT

Some precisions are given about the results derived in the above referenced paper by Berdaï and Garel (1994). The assumed restrictive condition, the so-called separation condition, can be removed. We give the statistic without this condition and we show how to compute the percentage points.

I. Le cas $(1-p)\mathcal{N}(0,1)+p\mathcal{N}(\theta,1)$

Soient $X_1,\ldots,X_n$ n variables aléatoires réelles indépendantes et de même loi de densité f. La statistique du test du rapport des vraisemblances maximales (TRVM) d'une hypothèse H_0 d'homogénéité :

$$\forall x\ f(x)=\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}=\varphi(x)$$

contre l'hypothèse de mélange : $f(x)=(1-p)\varphi(x)+p\varphi(x-\theta)$
avec $\theta\neq 0, p\in]0,1[, \theta\in\Theta$ où Θ est un intervalle compact de $\mathbb{R}$ et sous la condition :

$$|\theta|\geqslant\varepsilon_0, \tag{1.1}$$

est donnée par (cf. Berdaï et Garel (1994), formules 2.12 et 2.14) :

$$-2\log\lambda_n = \sup_{\substack{\theta\in\Theta \\ |\theta|\geqslant\varepsilon_0}} (T_n(\theta))^2 \cdot \mathbf{I}_{\{T_n(\theta)\geqslant 0\}} + o_p(1) \tag{1.2}$$

où

$$T_n(\theta) = n^{-1/2} \sum_{i=1}^{n} \frac{e^{x_i\theta-\theta^2/2} - 1}{(e^{\theta^2}-1)^{1/2}}. \tag{1.3}$$

1.1 Le nouveau résultat

Sous certaines hypothèses Dacunha-Castelle et Gassiat (1995) ont signalé qu'il était possible de lever la condition (1.1). En procédant à la maximisation de la log-vraisemblance sous l'hypothèse $|\theta| \leqslant \varepsilon_0$ et à l'aide des travaux de Redner (1981) on est conduit à examiner un voisinage de $\mathcal{D} = \{(p,\theta)/p \in [0,1] \text{ et } \theta = 0\}$. A l'aide de certains développements où des paramètres locaux apparaissent, on montre qu'il suffit de maximiser la log-vraisemblance sur un $n^{-1/2}$ voisinage de $\mathcal{D}$ pour obtenir le maximum global. On remarque également que :

$$\lim_{\theta\to 0^+} T_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i; \tag{1.4}$$

la limite à gauche ne diffère de la limite à droite que par le signe. La discontinuité ainsi mise en évidence est donc sans importance pour T_n^2. On obtient alors :

Théorème 1.1 Soit $\tilde{T}_n(.)$ le processus défini par :

$$\tilde{T}_n(\theta) = T_n(\theta) \text{ si } \theta \neq 0, \tilde{T}_n(0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

Alors la statistique du TRVM est donnée par

$$-2\log\lambda_n = \sup_{\theta\in\Theta} (\tilde{T}_n(\theta))^2 \cdot \mathbf{I}_{\{\tilde{T}_n(\theta)\geqslant 0\}} + o_p(1).$$

De plus $-2\log\lambda_n$ converge en loi sous H_0 vers $\sup_{\theta\in\Theta}(\tilde{T}(\theta))^2 \cdot \mathbf{I}_{\{\tilde{T}(\theta)\geqslant 0\}}$

où $\tilde{T}(.)$ est un processus gaussien de même fonction d'autocovariance que $\tilde{T}_n(.)$

1.2 Fonction d'autocovariance.

On retrouve sur la fonction d'autocovariance la discontinuité observée en (1.4) :

$$\operatorname{cov}(T_n(\xi), T_n(\eta)) = \frac{e^{\xi\eta} - 1}{(e^{\xi^2} - 1)^{1/2}(e^{\eta^2} - 1)^{1/2}} = \gamma(\xi, \eta)$$

et

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

Cependant la tabulation peut être obtenue à l'aide de la borne de Davies (1977). Nous avons besoin de la dérivée seconde :

