Statistique et analyse des données

Oreste Nasi

Tests d'hypothèses contiguës et statistiques de rangs sérielles linéaires

Statistique et analyse des données, tome 13, n° 3 (1988), p. 56-64

<http://www.numdam.org/item?id=SAD_1988__13_3_56_0>

Statistique et Analyse des Données
1988 - Vol. 13 n° 3 p.56-64

TESTS D'HYPOTHESES CONTIGUES ET STATISTIQUES DE RANGS SERIELLES LINEAIRES

Oreste NASI

U.F.R. "M.I.M.",
Université de Metz, île du Saulcy, 57045 Metz Cedex 01.

Résumé : *Nous donnons un lemme qui permet de construire simplement des tests asymptotiquement les plus puissants dans le cas de suites d'hypothèses contiguës. Cela permet de déterminer de nouvelles statistiques de rangs sérielles linéaires pour les tests d'une hypothèse de bruit blanc contre une hypothèse de dépendance sérielle de type ARMA.*

Abstract : *A lemma gives a simple method to obtain asymptotically most poverful tests for contigous hypotheses .This results leads to new optimal linear serial rank statistics for the problem of testing randomness against alternatives of ARMA serial dependence.*

Mots clés : *Contiguïté, processus ARMA, tests de rangs.*

Indice de classification STMA : 05-040, 12-060, 05-070.

0. INTRODUCTION

L'étude de la puissance asymptotique locale des tests de rangs à partir des résultats de Le Cam (et, notamment, à partir du "troisième Lemme de Le Cam" - dans la terminologie popularisée par [2] -) est bien connue depuis les travaux de Hájek. (Pour une illustration récente de ces méthodes, on peut se reporter à Puri et Sen, [6]).

Dans le premier paragraphe, en nous plaçant dans le cas général de statistiques quelconques (et non seulement des statistiques de rangs), nous montrons que pour ce type de problème, les

Manuscrit reçu le 18 avril 1988
Révisé le 7 juillet 1988

tests asymptotiquement les plus puissants s'obtiennent sans utiliser le troisième Lemme de Le Cam.

Plus précisément, pour les suites d'hypothèses $H_0^{(n)}$: l'échantillon $X^{(n)}$ est de densité p_n contre $H_1^{(n)}$: l'échantillon $X^{(n)}$ est de densité q_n, où les densités q_n sont contiguës aux densités p_n, le lemme du paragraphe 1 montre qu'il suffit que des statistiques $S_n(X^{(n)})$ soient, à des constantes près, asymptotiquement équivalentes (en probabilité) aux variables $\text{Log } L_n(X^{(n)}) = \text{Log } (q_n(X^{(n)}) / p_n(X^{(n)}))$ sous les hypothèses $H_0^{(n)}$, pour que les tests bâtis sur ces statistiques soient asymptotiquement les plus puissants. Il n'est par conséquent pas indispensable de se servir du troisième Lemme de Le Cam qui nécessite la détermination parfois délicate de la loi limite dans $\mathbb{R}^2$ des couples $(S_n(X^{(n)}) , \text{Log } L_n(X^{(n)}))$.

Ce résultat simple n'apparaît pas clairement chez Hájek et Šidák ([2]) et chez les auteurs qui par la suite utilisent le même type de raisonnement (voir par exemple [1] et [3]). En fait les tests asymptotiquement les plus puissants y sont toujours obtenus quand $(a_n S_n(X^{(n)}) + b_n , \text{Log } L_n(X^{(n)}) + d^2/2)$ converge en loi sous les hypothèses $H_0^{(n)}$ vers une loi normale centrée de matrice des covariances du type $d^2 I$ (où I est la matrice identité 2 x 2), ce qui implique immédiatement que les statistiques $a_n S_n(X^{(n)}) + b_n$ sont asymptotiquement équivalentes aux statistiques $\text{Log } L_n(X^{(n)}) + d^2/2$.

