C. R. Acad. Sci. Paris, Ser. I 334 (2002) 913–916

Statistique/*Statistics*
(Probabilités/*Probability Theory*)

Test asymptotiquement minimax pour une hypothèse nulle composite dans le modèle de densité

Christophe Pouet

L.A.T.P., Université de Provence, 39, rue F. Joliot-Curie, 13453 Marseille cedex 13, France

Reçu le 4 avril 2001 ; accepté après révision le 14 mars 2002

Note présentée par Paul Deheuvels.

Résumé Un échantillon de N variables aléatoires indépendantes et identiquement distribuées est considéré. Supposons que la densité appartienne à un espace de Hölder. Un test asymptotiquement minimax est construit pour le problème de test de l'hypothèse nulle : la densité appartient à un ensemble paramétrique, contre l'alternative : la densité est séparée de l'ensemble paramétrique pour la distance dans $\mathrm{L}_2[0,1]$. ***Pour citer cet article : C. Pouet, C. R. Acad. Sci. Paris, Ser. I 334 (2002) 913–916.*** © 2002 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

Asymptotically minimax test for a composite null hypothesis in the density model

Abstract Consider a sample of N random variables independent and identically distributed. Assume the density function belongs to a Hölder space. We construct an asymptotically minimax test and obtain the miximax rate of testing for the problem: the density function belongs to a parametric set versus the alternative: the distance in $\mathrm{L}_2[0,1]$ between the density function and the parametric set is bounded away from 0. ***To cite this article: C. Pouet, C. R. Acad. Sci. Paris, Ser. I 334 (2002) 913–916.*** © 2002 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

1. Introduction

Considérons un échantillon de N variables aléatoires réelles, $X_1, \ldots, X_N$, indépendantes et de même densité f. D'un point de vue pratique, il est intéressant pour un statisticien de commencer par réaliser un test d'adéquation à une famille de lois plutôt qu' à une loi précise. De tels exemples de tests sont fournis par les tests d'adéquation de Kolmogorov, von Mises–Smirnov ou du chi-deux (*voir* [1], Chapitre III). Le but poursuivi ici est d'étudier le problème de test d'une hypothèse nulle composite contre une alternative non-paramétrique avec l'approche minimax asymptotique.

Notons $\|\cdot\|$ la norme euclidienne de $\mathbb{R}^k$, $\|\cdot\|_2$ la norme usuelle sur l'espace $\mathrm{L}_2[0,1]$ et $f^{(k)}$ la k-ième dérivée d'une fonction f. Soit $\mathrm{H}(\beta, L)$ la classe de Hölder de paramètres $\beta = r + \alpha (0 < \alpha \leqslant 1)$ et $L : \mathrm{H}(\beta, L) = \{f \in \mathrm{L}_2[0,1] : |f^{(r)}(x) - f^{(r)}(y)| \leqslant L|x-y|^{\alpha}\}$, et soit D_1 l'ensemble des densités à support dans l'intervalle $[0,1]$.

Adresse e-mail : pouet@cmi.univ-mrs.fr (C. Pouet).

S1631-073X(02)02358-0/FLA

L'hypothèse nulle est constituée de l'ensemble $\Sigma(\Theta)$ famille paramétrique de densités indexée par Θ, sous-ensemble borné de $\mathbb{R}^k$, et telle que $\Sigma(\Theta) \subset \mathrm{H}(\beta, L) \cap D_1$. L'alternative est formée par $\Lambda_n(C)$, ensemble des densités qui est séparé de $\Sigma(\Theta)$ pour la distance usuelle dans $\mathrm{L}_2[0,1]$: $\Lambda(C\psi_N) = \{f \in \mathrm{H}(\beta, L) \cap D_1 : \inf_{\theta\in\Theta} \|f - f_\theta\|_2 \geqslant C\psi_N\}$, où C est une constante et ψ_N une suite décroissante, tendant vers 0 quand N tend vers l'infini. Le problème de test que nous considérons s'écrit

$$H_0 : \quad f \in \Sigma(\Theta), \tag{1}$$

contre

$$H_1 : \quad f \in \Lambda(C\psi_N). \tag{2}$$

DÉFINITION. – La *vitesse minimax de test* est définie comme étant la suite $\psi_N > 0$ qui vérifie les deux conditions suivantes : pour tout $0 \leqslant \delta \leqslant 1$ (niveau de test),

