Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 336 (2003) 955–958
COMPTES RENDUS MATHEMATIQUE

Statistique/Probabilités

Test d'indépendance nonparametrique

Armel Fabrice Yode

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

Reçu le 25 janvier 2003 ; accepté après révision le 15 avril 2003

Présenté par Paul Deheuvels

Résumé

Nous considérons le problème de test d'indépendance des coordonnées d'un vecteur aléatoire de dimension $d > 2$ et de densité f à support compact, contre une classe d'alternatives définie par la norme L_2. La résolution de ce problème se fait via l'approche minimax. Nous définissons la vitesse de test et une fonction de test dont la statistique est basée sur l'estimateur à noyau et qui atteint cette vitesse. L'erreur de première espèce est bornée par une suite positive pouvant tendre vers zéro quand le nombre d'observations devient assez grand. ***Pour citer cet article : A.F. Yode, C. R. Acad. Sci. Paris, Ser. I 336 (2003).***

Abstract

Asymptotically minimax testing of the hypothesis of independence. We consider the minimax problem of testing the independence of the components of a d-dimensional random vector against a set of alternatives defined by L_2-norm. We are interested in finding the minimax rate of testing and a test that attains this rate. The bound of the error of the first kind is a positive sequence which can decrease to zero as the number of observations increases. ***To cite this article: A.F. Yode, C. R. Acad. Sci. Paris, Ser. I 336 (2003).***

1. Introduction

Nous considérons l'expérience statistique engendrée par l'observation $X^n = (X - 1, \ldots, X - n)$, où $X_i = (X_i^{(1)}, \ldots, X_i^{(d)}) \in \mathbb{R}^d,\ d > 2,\ i = 1, \ldots, n,$ sont des vecteurs aléatoires indépendants identiquement distribués de densité commune $f(\cdot)$ par rapport à la mesure de Lebesgue μ. Soient Φ, l'ensemble de toutes les densités définies sur $\mathbb{R}^d$ et Φ_0 le sous ensemble de Φ défini par

$$\Phi_0 = \big\{ f(\cdot) \in \Phi\colon\ f(x_1, \ldots, x_d) = f_1(x_1) \cdots f_d(x_d) \big\}, \tag{1}$$

où les fonctions $f_k(\cdot),\ k = 1, \ldots, d,$ sont les densités marginales de $f(\cdot)$. En d'autres termes, les éléments de Φ_0 sont des densités qui s'écrivent comme produit de leurs densités marginales. De ce fait, si $f \in \Phi_0$ alors les coordonnées $X_i^{(j)},\ j = 1, \ldots, d,$ du vecteur aléatoire X_i sont indépendantes.

Adresse e-mail : yode@cmi.univ-mrs.fr (A.F. Yode).

1631-073X/03/$ – see front matter
doi:10.1016/S1631-073X(03)00207-3

Nous considérons, en se basant sur l'observation X^n, le problème de test nonparamétrique de l'hypothèse

$$H_0 : f(\cdot) \in \Phi_0 \tag{2}$$

contre la classe d'alternatives

$$H_n : f(\cdot) \in \Phi_n(\rho_n) = \Big\{ f(\cdot) \in \Phi \colon \inf_{f \in \Phi_0} d(f, f_0) \geqslant \varphi_n \Big\}, \tag{3}$$

où d est une distance et φ_n est une suite positive tendant vers 0 quand $n \to \infty$. De plus, nous supposerons dans la suite que les densités satisfont certaines conditions de régularité. Nous utiliserons l'approche minimax pour la résolution de ce problème.

Les études proches de cette Note dans la littérature, sont celles de Ingster [1,2]. Cependant, nous notons deux principales différences entre ses travaux et le nôtre. La première se situe au niveau du choix de la classe d'alternatives (3). Ingster [1,2], en considérant l'application $g(\cdot) : \Phi \to \Phi_0$ qui à toute densité f associe le produit de ses densités marginales $g(f) = \prod_{k=1}^{d} f_k(x_k)$, a introduit la famille d'applications $h_{g(f)}(\cdot) : \mathbb{R}^d \to \Delta = [0,1]^d,\ f \in \Phi$ définies par $h_{g(f)}(x_1, \ldots, x_d) = (F_1(x_1), \ldots, F_d(x_d))$, où $F_j(\cdot)$ est la fonction de répartition marginale associée à la densité marginale $f_j(\cdot)$, et qui génère des isomorphismes $H_{g(f)}(\cdot)$ des espaces de Hilbert $L_2(\Delta, \mathcal{B}, \mu)$ et $L_2(\mathbb{R}, \mathcal{B}_d, P_{g(f)})$ définis par