Il est plus aisé de déterminer directement des statistiques $S_n(X^{(n)})$ équivalentes à $\text{Log}L_n(X^{(n)})$. L'utilisation du troisième Lemme de Le Cam garde évidemment son utilité pour l'étude de l'efficacité relative asymptotique au sens de Pitman quand on examine différentes lois possibles pour les hypothèses alternatives.

Dans le deuxième paragraphe nous appliquons le lemme du paragraphe 1 aux statistiques de rangs sérielles linéaires introduites par Hallin, Ingenbleek et Puri ([3]) pour tester l'hypothèse de bruit blanc contre l'alternative d'une dépendance de type ARMA.

On obtient ainsi facilement, pour ces tests, de nouvelles statistiques de rangs sérielles linéaires plus simples que celles proposées dans [3],[4] et [5]. Nos statistiques sont asymptotiquement équivalentes (à des $O_p(n^{-1/2})$) à celles de ces auteurs.

1. TESTS ASYMPTOTIQUEMENT LES PLUS PUISSANTS POUR DES HYPOTHESES CONTIGUES

Soient une suite d'échantillons $(X^{(n)})_{n\geq 1}$ définis sur (Ω,τ, P) et à valeurs dans des espaces mesurés $(\aleph_n,\mathcal{A}_n,\mu_n)$ et une suite de statistiques $(S_n(X^{(n)}))_{n\geq 1}$ à valeurs dans $\mathbb{R}$.
Pour tout $n\geq 1$ on considère le test des hypothèses :
$H_0^{(n)}$: $X^{(n)}$ est de densité p_n contre $H_1^{(n)}$: $X^{(n)}$ est de densité q_n, où p_n et q_n sont deux densités sur $(\aleph_n,\mathcal{A}_n)$ par rapport à μ_n et où

(1) on rejette $H_0^{(n)}$ si $S_n(x^{(n)}) > c_n$ $(c_n \in \mathbb{R})$.

Notons $\text{Log } L_n : \aleph_n \to [-\infty,+\infty]$ les logarithmes des rapports des densités définis par :

$$\text{Log } L_n(x^{(n)}) = \text{Log } (q_n(x^{(n)}) / p_n(x^{(n)}))$$

(avec les conventions $\text{Log } L_n(x^{(n)}) = -\infty$, 0 ou $+\infty$ suivant que $q_n(x^{(n)}) = 0 < p_n(x^{(n)})$, $q_n(x^{(n)}) = p_n(x^{(n)}) = 0$ et $q_n(x^{(n)}) > p_n(x^{(n)}) = 0$) et désignons par $\mathcal{N}(m, \sigma^2)$ la loi normale dans $\mathbb{R}$ d'espérance m et de variance σ^2 et par Φ la fonction de répartition de $\mathcal{N}(0,1)$.
A partir du théorème de Neyman-Pearson et des propriétés bien connues de la contiguïté (voir par exemple [7]) on obtient le résultat suivant :

Lemme : Soit $\sigma > 0$. Si, sous les hypothèses $H_0^{(n)}$,

(a) les statistiques $\text{Log } L_n(X^{(n)})$ convergent en loi vers $\mathcal{N}(-\sigma^2/2, \sigma^2)$,

(b) il existe des suites de réels $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$, avec les $a_n>0$, telles que :

$$a_n S_n(X^{(n)}) + b_n - \text{Log } L_n(X^{(n)}) \to 0 \text{ en probabilité,}$$

alors, pour tout α dans $]0,1[$, toute suite de tests de $H_0^{(n)}$ contre $H_1^{(n)}$ définie par (1) et de seuil asymptotique α, est asymptotiquement la plus puissante au seuil α et de puissance asymptotique $\Phi(\sigma - \Phi^{-1}(1-\alpha))$.
C'est en particulier le cas si dans (1) on prend $c_n = a_n^{-1}[\sigma \Phi^{-1}(1-\alpha) - \sigma^2/2 - b_n]$.