$$\exists \widehat{C} > 0,\ \forall C < \widehat{C} : \lim_{N\to\infty} \inf_{\Delta_N} \Big\{ \sup_{\theta\in\Theta} P_{f_\theta}(\Delta_N = 1) + \sup_{f\in\Lambda(C\psi_N)} P_f(\Delta_N = 0) \Big\} \geqslant \delta, \tag{3}$$

et

$$\exists \overline{C},\ \exists \bar{\Delta}_N,\ \forall C > \overline{C} : \lim_{N\to\infty} \Big\{ \sup_{\theta\in\Theta} P_{f_\theta}(\bar{\Delta}_N = 1) + \sup_{f\in\Lambda(C\psi_N)} P_f(\bar{\Delta}_N = 0) \Big\} \leqslant \delta. \tag{4}$$

Le test $\overline{\Delta}_N$ est appelé *test asymptotiquement minimax*.

Remarque 1. – $\widehat{C}$ et $\overline{C}$ dépendent a priori du niveau de test fixé à l'avance et des paramètres caractérisant la régularité de l'espace fonctionnel.

Remarque 2. – La condition (3) signifie qu'il n'existe aucun test garantissant que la somme des erreurs soit plus petite qu'un niveau de test δ si l'alternative s'approche trop rapidement de l'hypothèse nulle. La condition (4) signifie qu'il existe un test qui nous garantit que la somme des erreurs ne dépasse pas le niveau de test δ si l'alternative s'approche à une vitesse « raisonnable » de l'hypothèse nulle.

2. Résultats

Avant de construire notre test et de formuler le résultat, nous présentons les hypothèses requises sur l'ensemble $\Sigma(\Theta)$. L'hypothèse (A.1) concerne l'existence d'une fonction particulière dans $\Sigma(\Theta)$:

HYPOTHÈSE (A.1). – Il existe une fonction $f_{\theta_0} \in \Sigma(\Theta)$ telle que $f_{\theta_0} \in \mathrm{H}(\beta, L')$ avec $L' < L$ et $d = \inf_{t\in[0,1]} f_{\theta_0}(t) > 0$.

L'hypothèse (A.2) impose une régularité en θ de l'espace $\Sigma(\Theta)$:

HYPOTHÈSE (A.2). – Il existe $\nu > 0$ et $Q > 0$ tels que pour tout $\theta \in \Theta$ et tout $\tau \in \Theta$, $\|f_\theta - f_\tau\|_2 \leqslant Q\|\theta - \tau\|^\nu$.

Soit δ le niveau du test, $0 < \delta < 1$.

Maintenant, nous construisons explicitement la procédure de test. Posons $n = \lfloor N^{2/(4\beta+1)} \rfloor$ (où $[\cdot]$ désigne la partie entière) et $\psi_N = N^{-2\beta/(4\beta+1)}$. L'intervalle $[0,1]$ est divisé en n sous-intervalles égaux, $A_{1,n}, \ldots, A_{n,n} : A_{j,n} = \left[\frac{j-1}{n}, \frac{j}{n}\right]$ pour $j = 1, \ldots, n$. Notons $1_A(t)$ la fonction indicatrice d'un ensemble A. Posons $c_{j,n}(f) = \int_{A_{j,n}} f(t)\,\mathrm{d}t$ et $k_{j,n} = \sum_{l=1}^N 1_{A_{j,n}}(X_l)$. Le test $\overline{\Delta}_N$ est défini de la manière suivante,

$$\overline{\Delta}_N = \begin{cases} 0 & \text{si } \inf_{\theta\in\Theta} \sum_{j=1}^n (nk_{j,n}/N - nc_{j,n}(f_\theta))^2 < n^2/N + (n^{3/2}/N) f_*(4/\delta)^{1/2}, \\ 1 & \text{si } \inf_{\theta\in\Theta} \sum_{j=1}^n (nk_{j,n}/N - nc_{j,n}(f_\theta))^2 \geqslant n^2/N + (n^{3/2}/N) f_*(4/\delta)^{1/2}, \end{cases}$$

où $f_* = \sup_{f\in\mathrm{H}(\beta,L)\cap D_1} \sup_{0\leqslant t\leqslant 1} |f(t)|$. Cette quantité peut être calculée dans certains cas (*voir* [7] pour un problème similaire).