$$(H_{g(f)}\phi)(x) = \phi\big(h_{g(f)}(x)\big).$$

Ici, $\mathcal{B}$ et $\mathcal{B}_d$ sont des σ-algèbres de Borel. De ce fait Ingster [1,2] a considéré la classe d'alternatives de la forme

$$\Phi_n = \left\{ f(\cdot) \in \Phi \colon H_{g(f)}^{-1}\left(\frac{f}{g(f)} - 1 \right) \in \Phi_{n,\Delta} \right\}. \tag{4}$$

Dans [1], $\Phi_{n,\Delta} \subset L_p(\Delta)$ est une classe de fonctions ϕ satisfaisant certaines conditions de regularité et $\|\phi\|_p \geqslant \varphi_n$, et dans [2], $\Phi_{n,\Delta} \subset l_p$ (ellipsoïde) et $\|\phi\|_{l_p} \geqslant \varphi_n$ pour tout $\phi \in \Phi_{n,\Delta},\ 1 \leqslant p \leqslant \infty$.

La deuxième différence concerne la borne de l'erreur de première espèce qui est, dans cette Note, une suite positive qui peut décroître vers 0 quand $n \to \infty$, contrairement à Ingster [1,2] où cette borne est une constante fixée. Des conditions pour que l'on soit capable de distinguer l'hypothèse nulle définie par (2) de la classe d'alternatives définies par (4) ont été établies dans [1,2].

2. Problème de test

Nous supposerons, dans cette Note, que $f(\cdot) \in H_d(r, L, S),\ r = m + \tau,\ \tau \in (0,1],\ m$ est un entier naturel, $L > 0,\ S$ où $H_d(r, L, S)$ est l'espace de Hölder isotrope sur $[0,1]^d$ qui est l'ensemble des fonctions bornées ϕ définies sur $[0,1]^d$ ayant la même régularité r dans chaque direction, i.e., pour tout $Z^{(l)} = (z_1^{(l)}, \ldots, z_d^{(l)}) \in [0,1]^d,\ l = 1, 2$, nous avons

$$\sup_{(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_d)} \left| \frac{\partial^m f}{\partial z_i^m}\big(Z^{(1)}\big) - \frac{\partial^m f}{\partial z_i^m}\big(Z^{(2)}\big) \right| \leqslant L \big| z_i^{(1)} - z_i^{(2)} \big|^{\tau}, \quad i = 1, \ldots, d,$$

$$\sup_{z \in [0,1]^d} \big| f(z) \big| \leqslant S.$$

Posons $\Sigma = \Phi \cap H_d(r, L, S)$ et $\Sigma_0 = \Phi_0 \cap H_d(r, L, S)$.

Nous considérons la classe d'alternatives définie par

$$\Phi_n(\varphi_n) = \Big\{ f(\cdot) \in \Sigma \colon \inf_{g \in \Sigma_0} \|f - g\|_2 \geqslant \varphi_n \Big\}, \tag{5}$$

où φ_n est une suite positive tendant vers 0 et $\|\cdot\|_2$ est la norme de $L_2([0,1]^d)$.

Soit $\psi_n = \psi_n(X^n)$, une suite de tests, i.e., de fonctions mesurables dépendant uniquement des observations à valeurs dans $\{0, 1\}$. Nous acceptons l'hypothèse nulle si $\psi_n = 0$ et nous acceptons l'alternative si $\psi_n = 1$. L'erreur de première espèce et celle de deuxième espèce sont respectivement définies par

$$\sup_{f\in\Phi_0} P_f^n\{\psi_n = 1\} \quad \text{et} \quad \sup_{f\in\Phi_n(\varphi_n)} P_f^n\{\psi_n = 0\}.$$

La mesure P_f^n est la loi de probabilité du n-échantillon associée à la densité f. Les propriétés du test sont caractérisées par les erreurs de première et deuxième espèce.