$$\rho_{11}(\eta) = \frac{\partial^2\gamma(\xi,\eta)}{\partial^2\xi} \text{ en } \xi = \eta \text{ donnée par } \left(e^{\eta^2} + \eta^2 e^{\eta^2} - e^{2\eta^2}\right) \Big/ (e^{\eta^2} - 1)^2 \tag{1.5}$$

et

$$\lim_{\eta\to 0}[-\rho_{11}(\eta)]^{1/2} = \frac{1}{\sqrt{2}},$$

ce qui montre qu'on ne retrouve pas sur la première diagonale la discontinuité signalée plus haut. Sur les intervalles du type $[-a, +a]$, $a > 0$ et compte tenu de la condition $\mathbf{I}_{\{\tilde{T}_n(\eta)\geqslant 0\}}$ la tabulation se fait sur le demi-intervalle $[0, a]$ à l'aide de la borne :

$$P\left[\sup_{\eta\in[0,a]} \tilde{T}_n(\eta) \geqslant c\right] \leqslant \Phi(-c) + \frac{1}{2\pi} e^{-1/2c^2} \int_0^a [-\rho_{11}(\eta)]^{1/2} d\eta. \tag{1.6}$$

2. Le cas $(1-p)\mathcal{N}(\theta_1, 1) + p\mathcal{N}(\theta_2, 1)$

On suppose ici que Θ est un ouvert relativement compact et que Θ_2 est un intervalle fermé inclus dans Θ. Nous voulons tester

$H_0 : \forall x\ f(x) = \varphi(x - \theta_1)$ où $\theta_1 \in \Theta$
contre $H_1 : f(x) = (1-p)\varphi(x-\theta_1) + p\varphi(x-\theta_2)$
$p \in]0, 1[, \theta_1 \in \Theta,\ \theta_2 \in \Theta_2,\ \theta_1 \neq \theta_2.$

2.1. Le nouveau résultat

Sous la condition $|\theta_2 - \theta_1| \geqslant \varepsilon_0$ la statistique du TRVM (cf. Berdaï et Garel (1994), formules 2.6 et 2.7) est donnée par

$$-2\log\lambda_n = \sup_{\substack{\eta_2 \in \Theta_2 \\ |\eta_2 - \theta_{10}| \geqslant \varepsilon_0}} (T_n(\eta_2))^2 \cdot \mathbf{I}_{\{T_n(\eta_2) \geqslant 0\}} + o_p(1) \tag{2.1}$$

où

$$T_n(\eta_2) = n^{-1/2} \sum_{i=1}^{n} \left(\frac{e^{-1/2(X_i - \eta_2)^2}}{e^{-1/2(X_i - \theta_{10})^2}} - 1 - (\eta_2 - \theta_{10})(X_i - \theta_{10}) \right) / D \tag{2.2}$$

où $D = \left[e^{(\eta_2 - \theta_{10})^2} - 1 - (\eta_2 - \theta_{10})^2 \right]^{1/2}$, et où θ_{10} désigne la vraie valeur de θ_1 sous H_0.

Pour l'étude de la statistique sans condition de séparation on peut se limiter, sans perte de généralité, à $p \in [0, 1/2]$.

A l'aide des résultats de Redner on peut se limiter à un voisinage de

$$\mathcal{D} = \{(p, \theta_1, \theta_2) \in [0, 1/2] \times \theta_{10} \times \theta_{10}\}.$$

A l'aide de développements pour $p \in [\varepsilon_0, 1/2]$ puis au voisinage de $(p, \theta_1, \theta_2) = (0, \theta_{10}, \theta_{10})$ où des paramètres locaux apparaissent, on montre qu'il suffit de se limiter à des $n^{-1/2}$ voisinages de $\mathcal{D}$ pour la maximisation de la log-vraisemblance. On obtient également que

$$\lim_{\eta_2 \to \theta_{10}} T_n(\eta_2) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(X_i - \theta_{10})^2 - 1}{\sqrt{2}}$$

puis le résultat suivant :

Théorème 2.1 Soit $\tilde{T}_n(.)$ le processus défini par :

$$\tilde{T}_n(\eta_2) = T_n(\eta_2) \text{ si } \eta_2 \neq \theta_{10}, \tilde{T}_n(\theta_{10}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(X_i - \theta_{10})^2 - 1}{\sqrt{2}}.$$

$$\text{Alors } -2\log\lambda_n = \sup_{\eta_2 \in \Theta_2} (\tilde{T}_n(\eta_2))^2 \cdot \mathbf{I}_{\{\tilde{T}_n(\eta_2) \geqslant 0\}} + o_p(1)$$