Exemple : soit le problème classique (voir [6], paragraphe 2.7) du test des hypothèses $H_0^{(n)}$: $X^{(n)}$ est un échantillon de n variables à valeurs dans $\mathbb{R}$, indépendantes et équidistribuées de fonction de répartition F_n continue, contre $H_1^{(n)}$: $X^{(n)}$ est de densité

$q_n(x^{(n)}) = \prod_{i=1}^{n} f(x_i - d_{ni})$, où f est une densité absolument continue de quantité d'information I(f) et les constantes de régression d_{ni} vérifient :

$$\lim_{n\to+\infty} \sum_{i=1}^{n} (d_{ni} - \bar{d}_n)^2 = D^2 \in]0,+\infty[\qquad \text{et} \qquad \lim_{n\to+\infty} \max_{1\le i\le n} (d_{ni} - \bar{d}_n)^2 = 0 \ .$$

A l'aide du lemme on retrouve facilement les tests asymptotiquement uniformément les plus puissants de Hajek et Sidak bâtis sur les statistiques de rangs

$S_n(R^{(n)}) = \sum_{i=1}^{n} (d_{n_i} - \bar{d}_n)\, a_n(R_i^{(n)})$ dont les scores vérifient :

$$\lim_{n\to+\infty} \int_0^1 [a_n(1 + [un]) - \varphi_f(u)]^2 \, du = 0 \qquad \text{avec} \qquad \varphi_f(u) = - \frac{f'(F^{-1}(u))}{f(F^{-1}(u))} \ .$$

En effet si les $X^{(n)}$ sont de densités $p_n(x^{(n)}) = \prod_{i=1}^{n} f(x_i - \bar{d}_n)$, on a :

$S_n(R^{(n)}) - T_n \to 0$ en probabilité et $T_n - \frac{1}{2} I(f) D^2 - \text{Log } L_n(X^{(n)}) \to 0$ en probabilité, avec

$$T_n = \sum_{i=1}^{n} (d_{n_i} - \bar{d}_n) \left[- \frac{f'(X_i^{(n)})}{f(X_i^{(n)})} \right].$$

2. STATISTIQUES DE RANGS SERIELLES POUR TESTER L'HYPOTHESE D'UN BRUIT BLANC CONTRE UNE ALTERNATIVE ARMA

Hallin, Ingenbleek et Puri proposent dans [3] des statistiques de rangs sérielles linéaires pour tester l'hypothèse qu'un échantillon $X^{(n)} = (X_1^{(n)},...,X_n^{(n)})$ est constitué de n variables indépendantes de même densité f, contre l'hypothèse qu'il s'agit de n observations consécutives d'un processus ARMA (p_1,p_2) avec $\min(p_1,p_2) \geq 1$.

Le lemme du paragraphe 1 permet de trouver de manière directe des statistiques de rangs sérielles optimales plus simples que celles proposées par ces auteurs.

Soit $(\xi_t)_{t \in \mathbb{Z}}$ un bruit blanc à valeurs dans $\mathbb{R}$ tel que les variables ξ_t sont indépendantes, centrées et de même densité f, et soient pour tout $n \geq 1$, les hypothèses :

$H_o^{(n)}$: $X^{(n)} = (\xi_1,...,\xi_n)$

$H_1^{(n)}$: $X^{(n)} = (X_1^{(n)}, ..., X_n^{(n)})$ est constitué de n observations consécutives du processus ARMA (p_1,p_2), noté $(X_t^{(n)})_{t \in \mathbb{Z}}$, de bruit blanc $(\xi_t)_{t \in \mathbb{Z}}$, défini par :

$$(1) \quad X_t^{(n)} - n^{-1/2} \sum_{i=1}^{p_1} a_i X_{t-i}^{(n)} = \xi_t + n^{-1/2} \sum_{i=1}^{p_2} b_i \xi_{t-i} \qquad \forall t \in \mathbb{Z},$$

où $a_1,...,a_{p1}$, $b_1,...,b_{p2}$ sont des réels non tous nuls.

Les statistiques de rangs sérielles linéaires sont définies par :

$$(2) \quad S_n = \frac{1}{n-p} \sum_{t=p+1}^{n} a_n(R_t^{(n)}, R_{t-1}^{(n)},..., R_{t-p}^{(n)})$$

où $p=\max(p_1,p_2)$, $(R_1^{(n)},...,R_n^{(n)})$ est le vecteur des rangs de $X^{(n)}$ et les scores $a_n(i_1,...,i_{p+1})$ sont des réels pour tous $i_1,...,i_{p+1} \in \{1,2,...,n\}$.