Nous dirons que ψ_n est un test de niveau asymptotique $\alpha_n \in (0, 1)$ si

$$\sup_{f\in\Phi_0} P_f^n\{\psi_n = 1\} \leqslant \alpha_n\big(1 + \mathrm{o}_n(1)\big) \quad n \to \infty. \tag{6}$$

Notre but est de construire un test ψ_{n,α_n} de niveau asymptotique α_n tel qu'il existe une suite de nombres positifs $\beta_n(\alpha_n) \in (0, 1)$ satisfaisant

$$\sup_{f\in\Phi_n(\lambda\varphi_n(\alpha_n))} P_f^n\{\psi_n = 0\} \leqslant \beta_n(\alpha_n)\big(1 + \mathrm{o}_n(1)\big), \quad n \to \infty, \tag{7}$$

où $\varphi_n(\alpha_n)$, appelée la vitesse de test, est une suite positive tendant vers 0 quand $n \to \infty$ caractérisant la distance minimale entre l'hypothèse nulle et la classe d'alternatives telle que l'on peut distinguer l'hypothèse nulle de l'alternative et λ est une constante positive (voir [3]).

Remarque 1. Habituellement, α_n est indépendant de n. Mais α_n peut dépendre de n (voir Lepski [4]). Nous considérerons plus tard le cas $\alpha_n = \mathrm{O}_n(n^{-\gamma})$ quand $n \to \infty$, où γ est une constante strictement positive et fixée.

3. Résultats

Nous proposons une procédure de test basée sur l'estimateur à noyau de $f(\cdot)$ défini par

$$\hat{f}_n(x) = \frac{1}{nh_n^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in [0, 1]^d,$$

où h_n est une suite positive tendant vers 0, $nh_n^d \to \infty$ quand $n \to \infty$ et $K(\cdot) : \mathbb{R}^d \to \mathbb{R}$ est un noyau.

Soit $\hat{f}_{0n}(\cdot) = \prod_{k=1}^{d} \hat{f}_{kn}(\cdot)$, l'estimateur de $f(\cdot)$ sur Σ_0, où $\hat{f}_{kn}(\cdot)$ est l'estimateur de la densité marginale $f_k(\cdot)$, obtenu par la méthode du noyau en se basant sur l'observation $(X_1^{(k)}, \ldots, X_n^{(k)})$ et défini par $\hat{f}_{kn}(x_k) = (1/nb_n)\sum_{i=1}^{n} K_*((x_k - X_i^{(k)})/b_n)$ pour tout $x_k \in [0, 1]$ où b_n étant une suite positive tendant vers 0 telle que $nb_n \to \infty$ quand $n \to \infty$ et $K_*(\cdot) : \mathbb{R} \to \mathbb{R}$ est un noyau.

Nous introduisons des conditions permettant d'établir nos résultats :

(C1) K et K_* sont des fonctions lipschitziennes et à support compact sur $\mathbb{R}^d$ et $\mathbb{R}$ respectivement,
(C2) $\int_{[0,1]^d} u_1^{a_1} \cdots u_d^{a_d} K(u)\,\mathrm{d}u = 0$ pour tout $(a_1, \ldots, a_d) \in \mathbb{N}^d$ tels que $\sum_{i=1}^{d} a_i < r$, $\int_{[0,1]} u^a K_*(u)\,\mathrm{d}u = 0$ pour tout $a \in \mathbb{N}$ tel que $a < r$.

Nous considérons le test ψ_{n,α_n} basé sur la statistique suivante

$$T_n = \left\| \hat{f}_n - \prod_{k=1}^{d} \hat{f}_{kn} \right\|_2^2 - \frac{1}{n^2h^{2d}} \sum_{i=1}^{n} \int_{[0,1]^d} K^2\left(\frac{x - X_i}{h}\right) \mathrm{d}x \tag{8}$$

et défini par $\psi_{n,\alpha_n} = \mathbb{I}_{\{T_n > \lambda_n(\alpha_n)\}}$, où $\lambda_n(\alpha_n)$ qui définit la région critique du test est choisie de telle sorte que l'inégalité (6) soit vraie et $\mathbb{I}$ est la fonction indicatrice.