THÉORÈME. – *Sous les hypothèses* (A.1) *et* (A.2), *le test* $\bar{\Delta}_N$ *défini ci-dessus est asymptotiquement minimax et* $\psi_N = N^{-2\beta/(4\beta+1)}$ *est la vitesse minimax pour le problème de test* (1), (2).

Ce résultat appelle plusieurs commentaires. Tout d'abord, ce théorème généralise le résultat énoncé dans [4] concernant le test d'hypothèse dans le modèle de densité avec la loi uniforme sur $[0, 1]$ comme hypothèse nulle. La vitesse minimax pour l'hypothèse nulle composite reste la même que dans le cas de l'hypothèse nulle simple ; nous ne connaissons pas la constante exacte de séparation qui est aussi inconnue dans le cas d'une hypothèse nulle simple.

Pour une niveau de test fixé par le statisticien, le théorème présente un test asymptotiquement minimax qui peut être construit explicitement et un ordre de grandeur pour la distance entre l'hypothèse nulle et l'alternative. Ce résultat est intéressant car dans de nombreux cas, les tests proposés permettent d'obtenir une erreur de première espèce donnée et ne sont évalués que par l'erreur de seconde espèce sur une suite d'alternatives simples (*voir* [2] et [11]). Cette approche tend à laisser de côté la richesse de l'ensemble des alternatives.

Le théorème présenté s'étend au cas de densités définies sur un intervalle $[a, b]$, $-\infty < a < b < \infty$. L'étude des preuves permet de conjecturer que le résultat peut aussi être étendu au cas des densités à support sur $\mathbb{R}$. Ce théorème est à notre connaissance l'un des premiers à présenter un test asymptotiquement minimax et la vitesse minimax dans le cadre du modèle de densité pour une hypothèse nulle composite avec la norme $L_2[0, 1]$ comme distance. Certains résultats existent dans le modèle de densité pour des hypothèses nulles composites spécifiques telles que la symétrie ou l'indépendance (*voir* [5]). De nombreux résultats existent dans le cadre du modèle de régression discrète (*voir* [3,6] et [10]) et dans le modèle de bruit blanc Gaussien (*voir* [9]).

3. Idée de la preuve pour la borne supérieure

La preuve de la borne supérieure s'effectue en deux temps : l'étude de l'erreur de première espèce et l'étude de l'erreur de deuxième espèce. Nous commencerons par l'erreur de première espèce.

Si le vrai paramètre est θ', la statistique de test vérifie

$$\inf_{\theta\in\Theta}\left(\sum_{j=1}^{n}\left(\frac{nk_{j,n}}{N}-nc_{j,n}(f_\theta)\right)^2\right)\leqslant\sum_{j=1}^{n}\left(\frac{nk_{j,n}}{N}-nc_{j,n}(f_{\theta'})\right)^2.$$

Le terme de droite de cette inégalité s'écrit comme la somme de deux termes,

$$\sum_{j=1}^{n}\left(\frac{nk_{j,n}}{N}-nc_{j,n}(f_{\theta'})\right)^2=L_1(\theta')+L_2(\theta'),$$

avec

$$L_1(\theta')=\sum_{k,l=1_{k\neq l}}^{N}\left(\frac{n}{N}\right)^2\sum_{j=1}^{n}\big(\mathbb{1}_{A_j}(X_k)-c_{j,n}(f_{\theta'})\big)\big(\mathbb{1}_{A_j}(X_l)-c_{j,n}(f_{\theta'})\big)+\frac{n^2}{N},$$

$$L_2(\theta')=\frac{n^2}{N}\sum_{j=1}^{n}c_{j,n}(f_{\theta'})^2-2\sum_{k=1}^{N}\sum_{j=1}^{n}\mathbb{1}_{A_j}(X_k)c_{j,n}(f_{\theta'}).$$

Alors, on a l'inégalité suivante

$$P_{f_{\theta'}}(\bar{\Delta}_N=1)\leqslant P_{f_{\theta'}}\big(L_1(\theta')\geqslant T_N\big)+P_{f_{\theta'}}\big(L_2(\theta')>0\big),$$

où $T_N=n^2/N+(n^{3/2}/N)f_*(4/\delta)^{1/2}$.