S'il existe une application $J :]0,1[^{p+1} \to \mathbb{R}$ telle que :

$$\int_{]0,1[^{p+1}} J^2(u_{p+1}, ...,u_1)\, du_{p+1} ...du_1 \in]o,+\infty[\quad \text{et}$$

$$(3) \quad \lim_{n \to +\infty} E_o\left(\left[a_n(R_{p+1}^{(n)}, ...,R_1^{(n)}) - J(F(X_{p+1}^{(n)}), ...,F(X_1^{(n)})) \right]^2 \right) = 0$$

où les espérances sont prises quand les hypothèses $H_o^{(n)}$ sont vérifiées et F désigne la fonction de répartition de la densité f, on dit que les statistiques de rangs sérielles définies en (2) sont de fonction génératrice des scores J. On supposera également que la densité f des variables ξ_t satisfait aux conditions suivantes :

i) $E(|\xi_t|^3) < +\infty$ et $E(\xi_t^2) = \sigma^2 > 0$

ii) f est absolument continue sur les intervalles finis avec $\int_{-\infty}^{+\infty} |f'(x)|\, dx < +\infty$, et de quantité d'information $I(f) = \int_{-\infty}^{+\infty} \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\, dx$ finie et non nulle,

iii) l'application φ, définie par $\varphi(x) = -f'(x)/f(x)$, est presque partout dérivable de dérivée vérifiant, pour une constante A, $|\varphi'(x) - \varphi'(y)| \leq A\, |x-y|$.

Complétons éventuellement par des 0 les suites $a_1,...,a_{p1}$ et $b_1,...,b_{p2}$ qui définissent le processus ARMA en (1), pour obtenir deux suites de longueur $p = \max(p_1,p_2)$ et notons

$$\| a+b \|^2 = \sum_{i=1}^{p} (a_i + b_i)^2.$$

Considérons maintenant la fonction génératrice :

$$(4) \qquad J(u_{p+1}, ...,u_1) = \varphi(F^{-1}(u_{p+1})) \sum_{i=1}^{p} (a_i + b_i)\ F^{-1}(u_{p+1-i}),$$

où $F^{-1}(u) = \inf\{x \in \mathbb{R} / F(x) \geq u\}$.

Proposition : Soient une suite de statistiques de rangs sérielles S_n de fonction génératrice des scores J donnée par (4) et m_n l'espérance de S_n quand $H_o^{(n)}$ est vérifiée.

Alors pour tout α dans]0,1[, les tests des hypothèses $H_o^{(n)}$ contre $H_1^{(n)}$ où on rejette $H_o^{(n)}$ pour les valeurs de $\sqrt{n-p}\ (S_n - m_n)$ supérieures à $\Phi^{-1}(1-\alpha)\ \| a+b \|\ \sigma\ I(f)^{1/2}$ sont asymptotiquement les plus puissants au seuil α et de puissance asymptotique égale à :

$\Phi(\| a+b \|\ \sigma\ I(f)^{1/2} - \Phi^{-1}(1-\alpha))$.

Preuve. Considérons les densités p_n et q_n des échantillons $X^{(n)}$ respectivement sous les hypothèses $H_o^{(n)}$ et $H_1^{(n)}$, et les statistiques correspondantes $\text{Log}\ L_n(X^{(n)})$ définies comme au paragraphe 1. Dans [3] (proposition 3.1 et paragraphes 4.1 et 4.2) il est établi que, sous $H_o^{(n)}$,