Posons $\lambda_n(\alpha_n) = (\lambda\varphi_n(\alpha_n))^2$, $\varphi_n(\alpha_n) = \big(n^{-1}\sqrt{\log\frac{2}{\alpha_n}}\big)^{2r/(4r+d)}$, $\lambda = 2^{(12r+d)/(8r+2d)} L_0^{d(4r+d)} \Upsilon^{r/(4r+d)}$,

$$L_0 = \frac{1}{m!}\sum_{j=1}^{d}\int_{[0,1]^d} |u_j|^r K(u)\,\mathrm{d}u, \qquad \Upsilon = S^2 \int_{[0,1]^{3d}} K(u)K(v)K(u+w)K(v+w)\,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w.$$

Pour toutes constantes positives A_*, B_*, γ et C_* vérifiant

$$A_* > \frac{8S\|K_*\|^2(r+2)}{2r+1}, \qquad B_* > \frac{8S\|K\|_2^2(4r+3d+4)}{2(4r+d)}, \tag{9}$$

$$\gamma < \min\left\{1, \frac{4r+3d+4}{2(4r+d)} - \frac{B_*}{8S\|K\|_2^2}; \frac{r+2}{2r+1} - \frac{A_*}{8S\|K_*\|_2^2}\right\}, \qquad C_* > 1+\sqrt{2}, \tag{10}$$

nous avons les résultats suivants

Théorème 3.1. *Supposons que les conditions C_1, C_2 sont satisfaites et que $r > \frac{d}{4}$ et $\alpha_n = \mathrm{O}_n(n^{-\gamma})$ quand $n \to \infty$. Alors,*

$$\sup_{f\in\Phi_0} P_f^n\big\{T_n \geqslant \big(\lambda\varphi_n(\alpha_n)\big)^2\big\} \leqslant \alpha_n\big(1+\mathrm{o}_n(1)\big), \quad n\to\infty. \tag{11}$$

Théorème 3.2. *Supposons que les conditions C_1, C_2 sont satisfaites et que $r > \frac{d}{4}$ et $\alpha_n = \mathrm{O}_n(n^{-\gamma})$. Posons*

$$\beta_n(\alpha_n) \triangleq \left(2\left(\frac{\alpha_n}{2}\right)^{(C_*^2-2C_*-1)^2} + \frac{8dQ_2}{A_*^{1/2}} n^{\frac{r+2}{2r+1}-\frac{A_*}{8S\|K_*\|_2^2}} + \frac{8Q_1}{B_*^{1/2}} n^{\frac{4r+3d+4}{2(4r+d)}-\frac{B_*}{8S\|K\|_2^2}} + \frac{16}{n}\right),$$

où Q_1 et Q_2 sont les constantes de Lipschitz respectives de K et K_*. Alors

$$\sup_{f\in\Phi_n(\lambda\varphi_n(\alpha_n))} P_f^n\big\{T_n \leqslant \big(\lambda\varphi_n(\alpha_n)\big)^2\big\} \leqslant \beta_n(\alpha_n)\big(1+\mathrm{o}_n(1)\big), \quad n\to\infty. \tag{12}$$

Les démonstrations de ces théorèmes sont basées sur des résultats de grandes déviations pour des U-statistiques dégénérées dont le noyau dépend de n et pour l'estimateur à noyau de la densité de probabilité.

Remarque 2. Sous les conditions (9), $\beta_n(\alpha_n) \to 0$ si $\alpha_n \to 0$ quand n tend vers l'infini.

Remarque 3. Si $\alpha_n \equiv \alpha$ pour tout n, nos résultats sont similaires à ceux de Ingster [1] pour le cas $p = 2$. De plus, sous les conditions (9), on a $\beta_n(\alpha) = 2(\frac{\alpha}{2})^{(C_*^2-2C_*-1)^2}(1+\mathrm{o}_n(1))$ quand $n\to\infty$.

Remerciements

Je tiens à remercier le Professeur Oleg V. Lepski pour son aide constante et Christophe Pouet pour l'intérêt qu'il a porté à mon travail.

Références

[1] Yu.I. Ingster, Asymptotically minimax testing of the hypothesis of independence, Zap. Nauchn. Sem. LOMI 153 (1986) 60–72. English translation: J. Soviet Math. 44 (1989) 466–476.

[2] Yu.I. Ingster, Minimax testing of the hypothesis of independence for ellipsoïds in l_p, Zap. Nauchn. Sem. POMI 207 (1993) 77–97. English translation: J. Math. Sci. 81 (1996) 2406–2420.

[3] O.V. Lepski, A.B. Tsybakov, Asymptotically exact nonparametric hypothesis testing in sup-norm and at fixed point, Probab. Theory Related Fields 117 (2000) 17–48.

[4] O.V. Lepski, How to improve the accuracy of estimation, Math. Methods Statist. 8 (1999) 441–486, 4 (2000).