De plus $-2\log\lambda_n$ converge vers $\sup_{\eta_2\in\Theta_2}(\tilde{T}(\eta_2))^2\cdot\mathbf{1}_{\{\tilde{T}(\eta_2)\geqslant 0\}}$

où $\tilde{T}(.)$ est un processus gaussien de même fonction d'autocovariance que $\tilde{T}_n(.)$.

2.2. Fonction d'autocovariance

Cette fonction est donnée par

$$r(\xi,\eta)=\frac{e^{(\xi-\theta_{10})(\eta-\theta_{10})}-1-(\xi-\theta_{10}).(\eta-\theta_{10})}{\left[e^{(\xi-\theta_{10})^2}-1-(\xi-\theta_{10})^2\right]^{1/2}\left[e^{\eta-\theta_{10})^2}-1-(\eta-\theta_{10})^2\right]^{1/2}}$$

En faisant tendre ξ vers θ_{10} on obtient

$$r(\theta_{10},\eta)=\frac{2^{-1/2}(\eta-\theta_{10})^2}{[e^{(\eta-\theta_{10})^2}-1-(\eta-\theta_{10})^2]^{1/2}}\text{ puis }\lim_{\eta\to\theta_{10}}r(\theta_{10},\eta)=1.$$

La dérivée seconde de cette fonction calculée pour $\theta_{10}=0$ en $\xi=\eta$ est donnée par

$$\left.\frac{\partial^2 r(\xi,\eta)}{\partial\xi^2}\right|_{\xi=\eta}=\frac{\eta^4e^{\eta^2}-e^{2\eta^2}+2e^{\eta^2}-1}{(e^{\eta^2}-1-\eta^2)^2}=r_{11}(\eta)$$

et au voisinage de $\eta=0,[-r_{11}(\eta)]^{1/2}$ admet le développement de Taylor : $\frac{1}{\sqrt{3}}+\frac{x^2}{2\sqrt{27}}+\frac{x^4}{40\sqrt{243}}+0(x^6)$. On peut donc utiliser, là encore, la borne de Davies (1.6) pour effectuer la tabulation au voisinage de $\theta_{10}=0$.

Remarque : La méthode que nous venons de développer ici dans deux cas particuliers s'étend au cas général, cf. Garel (1996, a et b), moyennant une condition sur la dérivée seconde de la densité f sous H_0 : $\left(\frac{\partial^2 f}{\partial\theta^2}\right)\frac{1}{f}$ doit être continue en moyenne quadratique au voisinage de θ_{10} sous H_0. Une condition un peu plus forte assure la tension du processus. La même méthode peut s'appliquer dans certains cas, comme le cas gaussien, où un paramètre de nuisance est présent sous H_0. Certains calculs de cette note ont été réalisés à l'aide de Mathematica.

Remerciements

B. Garel remercie Pierre Cazes pour les corrections et améliorations qu'il lui a suggérées.

Références bibliographiques

Berdaï A, Garel B. (1994), Performances d'un test d'homogénéité contre une hypothèse de mélange gaussien, Rev. Statistique Appliquée, XLII,1,63-79.

Dacunha-Castelle D., Gassiat E. (1995), Théorie asymptotique de tests de vraisemblance pour des modèles localement coniques. Comptes Rendus de l'Académie des Sciences, 320,1367-1372.

Davies R.B. (1977), Hypothesis Testing when a Nuisance Parameter is Present only under the Alternative, Biometrika, 64,247-254.

Garel B. (1996a), Théorie asymptotique du test du rapport des vraisemblances d'un mélange à deux composants, Comptes Rendus de l'Académie des Sciences, 323,199-202.

Garel B. (1996b), Asymptotic Theory of the Likelihood Ratio Test for the Identification of a Mixture, Soumis.

Redner R.A. (1981), Note on the Consistency of the Maximum Likelihood Estimate for non Identifiable Distributions, Ann. Statist., 19,225-228.