$\text{Log}\ L_n(X^{(n)})$ converge en loi vers $\eta(-\rho^2/2, \rho^2)$ avec $\rho^2 = \| a+b \|^2\ \sigma^2\ I(f)$,

et pour des statistiques de rangs sérielles S_n de fonction génératrice des scores J quelconque

$$(5) \qquad \sqrt{n-p}\ (S_n - m_n) - \rho^2/2 - \text{Log}\ L_n(X^{(n)}) - \{ \mathcal{S}_n - \mathcal{L}_n^0 - \sqrt{n-p}\ U_n \} \to 0 \text{ en probabilité,}$$

où, en négligeant les indices n des variables $X_i^{(n)}$,

$$\mathcal{S}_n = \frac{1}{\sqrt{n-p}} \sum_{k=p+1}^{n} J(F(X_k),F(X_{k-1}), ..., F(X_{k-p})) ,$$

$$\mathcal{L}_n^0 = \frac{1}{\sqrt{n-p}} \sum_{k=p+1}^{n} \varphi(X_k) \sum_{i=1}^{p} (a_i + b_i) X_{k-i}$$

et $$U_n = \frac{1}{\binom{n-p}{p+1}(p+1)!} \sum_{p+1 \leq k_1 \neq \ldots \neq k_{p+1} \leq n} J(F(X_{k_1}), \ldots, F(X_{k_{p+1}})) .$$

Avec J définie par (4) on obtient d'une part que $\mathcal{S}_n = \mathcal{L}_n^0$ presque sûrement sous $H_o^{(n)}$, et d'autre part, toujours sous $H_o^{(n)}$, que U_n est une U-statistique de noyau symétrique d'ordre p+1 définie à partir de variables indépendantes et équidistribuées. La convergence de $\sqrt{n-p}\ U_n$ vers 0 en probabilité s'en déduit facilement (voir par exemple Serfling, [8], lemme 5.2.2.B). En remplaçant dans (5) on obtient :

$$\sqrt{n-p}\ (S_n - m_n) - \rho^2/2 - \text{Log}\, L_n(X^{(n)}) \rightarrow 0 \quad \text{en probabilité,}$$

ce qui, avec le lemme du paragraphe 1 achève la démonstration.

Exemple : Supposons la densité f fortement unimodale (i.e. -Log f est une fonction convexe). Alors, pour J donnée par (4), les scores

$a_n(i_1, \ldots, i_{p+1}) = J(\frac{i_1}{n+1}, \ldots, \frac{i_{p+1}}{n+1})$ vérifient la condition (3) .

On obtient donc les tests optimaux de la proposition avec les statistiques :

$$(6) \qquad S_n = \frac{1}{n-p} \sum_{i=1}^{p} (a_i + b_i) \left\{ \sum_{t=p+1}^{n} \varphi(F^{-1}(\frac{R_t^{(n)}}{n+1}))\ F^{-1}(\frac{R_{t-i}^{(n)}}{n+1}) \right\} ,$$

alors qu'avec la fonction génératrice des scores plus compliquée proposée dans [3] :

$$J(u_{p+1}, \ldots, u_1) = \sum_{i=1}^{p} \frac{(a_i + b_i)}{p+1-i} \sum_{j=0}^{p-i} \varphi(F^{-1}(u_{p+1-j}))\ F^{-1}(u_{p+1-j-i}) ,$$

on utilise les statistiques :

$$S'_n = \frac{1}{n-p} \sum_{i=1}^{p} \frac{(a_i + b_i)}{p+1-i} \left\{ \sum_{t=p+1}^{n} \sum_{j=0}^{p-i} \varphi(F^{-1}(\frac{R_{t-j}^{(n)}}{n+1}))\ F^{-1}(\frac{R_{t-j-i}^{(n)}}{n+1}) \right\} .$$

Ainsi pour des alternatives $H_1^{(n)}$ où on considère le processus ARMA (2,1) :

$$X_t^{(n)} + \frac{1,4}{\sqrt{n}} X_{t-1}^{(n)} + \frac{0,5}{\sqrt{n}} X_{t-2}^{(n)} = \xi_t + \frac{0,2}{\sqrt{n}} \xi_{t-1} ,$$

de bruit blanc ξ_t de densité logistique (exemple traité par Hallin et al. dans [3]), on a respectivement :