L'inégalité de Chebyshev conduit à une majoration pour $P_{f_{\theta'}}(L_1(\theta')\geqslant T_N)$:

$$P_{f_{\theta'}}\big(L_1(\theta')\geqslant T_N\big)\leqslant\frac{2n^3(N-1)f_*^2}{N^3(T_N-n^2/N)^2}.$$

Le choix du seuil T_N implique que le membre de droite de l'inégalité précédente est majoré par $\delta/2$.

L'inégalité de Chebyshev permet aussi de majorer la probabilité $P_{f_{\theta'}}(L_2(\theta') > 0)$:

$$P_{f_{\theta'}}\big(L_2(\theta') > 0\big) \leqslant \frac{4f_*^3}{N(C_1 - C_2 n^{-\beta})^2},$$

où C_1 et C_2 sont des constantes ne dépendant que des paramètres β et L de l'espace de Hölder considéré.

Quant à l'erreur de deuxième espèce, il faut considérer une densité dans l'alternative. Soit f une telle densité. La statistique de test se décompose en deux parties, l'une ne dépendant que de f et l'autre faisant apparaître l'ensemble Θ,

$$P_f(\bar{\Delta}_N = 0) \leqslant P_f\left(R_1(f) \leqslant T_N - 2\frac{n^{3/2}}{N} f_*\left(\frac{4}{\delta}\right)^{1/2}\right) + P_f\left(R_2(f) \leqslant 2\frac{n^{3/2}}{N} f_*\left(\frac{4}{\delta}\right)^{1/2}\right), \qquad (5)$$

avec

$$R_1(f) = \sum_{j=1}^{n} \left(\frac{nk_{j,n}}{N} - nc_{j,n}(f)\right)^2,$$

$$R_2(f) = \inf_{\theta\in\Theta} \left\{ 2\sum_{j=1}^{n} \left(\frac{nk_{j,n}}{N} - c_{j,n}(f)\right) n\big(c_{j,n}(f) - c_{j,n}(f_\theta)\big) + n^2 \sum_{j=1}^{n} \big(c_{j,n}(f) - c_{j,n}(f_\theta)\big)^2 \right\}.$$

Le raisonnement qui a conduit à la borne supérieure pour l'erreur de première espèce s'applique au premier terme du membre de droite de l'inégalité précédente et permet de le majorer par $\delta/2$. L'inégalité de Bernstein (*voir* [8]) permet de montrer que le second terme du membre de droite de (5) tend vers 0 lorsque le nombre d'observations N tend vers l'infini.

4. Idée de la preuve pour la borne inférieure

La preuve de la borne inférieure s'apparente à la preuve de la borne inférieure dans le cas d'une hypothèse nulle simple (*voir* [4]).

Références bibliographiques

[1] A.A. Borovkov, Mathematical Statistics, Gordon and Breach, Amsterdam, 1998.
[2] R.L. Eubank, J.D. Hart, Testing goodness of fit in regression via order selection criteria, Ann. Statist. 20 (1992) 1412–1425.
[3] W. Härdle, E. Mammen, Comparing nonparametric versus parametric regression fits, Ann. Statist. 21 (1993) 1926–1947.
[4] Y.I. Ingster, Asymptotically minimax hypothesis testing for nonparametric alternatives I–II–III, Math. Methods Statist. (1993) 85–114, 171–189, 249–268.
[5] Y.I. Ingster, Minimax testing of the hypothesis of independence for ellipsoids in l_p, J. Math. Sci. 81 (1996) 2406–2420.
[6] J.L. Horowitz, V.G. Spokoiny, An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative, Econometrica 69 (2001) 599–631.
[7] O.V. Lepski, A.B. Tsybakov, Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point, Probab. Theory Related Fields 117 (2000) 17–48.
[8] D. Pollard, Convergence of Stochastic Processes, Springer, New York, 1984.
[9] C. Pouet, An asymptotically optimal test for a parametric set of regression functions against a non-parametric alternative, J. Statist. Plann. Inference 98 (2001) 177–189.
[10] V.G. Spokoiny, Testing a linear hypothesis using Haar transform, Rapport technique SFB 373, Humboldt Universität, Berlin, 1997.
[11] J.X. Zheng, A consistent test of functional form via nonparametric estimation techniques, J. Econometrics 75 (1996) 263–289.