$$S_n = \frac{1}{n-2} \sum_{t=3}^{n} \left(\frac{2R_t^{(n)}}{n+1} - 1 \right) \left[-1{,}2 \operatorname{Log}\left(\frac{R_{t-1}^{(n)}}{n+1-R_{t-1}^{(n)}} \right) - 0{,}5 \operatorname{Log}\left(\frac{R_{t-2}^{(n)}}{n+1-R_{t-2}^{(n)}} \right) \right] \quad \text{et}$$

$$S'_n = \frac{1}{n-2} \sum_{t=3}^{n} \left\{ -0{,}6 \left[\left(\frac{2R_t^{(n)}}{n+1} - 1 \right) \operatorname{Log}\left(\frac{R_{t-1}^{(n)}}{n+1-R_{t-1}^{(n)}} \right) + \left(\frac{2R_{t-1}^{(n)}}{n+1} - 1 \right) \operatorname{Log}\left(\frac{R_{t-2}^{(n)}}{n+1-R_{t-2}^{(n)}} \right) \right] - 0{,}5 \left(\frac{2R_t^{(n)}}{n+1} - 1 \right) \operatorname{Log}\left(\frac{R_{t-2}^{(n)}}{n+1-R_{t-2}^{(n)}} \right) \right\}$$

Les mêmes auteurs proposent dans d'autres articles ([4] et [5]) des tests asymptotiquement les plus puissants de $H_o^{(n)}$ contre $H_1^{(n)}$ où on rejette $H_o^{(n)}$ si les valeurs de

$\sqrt{n}\,(S''_n - m_n)/\sigma^{(n)}$ sont supérieures à $\Phi^{-1}(1-\alpha)\,\|a+b\|$, avec :

$$S''_n = \sum_{i=1}^{p} \frac{(a_i + b_i)}{n-i} \left\{ \sum_{t=i+1}^{n} \varphi\left(F^{-1}\left(\frac{R_t^{(n)}}{n+1}\right)\right) F^{-1}\left(\frac{R_{t-i}^{(n)}}{n+1}\right) \right\}$$

où $\sigma^{(n)}$ est l'écart-type de $\varphi(F^{-1}((n+1)^{-1} R_1^{(n)}))\, F^{-1}((n+1)^{-1} R_2^{(n)})$ sous $H_o^{(n)}$.

Ces tests où $\sigma^{(n)}$ dépend de n et où le nombre de termes de la deuxième somme qui définit S''_n dépend de i (comparer à (6)) restent plus compliqués que ceux que nous proposons. Cependant les coefficients d'autocorrélation des rangs qui servent à construire les statistiques S''_n sont la base de développements ultérieurs sur les statistiques de rangs quadratiques, permettant de traiter le cas où les coefficients du processus ARMA de l'alternative sont non spécifiés, au sens de [4] et [5].

Références :

[1] CONOVER, W.J., "Rank tests for one sample, two samples and k-samples without the assumption of a continuous distribution function", Annals of Statistics, 1973, vol. 1, n° 6, pp.1105-1125.

[2] HÁJEK, J. et SÎDÁK, Z. , Theory of rank tests, Academic Press, 1967.

[3] HALLIN, M. , INGENBLEEK, J.F. et PURI, M.L., "Linear serial rank tests for randomness against ARMA alternatives", Annals of Statistics, 1985, vol. 13, n° 3, pp.1156-1181.

[4] HALLIN, M. , INGENBLEEK, J.F. et PURI, M.L., "Linear and quadratic serial rank tests for ramdomness against serial dependence", Journal of Time Series Analysis, 1987, vol. 8, n° 4, pp. 409-424.

[5] HALLIN, M. et PURI, M.L. , "Optimal rank-based procedures for time series analysis : testing an ARMA model against other ARMA models", Annals of Statistics, 1988, vol. 16, n°1 pp. 402-432.

[6] PURI, M.L. et SEN, P.K., Nonparametric methods in general linear models, Wiley, 1985.

[7] RAOULT, J.P. , "Contiguity in asymptotical statistics", dans : Model Choise, Publications des Facultés Universitaires Saint-Louis,1985.

[8] SERFLING, R.J. , Approximation theorems of mathematical statistics, Wiley, 1980.